# Handbook of

# Human Factors and Ergonomics

**Fourth Edition**

EDITED BY

## Gavriel Salvendy

# HANDBOOK OF HUMAN FACTORS
# AND ERGONOMICS

# HANDBOOK OF HUMAN FACTORS AND ERGONOMICS

Fourth Edition

Edited by
## Gavriel Salvendy
*Purdue University*
*West Lafayette, Indiana*
and
*Tsinghua University*
*Beijing, People's Republic of China*

# CONTENTS

# ABOUT THE EDITOR

Gavriel Salvendy is professor emeritus of Industrial Engineering at Purdue University and Chair Professor and Head of the Department of Industrial Engineering at Tsinghua University, Beijing, China. He is the author or co-author of over 550 research publications, including over 300 journal papers and is the author or editor of 42 books. His publications have appeared in seven languages. He is the major professor to 67 former and current Ph.D. students. His main research deals with the human aspects of design, operation, and management of advanced engineering systems. Gavriel Salvendy is the founding editor of two journals: the *International Journal on Human–Computer Interaction and Human Factors* and *Ergonomics in Manufacturing and Service Industries*. He was the founding chair of the International Commission on Human Aspects in Computing, headquartered in Geneva, Switzerland. In 1990 he became the first member of either the Human Factors and Ergonomics Society or the International Ergonomics Association to be elected to the National Academy of Engineering. He was elected "for fundamental contributions to and professional leadership in human, physical, and cognitive aspects of engineering systems." In 1995 he received an Honorary Doctorate from the Chinese Academy of Sciences, "for great contributions to the development of science and technology and for the great influence upon the development of science and technology in China." He is the fourth person in all fields of science and engineering in the 45 years of the Academy ever to receive this award. In 2006, he received the Friendship Award presented by the People's Republic of China. The award is the highest honor the Chinese government confers on foreign experts. In 2007, he received the John Fritz Medal which is the engineering profession's highest award for his "fundamental international and seminal leadership and technical contributions to human engineering and industrial engineering education, theory, and practice." He is an honorary fellow and life member of the Ergonomics Society and fellow of Human Factors and Ergonomics Society, Institute of Industrial Engineers, and the American Psychological Association. He has advised organizations in 31 countries on the human side of effective design, implementation and management of advanced technologies in the workplace. He earned his Ph.D. in engineering production at the University of Birmingham, United Kingdom.

# CONTRIBUTORS

Bashar Alyousef
Research Assistant
Industrial & Systems Engineering
University of Wisconsin—Madison
Madison, Wisconsin

Johan Hviid Andersen
Professor
Department of Occupational Medicine
Danish Ramazzini Centre
Herning Hospital
Herning, Denmark

Margherita Antona
Coordinator of the Centre for Universal Access
    and Assistive Technologies
Institute of Computer Science
Foundation for Research and
    Technology—Hellas (FORTH)
Heraklion, Crete, Greece

Susan Archer
Senior Vice President
Alion Science and Technology
Boulder, Colorado

Grazyna Bartkowiak
Researcher, Department of Personal Protective
    Equipment
Lab Manager, Protective Clothing Laboratory
Central Institute for Labour Protection—National
    Research Institute
Warsaw, Poland

Krzysztof Baszczynski
Researcher, Department of Personal Protective
    Equipment
Lab Manager, Laboratory of Safety Helmets and
    Equipment Protecting against Falls from a
    Height
Central Institute for Labour Protection—National
    Research Institute
Warsaw, Poland

Kevin B. Bennett
Professor and Human Factors Area Leader
Department of Psychology
Wright State University
Dayton, Ohio

Kenneth R. Boff
Principal Scientist
Tennenbaum Institute
Georgia Institute of Technology
Atlanta, Georgia

Anna Bogdan
Researcher, Department of Ergonomics
Lab Manager, Thermal Load Laboratory
Central Institute for Labour Protection—National
    Research Institute
Warsaw, Poland

Walter R. Boot
Assistant Professor
Department of Psychology
Florida State University
Tallahassee, Florida

Walter C. Borman
Professor
Department of Psychology
University of South Florida
and Personnel Decisions Research Institute
Tampa, Florida

Peter R. Boyce
Consultant
Canterbury, United Kingdom

Martin Braun
Senior Scientist
Fraunhofer Institute for Industrial Engineering
Stuttgart, Germany

Agnieszka Brochocka
Researcher, Department of Personal Protective
    Equipment
Laboratory for Respiratory Protective Equipment
Central Institute for Labour Protection—National
    Research Institute
Warsaw, Poland

Patrick F. Bruning
Graduate Research Assistant
Organizational Behavior and Human Resources
Krannert School of Management
School of Business
Purdue University
West Lafayette, Indiana

Michael A. Campion
Herman C. Krannert Professor of Management
Krannert Graduate School of Management
Purdue University
Weat Lafayette, Indiana

Pascale Carayon
Proctor & Gamble Bascom Professor in Total
    Quality
Department of Industrial and Systems
    Engineering
Director of the Center for Quality and
    Productivity Improvement
University of Wisconsin—Madison
Madison, Wisconsin

C. Melody Carswell
Associate Professor
Department of Psychology
University of Kentucky
Lexington, Kentucky

John G. Casali
John Grado Professor
Grado Department of Industrial and Systems
    Engineering
Director, Auditory Systems Laboratory
Virginia Polytechnic Institute and State
    University
Blacksburg, Virginia

Yee-Yin Choong
Research Scientist
National Institute of Standards and Technology
Gaithersburg, Maryland

Joseph V. Cohn
Aerospace Experimental Psychologist
Office of Naval Research
U.S. Navy
Arlington, Virginia

Bradley T. Cook
Managing Engineer
Miller Engineering, Inc.
Ann Arbor, Michigan

Chris W. Coultas
Graduate Research Associate
Department of Psychology & Institute of
    Simulation and Training
University of Central Florida
Orlando, Florida

Sara J. Czaja
Professor and Co-Director
Department of Psychiatry and Behavioral Center
    on Aging Sciences
University of Miami Miller School
    of Medicine
Miami, Florida

Marvin Dainoff
Director, Center for Behavioral Sciences
Liberty Mutual Research Institute for Safety
Hopkinton, Massachusetts

Patrick G. Dempsey
Research Industrial Engineer
National Institute for Occupational Safety
    and Health
Pittsburgh, Pennsylvania

Konstantina Doulgeraki
Research Engineer
Institute of Computer Science
Foundation for Research and
    Technology—Hellas (FORTH)
Heraklion, Crete, Greece

Colin G. Drury
Distinguished Professor Emeritus
Department of Industrial Engineering
University at Buffalo: State University of
    New York
Buffalo, NewYork

Vincent G. Duffy
Associate Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana

Paula J. Edwards
Director
HIMformatics
Atlanta, Georgia

Mica R. Endsley
President
SA Technologies, Inc.
Marietta, Georgia

Edna R. Fiedler
Aerospace Psychology Consultants
Tucson, Arizona

Donald L. Fisher
Professor and Head
Department of Mechanical & Industrial
    Engineering
University of Massachusetts—Amherst
Amherst, Massachusetts

Arthur D. Fisk
Professor
School of Psychology
Georgia Institute of Technology
Atlanta, Goergia

John M. Flach
Professor and Department Chair
Department of Psychology
Wright State University
Dayton, Ohio

Sven Fuchs
Senior Research Associate
Human-Systems Integration
Design Interactive, Inc.
Oviedo, Florida

Fredrick P. Garcia
Lab Manager/Lead Usability Specialist
Center for Usability in Design and Accessibility
Long Beach, California

Paul A. Green
Research Professor
Transportation Research Institute
University of Michigan
Ann Arbor, Michigan

Michael J. Griffin
Professor
Human Factors Research Unit
Institute of Sound and Vibration Research
University of Southampton
Southampton, United Kingdom

Rebecca Grossman
Graduate Research Associate
Department of Psychology
Institute for Simulation & Training
University of Central Florida
Orlando, Florida

Kelly S. Hale
Senior Vice President
Design Interactive, Inc.
Oviedo, Florida

Jerry W. Hedge
Senior Research Manager
Survey Research Division
RTI International
Research Triangle Park, North Carolina

Martin G. Helander
Professor
School of Mechanical and Aerospace
    Engineering
Nanyang Technological University
Singapore

Fabian Hermann
Competence-Center Human-Computer
    Interaction
Fraunhofer Institute for Industrial
    Engineering
Stuttgart, Germany

Erik Hollnagel
Professor
University of Southern Denmark
Odense, Denmark

Peter Hoonakker
Research Scientist
Center for Quality & Productivity Improvement
    (CQPI)
University of Wisconsin—Madison
Madison, Wisconsin

Rafal Hrynyk
Assistant Researcher
Laboratory of Head and Foot Protection
Department of Personal Protective Equipment
Central Institute for Labour Protection—National
    Research Institute
Warsaw, Poland

Emilia Irzmanska
Researcher
Laboratory of Head and Foot Protection
Department of Personal Protective Equipment
Central Institute for Labour Protection—National
    Research Institute
Warsaw, Poland

Dan Ispas
Assistant Professor
Department of Psychology
Illinois State University
Normal, Illinois

Julie A. Jacko
Professor
School of Public Health Faculty Fellow
Institute for Health Informatics
Principal Investigator and Director
University Partnership for Health Informatics
University of Minnesota
Minneapolis, Minnesota

J. Bern Jordan
Researcher
Industrial & Systems Engineering
University of Wisconsin—Madison
Madison, Wisconsin

Tanja Kabel
RWTH Aachen University
Aachen, Germany

Waldemar Karwowski
Professor and Chair
Department of Industrial & Management Systems
University of Central Florida
Orlando, Florida

Halimahtun M. Khalid
Executive Director
Principal Scientist and Consultant
Damai Sciences
Kuala Lampur, Malaysia

Thitima Kongnakorn
Senior Research Associate
Center for Health Economics & Science Policy
United BioSource Corp.
Lexington, Massachusetts

Danuta Koradecka
Professor and Director
Researcher
Central Institute for Labour Protection—National
    Research Institute
Warsaw, Poland

Vasilis Kouroumalis
Research Engineer
Institute of Computer Science
Foundation for Research and
    Technology—Hellas (FORTH)
Heraklion, Crete, Greece

Steven J. Landry
Assistant Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana

Kenneth R. Laughery, Sr.
Professor Emeritus
Department of Psychology
Rice University
Houston, Texas

K. Ronald Laughery, Jr.
President
Alion Science and Technology
Boulder, Colorodo

Christian Lebiere
Research Faculty
Department of Psychology
Carnegie Mellon University
Pittsburgh, Pennsylvania

John D. Lee
Professor
Industrial & Systems Engineering Department
University of Wisconsin—Madison
Madison, Wisconsin

Mark R. Lehto
Professor and Associate Head
School of Industrial Engineering
Purdue University
West Lafayette, Indiana

Asterios Leonides
Research Engineer
Institute of Computer Science
Foundation for Research and
    Technology—Hellas (FORTH)
Heraklion, Crete, Greece

James R. Lewis
Senior Human Factors Engineer
IBM Software Group
Boca Raton, Florida

Torsten Licht
RWTH Aachen University
Aachen, Germany

Holger Luczak
Professor Emeritus
Institute of Industrial Engineering and
    Ergonomics
Aachen, Germany

Katarzyna Majchrzycka
Researcher
Head, Department of Personal Protective
    Equipment
Central Institute for Labour Protection—National
    Research Institute
Warsaw, Poland

Krzysztof Makowski
Assistant Researcher
Laboratory for Respiratory Protective Equipment
Department of Personal Protective Equipment
Central Institute for Labour Protection—National
    Research Institute
Warsaw, Poland

Nicolas Marmaras
Professor
School of Mechanical Engineering
National Technical University of Athens
Athens, Greece

William S. Marras
Honda Chair Professor
Director, Biodynamics Laboratory
Director, COHAM
Executive Director, Institute for Ergonomics
Department of Integrated Systems Engineering
Ohio State University
Columbus, Ohio

Anna Marszalek
Researcher
Thermal Load Laboratory
Department of Ergonomics
Central Institute for Labour Protection—National
    Research Institute
Warsaw, Poland

Michael Matessa
Lead Cognitive Scientist
MA&D Operation
Alion Science and Technology
Boulder, Colorado

Christopher B. Mayhorn
Associate Professor
Department of Psychology
North Carolina State University
Raleigh, North Carolina

Wayne Maynard
Manager, Technical Services
Loss Control Advisory Services
Liberty Mutual Research Institute of Safety
Hopkinton, Massachusetts

Molly A. McClellan
Business Analyst
Office of Occupational Health and Safety
University of Minnesota
Minneapolis, Minnesota

Katrin Meinken
Dipl.-Biol.
Human Factors Engineering
Fraunhofer Institute for Industrial
    Engineering
Stuttgart, Germany

Frederick P. Morgeson
Professor and Valade Research Scholar
Department of Management
Michigan State University
East Lansing, Michigan

Allen L. Nagy
Professor of Psychology
Department of Psychology
Wright State University
Dayton, Ohio

Fiona Fui-Hoon Nah
Associate Professor of Management Information
    Systems
Department of Management
University of Nebraska—Lincoln
Lincoln, Nebraska

Sankaran N. Nair
Director, Research Data
Center on Aging
University of Miami Miller School of Medicine

Dimitris Nathanael
Lecturer
School of Mechanical Engineering
National Technical University of Athens
Athens, Greece

Timothy A. Nichols
User Research Engineer
Microsoft Game Studios
Microsoft Corporation
Redmond, Washington

Chris North
Associate Professor
Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

Grzegorz Owczarek
Lab Manager, Eye and Face Protection
    Laboratory
Researcher, Department of Personal Protective
    Equipment
Central Institute for Labour Protection—National
    Research Institute
Warsaw, Poland

A. Ant Ozok
Associate Professor
Department of Information Systems
University of Maryland Baltimore County
Baltimore, Maryland

Nikolaos Partarakis
Research Engineer
Institute of Computer Science
Foundation for Research and
    Technology—Hellas (FORTH)
Heraklion, Crete, Greece

Matthias Peissner
Head of Competence Center Human-Computer
    Interaction
Competence Center Human-Computer Interaction
Fraunhofer Institute for Industrial Engineering
Stuttgart, Germany

Tom Plocher
Staff Scientist
Automation and Control Solutions Laboratory
Honeywell International
Golden Valley, Minnesota

Beth M. Plott
Division Manager
Development Division
Alion Science & Technology
Boulder, Colorado

Janet D. Proctor
Lead Academic Advisor, Psychology
Liberal Arts
Purdue University
West Lafayette, Indiana

Robert W. Proctor
Distinguished University Professor
Department of Psychological Sciences
Purdue University
Wet Lafayette, Indiana

Pei-Luen Patrick Rau
Director, Human Factors Institute and
Professor, Department of Industrial Engineering
Tsinghua University
Beijing, China

Valerie J. Berg Rice
General Ergonomics
Selma, Texas

Michelle Robertson
Research Scientist
Center for Behavioral Science
Liberty Mutual Research Institute for Safety
Hopkinton, Massachusetts

Kathleen M. Robinette
Principal Research Anthropologist
Air Force Research Laboratory
711th Human Performance Wing/RHPA
Wright-Patterson Air Force Base, Ohio

David Rodrick
Assistant Professor
Department of Industrial & Manufacturing
    Systems Engineering
The University of Michigan—Dearborn
Dearborn, Michigan

Wendy A. Rogers
Professor
School of Psychology
Georgia Institute of Technology
Atlanta, Georgia

William B. Rouse
Executive Director & Professor
Tennenbaum Institute
Georgia Institute of Technology
Atlanta, Georgia

François Sainfort
Mayo Professor
Biomedical Health Informatics
University of Minnesota
Minneapolis, Minnesota

Eduardo Salas
Pegasus Professor & University
    Trustee Chair
Department of Psychology
Institute for Simulation & Training
University of Central Florida
Orlando, Florida

Anthony Savidis
Associate Professor, Department of Computer
    Science
University of Crete
Researcher
Institute of Computer Science
Foundation for Research and
    Technology—Hellas (FORTH)
Heraklion, Crete, Greece

Dylan Schmorrow
U.S. Navy
Arlington, Virginia

E. Eugene Schultz
Chief Technology Officer
Emagined Security
San Carlos, California

Richard Schweickert
Professor of Psychological Sciences
Department of Psychology
Purdue University
West Lafayette, Indiana

Bobbie D. Seppelt
Research Associate
Industrial & Systems Engineering Department
University of Wisconsin—Madison
Madison, Wisconsin

Joseph Sharit
Research Professor
Department of Industrial Engineering
University of Miami
Coral Gables, Florida

Walter E. Sipes
NASA Johnson Space Center
Houston, Texas

Bohdana Sherehiy
Independent Consultant
Washington, District of Columbia

Thomas B. Sheridan
Ford Professor of Engineering and Applied
    Psychology Emeritus
Department of Mechanical Engineering and
    Department of Aeronautics and Astronautics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Michael J. Smith
The Duane H. and Dorothy M. Bluemke
    Professor Emeritus
Department of Industrial and Systems
    Engineering
University of Wisconsin—Madison
Madison, Wisconsin

Dieter Spath
Professor and Head of Institute
Fraunhofer Institute for Industrial Engineering
Stuttgart, Germany

Sandra Sproll
Senior Scientist
Fraunhofer Institute for Industrial Engineering
Stuttgart, Germany

Kay M. Stanney
President
Design Interactive, Inc.
Oviedo, Florida

Constantine Stephanidis
Professor, Department of Computer Science
University of Crete
Director, Institute of Computer Science
Foundation for Research and
    Technology—Hellas (FORTH)
Heraklion, Crete, Greece

Alvaro D. Taveira
Professor and Chair
Department of Occupational and Environmental
    Safety and Health
University of Wisconsin—Whitewater
Whitewater, Wisconsin

Pamela S. Tsang
Professor
Department of Psychology
Wright State University
Dayton, Ohio

Gregg C. Vanderheiden
Professor, Department of Industrial & Systems
    Engineering and Department of Biomedical
    Engineering Director, Trace Center
University of Wisconsin—Madison
Madison, Wisconsin

Michael A. Vidulich
Principal Research Psychologist
Air Force Research Laboratory
Wright-Patterson Air Force Base, Ohio

Kim-Phuong L. Vu
Associate Professor
Department of Psychology
California State University—Long Beach
Long Beach, California

Christopher D. Wickens
Professor Emeritus, Aviation and Psychology
Institute of Aviation
University of Illinois
Senior Scientist
Alionscience
Savoy, Illinois

Glenn Wilson
Emeritus Principal Research Psychologist
Air Force Research Laboratory
Wright-Patteron Air Force Base, Ohio
Chief Scientist
Physiometrex
Keizer, Oregon

Michael S. Wogalter
Professor
Department of Psychology
North Carolina State University
Raleigh, North Carolina

Barbara Woolford
Manager, Space Human Factors and Habitability
    Research
NASA Johnson Space Center
Houston, Texas

Anping Xie
Research Assistant and Doctoral Student
Industrial & Systems Engineering
Center for Quality and Productivity Improvement
University of Wisconsin—Madison
Madison, Wisconsin

Ji Soo Yi
Assistant Professor
School of Industrial Engineering
Purdue University
West Lafayette, Indiana

Peter Young
Professor
Department of Electrical and Computer
    Engineering
Colorado State University
Fort Collins, Colorado

Panayiotis Zaphiris
Associate Professor
Department of Multimedia and Graphic Arts
Cyprus University of Technology
Limassol, Cyprus

Jan Zera
Lab Manager, Laboratory of Hearing Protection
Researcher, Department of Vibroacoustic Hazards
Central Institute for Labour Protection—National
    Research Institute
Warsaw, Poland

# PREFACE

This handbook is concerned with the role of humans in complex systems, the design of equipment and facilities for human use, and the development of environments for comfort and safety. The first, second, and third editions of the handbook were a major success and profoundly influenced the human factors profession. It was translated and published in Japanese and Russian and won the Institute of Industrial Engineers Joint Publishers Book of the Year Award. It has received strong endorsement from top management; the late Eliot Ester, retired president of General Motors Corporation, who wrote the forward to the first edition of the handbook, indicated that "regardless of what phase of the economy a person is involved in, this handbook is a very useful tool. Every area of human factors from environmental conditions and motivation to the use of new communication systems ... is well covered in the handbook by experts in every field."

In a literal sense, human factors and ergonomics is as old as the machine and environmental design, for it was aimed at designing them for human use. However, it was not until World War II that human factors emerged as a separate discipline. The field of human factors and ergonomics has developed and broadened considerably since its inception 70 years ago and has generated a body of knowledge in the following areas of specializations:

- Human factors profession
- Human factors fundamentals
- Design of tasks and jobs
- Equipment, workplace, and environmental design
- Design for health, safety, and comfort
- Performance modeling
- Evaluation
- Human–computer interaction
- Design for individual differences
- Selected applications

The foregoing list shows how broad the field has become. As such, this handbook should be of value to all human factors and ergonomics specialists, engineers, industrial hygienists, safety engineers, and human–computer interaction specialists. The 61 chapters constituting the fourth edition of the handbook were written by 131 experts. In creating this handbook, the authors gathered information from over 7500 references and presented over 500 figures and 200 tables to provide theoretically based and practically oriented material for use by both practitioners and researchers. In the fourth edition of the *Handbook of Human Factors and Ergonomics*, the chapters have been completely, newly written. This fourth edition

of the handbook covers totally new areas that were not included in the third edition. These include the following subjects:

- Managing low-back disorder risk in the workplace
- Neuroergonomics
- Social networking
- User requirements
- Human factors in ambient intelligent environments
- Online interactivity
- Office ergonomics
- Human factors and ergonomics in motor vehicle transportation
- Human factors and ergonomics in aviation

The main purpose of this handbook is to serve the needs of the human factors and ergonomics researchers, practitioners, and graduate students. Each chapter has a strong theory and science base and is heavily tilted toward application orientation. As such, a significant number of case studies, examples, figures, and tables are utilized to facilitate usability of the presented material.

The many contributing authors came through magnificently. I thank them all most sincerely for agreeing so willingly to create this handbook with me. I had the privilege of working with Robert L. Argentieri, our Wiley executive editor, who significantly facilitated my editorial work with his assistant Dan Magers. I was truly fortunate to have during the preparation of this handbook the most able contribution of Myrna Kasdorf, editorial coordinator of the handbook, who has done a truly outstanding job.

GAVRIEL SALVENDY
*January 2011*

# PART 1
# HUMAN FACTORS FUNCTION

# CHAPTER 1

# THE DISCIPLINE OF HUMAN FACTORS AND ERGONOMICS

**Waldemar Karwowski**
**University of Central Florida**
**Orlando, Florida**

*The purpose of science is mastery over nature.*

F. Bacon (*Novum Organum,* 1620)

## 1 INTRODUCTION

Over the last 60 years human factors, a term that is used here synonymously with ergonomics [and denoted as human factors ergonomics (HFE)], has been evolving as a unique and independent discipline that focuses on the nature of human–artifact interactions, viewed from the unified perspective of the science, engineering, design, technology, and management of human-compatible systems, including a variety of natural and artificial products, processes, and living environments (Karwowski, 2005). The various dimensions of such defined ergonomics discipline are shown in Figure 1. The International Ergonomics Association (IEA, 2003) defines ergonomics (or human factors) as the scientific discipline concerned with the understanding of the interactions among humans and other elements of a system and the profession that applies theory, principles, data, and methods to design in order to optimize human well-being and overall system performance. Human factors professionals contribute to the design

and evaluation of tasks, jobs, products, environments, and systems in order to make them compatible with the needs, abilities, and limitations of people. Ergonomics discipline promotes a holistic, human-centered approach to work systems design that considers the physical, cognitive, social, organizational, environmental, and other relevant factors (Grandjean, 1986; Wilson and Corlett, 1995; Sanders and McCormick, 1993; Chapanis, 1995, 1999; Salvendy, 1997; Karwowski, 2001; Vicente, 2004; Stanton et al., 2004).

Historically, *ergonomics* (*ergon + nomos*), or "the study of work," was originally and proposed and defined by the Polish scientist B. W. Jastrzebowski (1857a-d) as the scientific discipline with a very broad scope and wide subject of interests and applications, encompassing all aspects of human activity, including labor, entertainment, reasoning, and dedication (Karwowski (1991, 2001). In his paper published in the journal *Nature and Industry* (1857), Jastrzebowski divided work into two main categories: the *useful work*, which brings improvement for the common good, and the *harmful work* that
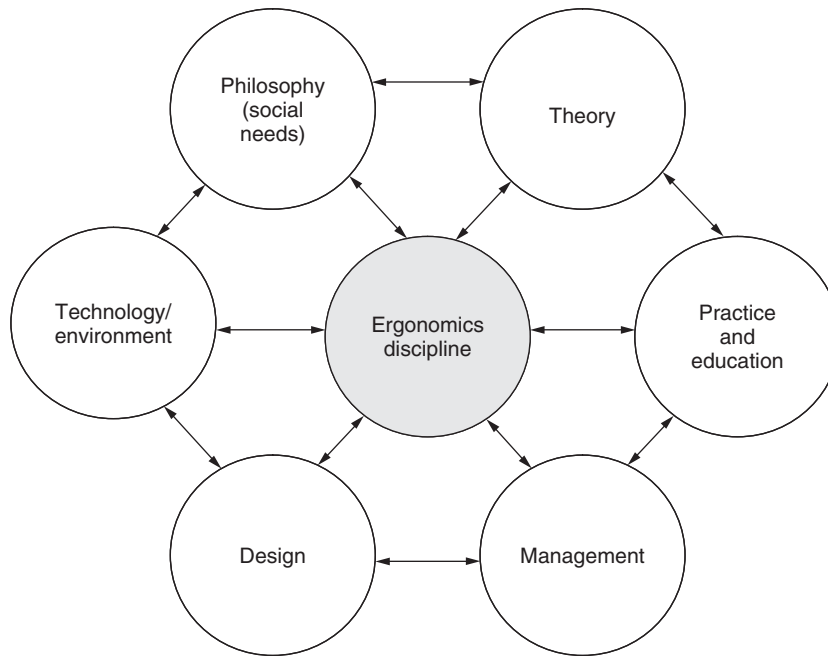
**Figure 1**   General dimensions of ergonomics discipline (after Karwowski, 2005).

brings deterioration (discreditable work). Useful work, which aims to improve things and people, is classified into physical, aesthetic, rational, and moral work. According to Jastrzebowski, such work requires utilization of the motor forces, sensory forces, forces of reason (thinking and reasoning), and the spiritual force. The four main benefits of the useful work are exemplified through the property, ability, perfection, and felicity.

The contemporary ergonomics discipline, independently introduced by Murrell in 1949 (Edholm and Murrell, 1973), was viewed at that time as an applied science, the technology, and sometimes both. The British scientists had founded the Ergonomics Research Society in 1949. The development of ergonomics internationally can be linked to a project initiated by the *European Productivity Agency* (EPA), a branch of the *Organization for European Economic Cooperation*, which first established a Human Factors Section in 1955 (Kuorinka, 2000). Under the EPA project, in 1956 specialists from European countries visited the United States to observe human factors research. In 1957 the EPA organized a technical seminar on "Fitting the Job to the Worker" at the University of Leiden, The Netherlands, during which a set of proposals was presented to form an international association of work scientists. A steering committee consisting of H.S. Belding, G.C.E. Burger, S. Forssman, E. Grandjean, G. Lehman, B. Metz, K.U. Smith, and R.G. Stansfield, was charged to develop specific proposal for such association. The committee decided to adopt the name International Ergonomics Association. At the meeting in Paris in 1958 it was decided to proceed with forming the new association. The steering committee designated itself

as the Committee for the International Association of Ergonomic Scientists and elected G.C.E. Burger as its first president, K.U. Smith as treasurer, and E. Grandjean as secretary. The Committee for the International Association of Ergonomic Scientists met in Zurich in 1959 during a conference organized by the EPA and decided to retain the name International Ergonomics Association. On April 6, 1959, at the meeting in Oxford, England, E. Grandjean declared the founding of the IEA. The committee met again in Oxford, England, later in 1959 and agreed upon the set of bylaws or statutes of the IEA. These were formally approved by the IEA General Assembly at the first International Congress of Ergonomics held in Stockholm in 1961.

Traditionally, the most often cited domains of specialization within HFE are the physical, cognitive, and organizational ergonomics. Physical ergonomics is mainly concerned with human anatomical, anthropometric, physiological, and biomechanical characteristics as they relate to physical activity [Chaffin et al., 2006, Pheasant, 1986; Kroemer et al., 1994; Karwowski and Marras, 1999; National Research Council (NRC), 2001; Marras, 2008]. Cognitive ergonomics focuses on mental processes such as perception, memory, information processing, reasoning, and motor response as they affect interactions among humans and other elements of a system (Vicente, 1999; Hollnagel, 2003; Diaper and Stanton, 2004). Organizational ergonomics (also known as macroergonomics) is concerned with the optimization of sociotechnical systems, including their organizational structures, policies, and processes (Reason, 1997; Hendrick and Kleiner, 2002a,b; Hollman et al., 2003; Nemeth, 2004). Examples of the relevant topics include

**Table 1  Exemplary Domains of Disciplines of Medicine, Psychology, and Human Factors**

| Medicine | Psychology | Human factors |
| --- | --- | --- |
| Cardiology | Applied psychology | Physical ergonomics |
| Dermatology | Child psychology | Cognitive ergonomics |
| Gastroenterology | Clinical psychology | Macroergonomics |
| Neurology | Cognitive psychology | Knowledge ergonomics |
| Radiology | Community psychology | Rehabilitation ergonomics |
| Endocrinology | Counseling psychology | Participatory ergonomics |
| Pulmonology | Developmental psychology | Human–computer interaction |
| Gerontology | Experimental psychology | Neuroergonomics |
| Neuroscience | Educational psychology | Affective ergonomics |
| Nephrology | Environmental psychology | Ecological ergonomics |
| Oncology | Forensic psychology | Forensic ergonomics |
| Ophthalmology | Health psychology | Consumer ergonomics |
| Urology | Positive psychology | Human–systems integration |
| Psychiatry | Organizational psychology | Ergonomics of aging |
| Internal medicine | Social psychology | Information ergonomics |
| Community medicine | Quantitative psychology | Community ergonomics |
| Physical medicine | Social psychology | Nanoergonomics |
| | | Service ergonomics |

communication, crew resource management, design of working times, teamwork, participatory work design, community ergonomics, computer-supported cooperative work, new work paradigms, virtual organizations, telework, and quality management. The above traditional domains as well as new domains are listed in Table 1. According to the above discussion, the paramount objective of HFE is to understand the interactions between people and everything that surrounds us and based on such knowledge to optimize the human well-being and overall system performance. Table 2 provides a summary of the specific HFE objectives as discussed by Chapanis (1995). As recently pointed out by the National Academy of Engineering (NAE, 2004), in the future, ongoing developments in engineering will *expand toward tighter connections between technology and the human experience, including new products customized to the physical dimensions and capabilities of the user, and ergonomic design of engineered products*.

## 2  HUMAN–SYSTEM INTERACTIONS

While in the past ergonomics has been driven by technology (reactive design approach), in the future ergonomics should drive technology (proactive design approach). While technology is a product and a process involving both science and engineering, science aims to understand the "why" and "how" of nature through a process of scientific inquiry that generates knowledge about the natural world (Mitchem, 1994; NRC 2001). Engineering is the "design under constraints" of cost, reliability, safety, environmental impact, ease of use, available human and material resources, manufacturability, government regulations, laws, and politics (Wulf, 1998). Engineering, as a body of knowledge of design and creation of human-made products and a process for

**Table 2  Objectives of HFE Discipline**

Basic Operational Objectives
  Reduce errors
  Increase safety
  Improve system performance
Objectives Bearing on Reliability, Maintainability, and
  Availability (RMA) and Integrated Logistic Support (ILS)
  Increase reliability
  Improve maintainability
  Reduce personnel requirements
  Reduce training requirements
Objectives Affecting Users and Operators
  Improve the working environment
  Reduce fatigue and physical stress
  Increase ease of use
  Increase user acceptance
  Increase aesthetic appearance
Other Objectives
  Reduce losses of time and equipment
  Increase economy of production

*Source*: Chapanis (1995).

solving problems, seeks to shape the natural world to meet human needs and wants.

Contemporary HFE discovers and applies information about human behavior, abilities, limitations, and other characteristics to the design of tools, machines, systems, tasks, jobs, and environments for productive, safe, comfortable, and effective human use (Sanders and Mccormick, 1993; Helander, 1997). In this context, HFE deals with a broad scope of problems relevant to the design and evaluation of work systems, consumer

products, and working environments, in which human−machine interactions affect human performance and product usability (Carayon, 2006; Dekker, (2007; Karwowski, 2006; Bedny and Karwowski, 2007; Weick and Sutcliffe, 2007; Sears and Jacko, 2009; Wogalter, 2006; Reason, 2008; Bisantz and Burns, 2009; Karwowski et al., 2010). The wide scope of issues addressed by the contemporary HFE discipline is presented in Table 3. Figure 2 illustrates the evolution of the scope of HFE with respect to the nature of human−system interactions and applications of human− system integration in a large variety of domains (Vicente, 2004; Karwowski, 2007; Lehto and Buck, 2007; Marras and Karwowski 2006a,b; Rouse, 2007; Guerin et al., 2007; Dekker, 2007; Schmorrow and Stanney, 2008; Pew and Mavor, 2008.; Cook and Durso, 2008; Zacharias et al., 2008; Salas et al., 2008; Marras, 2008, Chebbykin et al., 2008; Salvendy and Karwowski, 2010; Kaber and Boy, 2010; Marek et al., 2010).

Originally, HFE focused on the local human− machine interactions, while today the main focus is on the broadly defined human−technology interactions. In

**Table 3  Classification Scheme for Human Factors/Ergonomics**

1. General

### Human Characteristics

2. Psychological aspects
3. Physiological and anatomical aspects
4. Group factors
5. Individual differences
6. Psychophysiological state variables
7. Task-related factors

### Information Presentation and Communication

8. Visual communication
9. Auditory and other communication modalities
10. Choice of communication media
11. Person−machine dialogue mode
12. System feedback
13. Error prevention and recovery
14. Design of documents and procedures
15. User control features
16. Language design
17. Database organization and data retrieval
18. Programming, debugging, editing, and programming aids
19. Software performance and evaluation
20. Software design, maintenance, and reliability

### Display and Control Design

21. Input devices and controls
22. Visual displays
23. Auditory displays
24. Other modality displays
25. Display and control characteristics

**Table 3  (continued)**

### Workplace and Equipment Design

26. General workplace design and buildings
27. Workstation design
28. Equipment design

### Environment

29. Illumination
30. Noise
31. Vibration
32. Whole-body movement
33. Climate
34. Altitude, depth, and space
35. Other environmental issues

### System Characteristics

36. General system features

### Work Design and Organization

37. Total system design and evaluation
38. Hours of work
39. Job attitudes and job satisfaction
40. Job design
41. Payment systems
42. Selection and screening
43. Training
44. Supervision
45. Use of support
46. Technological and ergonomic change

### Health and Safety

47. General health and safety
48. Etiology
49. Injuries and illnesses
50. Prevention

### Social and Economic Iimpact of the System

51. Trade unions
52. Employment, job security, and job sharing
53. Productivity
54. Women and work
55. Organizational design
56. Education
57. Law
58. Privacy
59. Family and home life
60. Quality of working life
61. Political comment and ethical considerations

### Methods and Techniques

62. Approaches and methods
63. Techniques
64. Measures

*Source*: Ergonomics Abstracts (2004).

```
┌─────────────────────────────────────┐
│   Human–technology relationships    │
└─────────────────────────────────────┘
                   ▲
                   ▼
┌─────────────────────────────────────┐
│   Technology–system relationships   │
└─────────────────────────────────────┘
                   ▲
                   ▼
┌─────────────────────────────────────┐
│     Human–system relationships      │
└─────────────────────────────────────┘
                   ▲
                   ▼
┌─────────────────────────────────────┐
│     Human–machine relationships     │
└─────────────────────────────────────┘
```
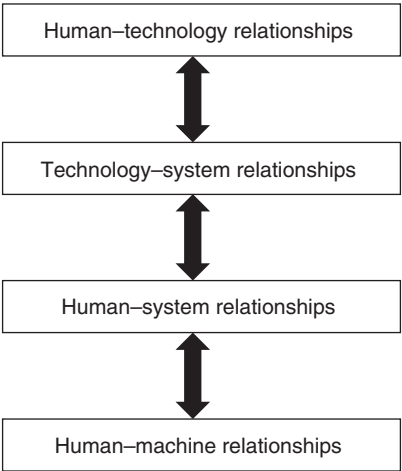
**Figure 2** Expanded view of the human–technology relationships (modified after Meister, 1999).

this view, the HFE can also be called the discipline of *technological ecology*. Tables 4 and 5 present taxonomy of the human-related and technology-related components, respectively, which are of great importance to HFE discipline. According to Meister (1987), the traditional concept of the human–machine system is an organization of people and the machines they operate and maintain in order to perform assigned jobs that implement the purpose for which the system was developed (Meister, 1987). In this context, a system

is a *construct* whose characteristics are manifested in physical and behavioral phenomena Meister (1991). The system is critical to HFE theorizing because it describes the substance of the human–technology relationship. General system variables of interest to HFE discipline are shown in Table 6.

The human functioning in human–machine systems can be described in terms of perception, information processing, decision making, memory, attention, feedback, and human response processes. Furthermore, the human work taxonomy can be used to describe five distinct levels of human functioning, ranging from primarily physical tasks to cognitive tasks (Karwowski, 1992a). These basic but universal human activities are (1) tasks that produce force (primarily muscular work),

**Table 5  Taxonomy of HFE Elements: Technology**

| Technology Elements | Effects of Technology on the Human |
|---|---|
| Components | Changes in human role |
| Tools | Changes in human behavior |
| Equipments | |
| Systems | |
| Degree of Automation | Organization–Technology Relationships |
| Mechanization | Definition of organization |
| Computerization | Organizational variables |
| Artificial intelligence | |
| System Characteristics | |
| Dimensions | |
| Attributes | |
| Variables | |

*Source*: Meister (1999).

**Table 4  Taxonomy of HFE Elements: The Human Factor**

| | |
|---|---|
| Human Elements | Effects of the Human on Technology |
|   Physical/sensory |   Improvement in technology effectiveness |
|   Cognitive |   Absence of effect |
|   Motivational/emotional |   Reduction in technological effectiveness |
| Human Conceptualization | Human Operations in Technology |
|   Stimulus–response orientation (limited) |   Equipment operation |
|   Stimulus–conceptual–response orientation (major) |   Equipment maintenance |
|   Stimulus–conceptual–motivational–response orientation (major) |   System management |
| Human Technological Relationships |   Type/degree of human involvement |
|   Controller relationship |     Direct (operation) |
|   Partnership relationship |     Indirect (recipient) |
|   Client relationship |     Extensive |
| Effects of Technology on the Human |     Minimal |
|   Performance effects |     None |
|     Goal accomplishment | |
|     Goal nonaccomplishment | |
|     Error/time discrepancies | |
|     Feeling effect | |
|     Technology acceptance | |
|     Technology indifference | |
|     Technology rejection | |
|     Demand effects | |
|     Resource mobilization | |
|     Stress/trauma | |

*Source*: Meister (1999).

**Table 6  General System Variables**

1. Requirement constraints imposed on the system
2. Resources required by the system
3. Nature of its internal components and processes
4. Functions and missions performed by the system
5. Nature, number, and specificity of goals
6. Structural and organizational characteristics of the system (e.g., its size, number of subsystems and units, communication channels, hierarchical levels, and amount of feedback)
7. Degree of automation
8. Nature of the environment in which the system functions
9. System attributes (e.g., complexity, sensitivity, flexibility, vulnerability, reliability, and determinacy)
10. Number and type of interdependencies (human–machine interactions) within the system and type of interaction (degree of dependency)
11. Nature of the system's terminal output(s) or mission effects

*Source*: Meister (1999).

(2) tasks of continuously coordinating sensorimonitor functions (e.g., assembling or tracking tasks), (3) tasks of converting information into motor actions (e.g., inspection tasks), (4) tasks of converting information into output information (e.g., required control tasks), and (5) tasks of producing information (primarily creative work) (Grandjean, 1986; Luczak et al., 1999). Any task in a human–machine system requires processing of information that is gathered based on perceived and interpreted relationships between system elements. The processed information may need to be stored by either a human or a machine for later use.

The scope of HFE factors that need to be considered in the design, testing, and evaluation of any human–system interactions is shown in Table 7 in the form of the exemplary ergonomics checklist. It should be noted that such checklists also reflect practical application of the discipline. According to the Board of Certification in Professional Ergonomics (BCPE), a practitioner of ergonomics is a person who (1) has a mastery of a body of ergonomics knowledge, (2) has a command of the methodologies used by ergonomists in applying that knowledge to the design of a product, system, job, or environment, and (3) has applied his or her knowledge to the analysis, design testing, and evaluation of products, systems, and environments. The areas of current practice in the field can be best described by examining the focus of Technical Groups of the Human Factors and Ergonomics Society, as illustrated in Table 8.

**Table 7  Examples of Factors to Be Used in Ergonomics Checklists**

*I. Anthropometric, Biomechanical, and Physiological Factors*

1. Are the differences in human body size accounted for by the design?
2. Have the right anthropometric tables been used for specific populations?
3. Are the body joints close to neutral positions?
4. Is the manual work performed close to the body?
5. Are there any forward-bending or twisted trunk postures involved?
6. Are sudden movements and force exertion present?
7. Is there a variation in worker postures and movements?
8. Is the duration of any continuous muscular effort limited?
9. Are the breaks of sufficient length and spread over the duration of the task?
10. Is the energy consumption for each manual task limited?

*II. Factors Related to Posture (Sitting and Standing)*

1. Is sitting/standing alternated with standing/sitting and walking?
2. Is the work height dependent on the task?
3. Is the height of the work table adjustable?
4. Are the height of the seat and backrest of the chair adjustable?
5. Is the number of chair adjustment possibilities limited?
6. Have good seating instructions been provided?
7. Is a footrest used where the work height is fixed?
8. Has the work above the shoulder or with hands behind the body been avoided?
9. Are excessive reaches avoided?
10. Is there enough room for the legs and feet?
11. Is there a sloping work surface for reading tasks?
12. Have the combined sit–stand workplaces been introduced?
13. Are handles of tools bent to allow for working with the straight wrists?

**Table 7 (*continued*)**

### III. Factors Related to Manual Materials Handling (Lifting, Carrying, Pushing, and Pulling Loads)

1. Have tasks involving manual displacement of loads been limited?
2. Have optimum lifting conditions been achieved?
3. Is anybody required to lift more than 23 kg?
4. Have lifting tasks been assessed using the NIOSH (1991) method?
5. Are handgrips fitted to the loads to be lifted?
6. Is more than one person involved in lifting or carrying tasks?
7. Are there mechanical aids for lifting or carrying available and used?
8. Is the weight of the load carried limited according to the recognized guidelines?
9. Is the load held as close to the body as possible?
10. Are pulling and pushing forces limited?
11. Are trolleys fitted with appropriate handles and handgrips?

### IV. Factors Related to Design of Tasks and Jobs

1. Does the job consist of more than one task?
2. Has a decision been made about allocating tasks between people and machines?
3. Do workers performing the tasks contribute to problem solving?
4. Are the difficult and easy tasks performed interchangeably?
5. Can workers decide independently on how the tasks are carried out?
6. Are there sufficient possibilities for communication between workers?
7. Is there sufficient information provided to control the assigned tasks?
8. Can the group take part in management decisions?
9. Are the shift workers given enough opportunities to recover?

### V. Factors Related to Information and Control Tasks

*Information*

1. Has an appropriate method of displaying information been selected?
2. Is the information presentation as simple as possible?
3. Has the potential confusion between characters been avoided?
4. Has the correct character/letter size been chosen?
5. Have texts with capital letters only been avoided?
6. Have familiar typefaces been chosen?
7. Is the text/background contrast good?
8. Are the diagrams easy to understand?
9. Have the pictograms been properly used?
10. Are sound signals reserved for warning purposes?

*Control*

1. Is the sense of touch used for feedback from controls?
2. Are differences between controls distinguishable by touch?
3. Is the location of controls consistent and is sufficient spacing provided?
4. Have the requirements for the control-display compatibility been considered?
5. Is the type of cursor control suitable for the intended task?
6. Is the direction of control movements consistent with human expectations?
7. Are the control objectives clear from the position of the controls?
8. Are controls within easy reach of female workers?
9. Are labels or symbols identifying controls properly used?
10. Is the use of color in controls design limited?

**Table 7  (*continued*)**

*Human–Computer Interaction*

1. Is the human–computer dialogue suitable for the intended task?
2. Is the dialogue self-descriptive and easy to control by the user?
3. Does the dialogue conform to the expectations on the part of the user?
4. Is the dialogue error tolerant and suitable for user learning?
5. Has command language been restricted to experienced users?
6. Have detailed menus been used for users with little knowledge and experience?
7. Is the type of help menu fitted to the level of the user's ability?
8. Has the QWERTY layout been selected for the keyboard?
9. Has a logical layout been chosen for the numerical keypad?
10. Is the number of function keys limited?
11. Have the limitations of speech in human–computer dialogue been considered?
12. Are touch screens used to facilitate operation by inexperienced users?

### *VI. Environmental Factors*

*Noise and Vibration*

1. Is the noise level at work below 80 dBA?
2. Is there an adequate separation between workers and source of noise?
3. Is the ceiling used for noise absorption?
4. Are the acoustic screens used?
5. Are hearing conservation measures fitted to the user?
6. Is personal monitoring to noise/vibration used?
7. Are the sources of uncomfortable and damaging body vibration recognized?
8. Is the vibration problem being solved at the source?
9. Are machines regularly maintained?
10. Is the transmission of vibration prevented?

*Illumination*

1. Is the light intensity for normal activities in the range of 200–800 lux?
2. Are large brightness differences in the visual field avoided?
3. Are the brightness differences between task area, close surroundings, and wider surroundings limited?
4. Is the information easily legible?
5. Is ambient lighting combined with localized lighting?
6. Are light sources properly screened?
7. Can the light reflections, shadows, or flicker from the fluorescent tubes be prevented?

*Climate*

1. Are workers able to control the climate themselves?
2. Is the air temperature suited to the physical demands of the task?
3. Is the air prevented from becoming either too dry to too humid?
4. Are draughts prevented?
5. Are the materials/surfaces that have to be touched neither too cold nor too hot?
6. Are the physical demands of the task adjusted to the external climate?
7. Are undesirable hot and cold radiation prevented?
8. Is the time spent in hot or cold environments limited?

*Source*: Dul and Weerdmeester (1993).

**Table 8  Subject Interests of Technical Groups of Human Factors and Ergonomics Society**

| Technical Group | Description/Areas of Concerns |
| --- | --- |
| Aerospace systems | Application of human factors to the development, design, certification, operation, and maintenance of human–machine systems in aviation and space environments. The group addresses issues for civilian and military systems in the realms of performance and safety. |
| Aging | Human factors applications appropriate to meeting the emerging needs of older people and special populations in a wide variety of life settings. |
| Augmented cognition | Fostering the development and application of real-time physiological and neurophysiological sensing technologies that can ascertain a human's cognitive state while interacting with computing-based systems; data classification and integration architectures that enable closed-loop system applications; mitigation (adaptive) strategies that enable efficient and effective system adaptation based on a user's dynamically changing cognitive state; individually tailored training systems. |
| Cognitive engineering and decision making | Research on human cognition and decision making and the application of this knowledge to the design of systems and training programs. Emphasis is on considerations of descriptive models, processes, and characteristics of human decision making, alone or in conjunction with other individuals or intelligent systems; factors that affect decision making and cognition in naturalistic task settings; technologies for assisting, modifying, or supplementing human decision making; and training strategies for assisting or influencing decision making. |
| Communications | All aspects of human-to-human communication, with an emphasis on communication mediated by telecommunications technology, including multimedia and collaborative communications, information services, and interactive broadband applications. Design and evaluation of both enabling technologies and infrastructure technologies in education, medicine, business productivity, and personal quality of life. |
| Computer systems | Human factors in the design of computer systems. This includes the user-centered design of hardware, software, applications, documentation, work activities, and the work environment. Practitioners and researchers in the CSTG community take a holistic, systems approach to the design and evaluation of all aspects of user–computer interactions. Some goals are to ensure that computer systems are useful, usable, safe, and, where possible, fun and to enhance the quality of work life and recreational/educational computer use by ensuring that computer interface, function, and job design are interesting and provide opportunities for personal and professional growth. |
| Consumer products | Development of consumer products that are useful, usable, safe, and desirable. Application of the principles and methods of human factors, consumer research, and industrial design to ensure market success. |
| Education | Education and training of human factors and ergonomics specialists. This includes undergraduate, graduate, and continuing education needs, issues, techniques, curricula, and resources. In addition, a forum is provided to discuss and resolve issues involving professional registration and accreditation |
| Environmental design | Relationship between human behavior and the designed environment. Common areas of research and interest include ergonomic and macroergonomic aspects of design within home, office, and industrial settings. An overall objective of this group is to foster and encourage the integration of ergonomics principles into the design of environments |
| Forensics professional | Application of human factors knowledge and technique to ''standards of care'' and accountability established within legislative, regulatory, and judicial systems. The emphasis on providing a scientific basis to issues being interpreted by legal theory. |
| Health care | Maximizing the contributions of human factors and ergonomics to medical systems effectiveness and the quality of life of people who are functionally impaired |
| Individual differences | A wide range of personality and individual difference variables that are believed to mediate performance. |
| Industrial ergonomics | Application of ergonomics data and principles for improving safety, productivity, and quality of work in industry. Concentration on service and manufacturing processes, operations, and environments. |

**Table 8  (*continued*)**

| Technical Group | Description/Areas of Concerns |
|---|---|
| Internet | Human factor aspects of user interface design of Web content, Web-based applications, Web browsers, Webtops, Web-based user assistance, and Internet devices; behavioral and sociological phenomena associated with distributed network communication; human reliability in administration and maintenance of data networks; and accessibility of Web-based products. |
| Macroergonomics | Organizational design and management issues in human factors and ergonomics as well as work system design and human–organization interface technology. The Technical Group is committed to improving work system performance (e.g., productivity, quality, health and safety, quality of work life) by promoting work system analysis and design practice and the supporting empirical science concerned with the technological subsystem, personnel subsystem, external environment, organizational design, and their interactions. |
| Perception and performance | Perception and its relation to human performance. Areas of concern include the nature, content, and quantification of sensory information and the context in which it is displayed; the physics and psychophysics of information display; perceptual and cognitive representation and interpretation of displayed information; assessment of workload using tasks having a significant perceptual component; and actions and behaviors that are consequences of information presented to the various sensory systems. |
| Product design | Developing consumer products that are useful, usable, safe, and desirable. By applying the principles and methods of human factors, consumer research, and industrial design, the group works to ensure the success of products sold in the marketplace |
| Safety | Development and application of human factors technology as it relates to safety in all settings and attendant populations. These include, but are not limited to, aviation, transportation, industry, military, office, public building, recreation, and home environment |
| System development | Fostering research and exchanging information on the integration of human factors and ergonomics into the development of systems. Members are concerned with defining human factors/ergonomics activities and integrating them into the system development process in order to enable systems that meet user requirements. Specific topics of interest include the system development process itself; developing tools and methods for predicting and assessing human capabilities and limitations, notably modeling and simulation; creating principles that identify the role of humans in the use, operation, maintenance, and control of systems; applying human factors and ergonomics data and principles to the design of human–system interfaces; and the full integration of human requirements into system and product design through the application of HSI methods to ensure technical and programmatic integration of human considerations into systems acquisition and product development processes; the impact of increasing computerization and stress and workload effects on performance. |
| Surface transportation | Human factors related to the international surface transportation field. *Surface transportation* encompasses numerous mechanisms for conveying humans and resources: passenger, commercial, and military vehicles, on- and off-road; mass transit; maritime transportation; rail transit, including vessel traffic services (VTSs); pedestrian and bicycle traffic; and highway and infrastructure systems, including intelligent transportation systems (ITSs). |
| Test and evaluation | All aspects of human factors and ergonomics as applied to the evaluation of systems. Evaluation is a core skill for all human factors professionals and includes measuring performance, workload, situational awareness, safety, and acceptance of personnel engaged in operating and maintaining systems. Evaluation is conducted during system development when prototype equipment and systems are being introduced to operational usage and at intervals thereafter during the operational life of these systems. |
| Training | Fosters information and interchange among people interested in the fields of training and training research. |
| Virtual environment | Human factors issues associated with human–virtual environment interaction. These issues include maximizing human performance efficiency in virtual environments, ensuring health and safety, and circumventing potential social problems through proactive assessment. For VE/VR systems to be effective and well received by their users, researchers need to focus significant efforts on addressing human factors issues. |

*Source*: www.hfes.org.

## 3  HFE AND ECOLOGICAL COMPATIBILITY

The HFE discipline advocates systematic use of the knowledge concerning relevant human characteristics in order to achieve compatibility in the design of interactive systems of people, machines, environments, and devices of all kinds to ensure specific goals [Human Factors and Ergonomics Society (HFES), 2003)]. Typically such goals include improved (system) effectiveness, productivity, safety, ease of performance, and the contribution to overall human well-being and quality of life. Although the term *compatibility* is a key word in the above definition, it has been mainly used in a narrow sense only, often in the context of the design of displays and controls, including the studies of spatial (location) compatibility or the intention–response–stimulus compatibility related to movement of controls (Wickens and Carswell, 1997). Karwowski and his co-workers (Karwowski et al., 1988; Karwowski, 1985, 1991) advocated the use of compatibility in a greater context of the ergonomics system. For example, Karwowski (1997) introduced the term *human-compatible systems* in order to focus on the need for comprehensive treatment of compatibility in the human factors discipline.

The *American Heritage Dictionary of English Language* (Morris, 1978) defines "compatible" as (1) capable of living or performing in harmonious, agreeable, or congenial combination with another or others and (2) capable of orderly, efficient integration and operation with other elements in a system. From the beginning of contemporary ergonomics, the measurements of compatibility between the system and the human and evaluation of the results of ergonomics interventions were

based on the measures that best suited specific purposes (Karwowski, 2001). Such measures included the specific psychophysiological responses of the human body (example.g., heart rate, EMG, perceived human exertion, satisfaction, comfort or discomfort) as well as a number of indirect measures, such as the incidence of injury, economic losses or gains, system acceptance, or operational effectiveness, quality, or productivity. The lack of a universal matrix to quantify and measure human–system compatibility is an important obstacle in demonstrating the value of ergonomics science and profession (Karwowski, 1997). However, even though 20 years ago ergonomics was perceived by some (e.g., see Howell, 1986) as a highly unpredictable area of human scientific endeavor, today HFE has positioned itself as a unique, *design-oriented* discipline, independent of engineering and medicine (Moray, 1984; Sanders and McCormick, 1993; Helander, 1997; Karwowski, 1991, 2003).

Figure 3 illustrates the human–system compatibility approach to ergonomics in the context of quality of working life and system (an enterprise or business entity) performance. This approach reflects the nature of complex compatibility relationships between the human operator (human capacities and limitations), technology (in terms of products, machines, devices, processes, and computer-based systems), and the broadly defined environment (business processes, organizational structure, nature of work systems, and effects of work-related multiple stressors). The operator's performance is an outcome of the compatibility matching between individual human characteristics (capacities and limitations) and the requirements and affordances of both the technology
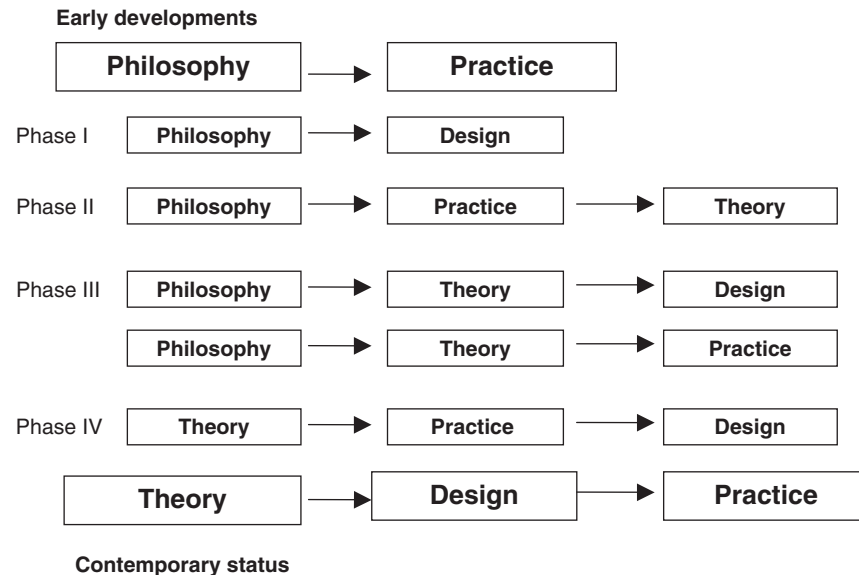


Figure 3   Evolution in development of HFE discipline (after Karwowski, 2005).

and environment. The quality of working life and the system (enterprise) performance is affected by matching the positive and negative outcomes of the complex compatibility relationships between the human operator, technology, and environment. Positive outcomes include such measures as work productivity, performance times, product quality, and subjective psychological (desirable) behavioral outcomes such as job satisfaction, employee morale, human well-being, and commitment. The negative outcomes include both human and system-related errors, loss of productivity, low quality, accidents, injuries, physiological stresses, and subjective psychological (undesirable) behavioral outcomes such as job dissatisfaction, job/occupational stress, and discomfort.

## 4  DISTINGUISHING FEATURES OF CONTEMPORARY HFE DISCIPLINE AND PROFESSION

The main focus of the HFE discipline in the twenty-first century will be the design and management of systems that satisfy customer demands in terms of human compatibility requirements. Karwowski (2005) has discussed 10 characteristics of contemporary HFE discipline and profession. Some of the distinguishing features are as follows:

- HFE experiences continuing evolution of its "fit" philosophy, including diverse and ever-expanding human-centered design criteria (from safety to comfort, productivity, usability, or affective needs like job satisfaction or life happiness).
- HFE covers extremely diverse subject matters, similarly to medicine, engineering, and psychology (see Table 1).
- HFE deals with very complex phenomena that are not easily understood and cannot be simplified by making nondefendable assumptions about their nature.
- Historically, HFE has been developing from the "philosophy of fit" toward practice. Today, HFE is developing a sound theoretical basis for design and practical applications (see Figure 4).
- HFE attempts to "by-step" the need for the fundamental understanding of human−system interactions without separation from the consideration of knowledge utility for practical applications in the quest for immediate and useful solutions (also see Figure 5).
- HFE has limited recognition by decisionmakers, the general public, and politicians as to its value that it can bring to a global society at large, especially in the context of facilitating the socioeconomic development.
- HFE has a relatively limited professional educational base.
- The impact of HFE is affected by the ergonomics illiteracy of the students and professionals in other disciplines, the mass media, and the public at large.

Theoretical ergonomics is interested in the fundamental understanding of the interactions between people and their environments. Central to HFE interests is also an understanding of how human−system interactions should be designed. On the other hand, HFE also falls under the category of applied research. The taxonomy of research efforts with respect to the quest for a fundamental understanding and the consideration of use, originally proposed by Stokes (1997), allows for differentiation of the main categories of research dimensions as follows: (1) pure basic research, (2) use-inspired basic research, and (3) pure applied research. Figure 5 illustrates the interpretation of these categories for the HFE theory, design, and applications. Table 9 presents relevant specialties and subspecialties in HFE research as outlined by Meister (1999), who classified them into three main categories: (1) system/technology-oriented specialties, (2) process-oriented specialties, and (3) behaviorally oriented specialties. In addition, Table 10 presents a list of contemporary HFE research methods that can be used to advance the knowledge discovery and utilization through its practical applications.

## 5  PARADIGMS FOR ERGONOMICS DISCIPLINE

The paradigms for any scientific discipline include theory, abstraction, and design (Pearson and Young, 2002). Theory is a foundation of the mathematical sciences. Abstraction (modeling) is a foundation of the natural sciences, where progress is achieved by formulating hypotheses and systematically following the modeling process to verify and validate them. Design is the basis for engineering, where progress is achieved primarily by posing problems and systematically following the design process to construct systems that solve them.

In view of the above, Karwowski (2005) discussed the paradigms for HFE discipline: (1) ergonomics theory, (2) ergonomics abstraction, and (3) ergonomics design. Ergonomics theory is concerned with the ability to identify, describe, and evaluate human−system interactions. Ergonomics abstraction is concerned with the ability to use those interactions to make predictions that can be compared with the real world. Ergonomics design is concerned with the ability to implement knowledge about those interactions and use them to develop systems that satisfy customer needs and relevant human compatibility requirements. Furthermore, the pillars for any scientific discipline include a definition, a teaching paradigm, and an educational base (NRC, 2001). A definition of the ergonomics discipline and profession adopted by the IEA (2003) emphasizes fundamental questions and significant accomplishments, recognizing that the HFE field is constantly changing. A teaching paradigm for ergonomics should conform to established scientific standards, emphasize the development of competence in the field, and integrate theory, experimentation, design, and practice. Finally, an introductory course sequence in ergonomics should be based on the curriculum model and the disciplinary description.
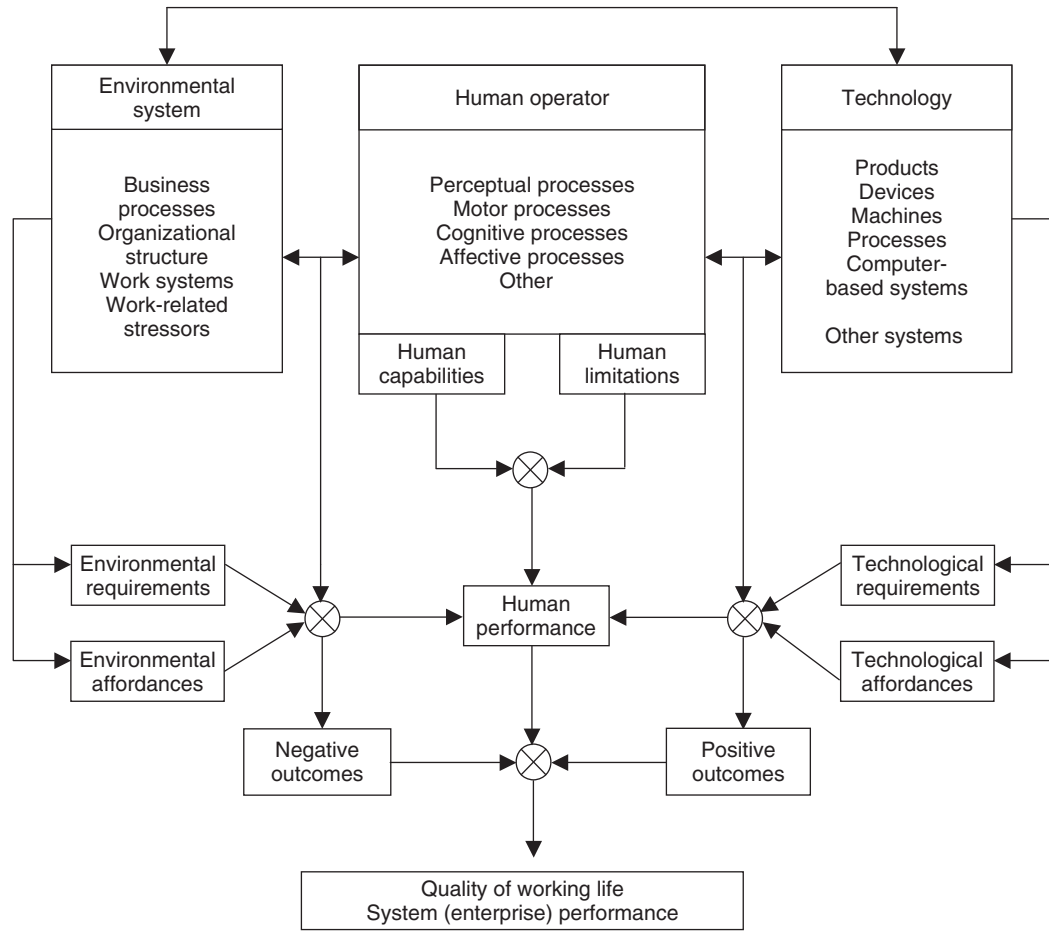
Note: ⊗—Matching of compatibility relationships

**Figure 4**   Human–system compatibility approach to ergonomics (Karwowski, 2005).



**Figure 5**   Considerations of fundamental understanding and use in ergonomics research (Karwowski 2005).

**Table 9  Specialties and Subspecialties in HFE Research**

### *System/Technology-Oriented Specialties*

1.  Aerospace: civilian and military aviation and outer space activities.
2.  Automotive: automobiles, buses, railroads, transportation functions (e.g., highway design, traffic signs, ships).
3.  Communication: telephone, telegraph, radio, direct personnel communication in a technological context.
4.  Computers: anything associated with the hardware and software of computers.
5.  Consumer products: other than computers and automobiles, any commercial product sold to the general public (e.g., pens, watches, TV).
6.  Displays: equipment used to present information to operators (e.g., HMO, HUD, meters, scales).
7.  Environmental factors/design: the environment in which human–machine system functions are performed (e.g., offices, noise, lighting).
8.  Special environment: this turns out to be underwater.

### *Process-Oriented. Specialties*

*The emphasis is on how human functions are performed and methods of improving or analyzing that performance:*

1.  Biomechanics: human physical strength as it is manifested in such activities as lifting, pulling, and so on.
2.  Industrial ergonomics (IE): papers related primarily to manufacturing; processes and resultant problems (e.g., carpal tunnel syndrome).
3.  Methodology/measurement: papers that emphasize ways of answering HFE questions or solving HFE problems.
4.  Safety: closely related to IE but with a major emphasis on analysis and prevention of accidents.
5.  System design/development: papers related to the processes of analyzing, creating, and developing systems.
6.  Training: papers describing how personnel are taught to perform functions/tasks in the human–machine system.

### *Behaviorally Oriented Specialties*

1.  Aging: the effect of this process on technological performance.
2.  Human functions: emphasizes perceptual-motor and cognitive functions. The latter differs from training in the sense that training also involves cognition but is the process of implementing cognitive capabilities. (The HFE specialty called *cognitive ergonomics/decision making* has been categorized.)
3.  Visual performance: how people see. They differ from displays in that the latter relate to equipment for seeing, whereas the former deals with the human capability and function of seeing.

*Source*: Meister (1999).

## 6  ERGONOMICS COMPETENCY AND LITERACY

As pointed out by the National Academy of Engineering (Pearson and Young, 2002), many consumer products and services promise to make people's lives easier, more enjoyable, more efficient, or healthier but very often do not deliver on this promise. Design of interactions with technological artifacts and work systems requires involvement of ergonomically competent people—people with ergonomics proficiency in a certain area, although not generally in other areas of application, similarly to medicine or engineering.

One of the critical issues in this context is the ability of users to understand the utility and limitations of technological artifacts. Ergonomics literacy prepares individuals to perform their roles in the workplace and outside the working environment. Ergonomically literate people can learn enough about how technological systems operate to protect themselves by making informed choices and making use of beneficial affordances of the artifacts and environment. People trained in ergonomics typically possess a high level of knowledge and skill related to one or more specific area of ergonomics application. Ergonomics literacy is a prerequisite to

ergonomics competency. The following can be proposed as dimensions for ergonomics literacy:

1.  *Ergonomics Knowledge and Skills*. An individual has the basic knowledge of the philosophy of human-centered design and principles for accommodating human limitations.
2.  *Ways of Thinking and Acting*. An individual seeks information about benefits and risks of artifacts and systems (consumer products, services, etc.) and participates in decisions about purchasing and use and/or development of artifacts/systems
3.  *Practical Ergonomics Capabilities*. An individual can identify and solve simple task (job)-related design problems at work or home and can apply basic concepts of ergonomics to make informed judgments about usability of artifacts and the related risks and benefits of their use.

Table 11 presents a list of 10 standards for ergonomics literacy which were proposed by Karwowski (2003) in parallel to a model of technological literacy reported by the NAE (Pearson and Young, 2002). Eight of these standards are related to developing an

**Table 10 Contemporary HFE Research Methods**

### *Physical Methods*

PLIBEL: method assigned for identification of ergonomic hazards musculoskeletal discomfort surveys used at NIOSH

Dutch musculoskeletal questionnaire (DMQ)

Quick exposure checklist (QEC) for assessment of workplace risks for work-related musculoskeletal disorders (WMSDs)

Rapid upper limb assessment (RULA)

Rapid entire body assessment

Strain index

Posture checklist using personal digital assistant (PDA) technology

Scaling experiences during work: perceived exertion and difficulty

Muscle fatigue assessment: functional job analysis technique

Psychophysical tables: lifting, lowering, pushing, pulling, and carrying

Lumbar motion monitor

Occupational repetitive-action (OCRA) methods: OCRA index and OCRA checklist

Assessment of exposure to manual patient handling in hospital wards: MAPO index (movement and assistance of hospital patients)

### *Psychophysiological Methods*

Electrodermal measurement

Electromyography (EMG)

Estimating mental effort using heart rate and heart rate variability

Ambulatory EEG methods and sleepiness

Assessing brain function and mental chronometry with event-related potentials (ERPs)

EMG and functional magnetic resonance imaging (fMRI)

Ambulatory assessment of blood pressure to evaluate workload

Monitoring alertness by eyelid closure

Measurement of respiration in applied human factors and ergonomics research

### *Behavioral and Cognitive Methods*

Observation

Heuristics

Applying interviews to usability assessment

Verbal protocol analysis

Repertory grid for product evaluation

Focus groups

Hierarchical task analysis (HTA)

Allocation of functions

Critical decision method

Applied cognitive work analysis (ACWA)

Systematic human error reduction and prediction approach (SHERPA)

Predictive human error analysis (PHEA)

Hierarchical task analysis

Mental workload

Multiple resource time sharing

Critical path analysis for multimodal activity

Situation awareness measurement and situation awareness

Keystroke level model (KLM)

GOMS

Link analysis

Global assessment technique

**Table 10** (*continued*)

### *Team Methods*

Team training
Distributed simulation training for teams
Synthetic task environments for teams: CERTTs UAV-STE
Event-based approach to training (EBAT)
Team building
Measuring team knowledge
Team communications analysis
Questionnaires for distributed assessment of team mutual awareness
Team decision requirement exercise: making team decision requirements explicit
Targeted acceptable responses to generated events or tasks (TARGETs)
Behavioral observation scales (BOS)
Team situation assessment training for adaptive coordination
Team task analysis
Team workload
Social network analysis

### *Environmental Methods*

Thermal conditions measurement
Cold stress indices
Heat stress indices
Thermal comfort indices
Indoor air quality: chemical exposures
Indoor air quality: biological/particulate-phase contaminant
Exposure assessment methods
Olfactometry: human nose as detection instrument
Context and foundation of lighting practice
Photometric characterization of luminous environment
Evaluating office lighting
Rapid sound quality assessment of background noise
Noise reaction indices and assessment
Noise and human behavior
Occupational vibration: concise perspective
Habitability measurement in space vehicles and Earth analogs

### *Macroergonomic Methods*

Macroergonomic organizational questionnaire survey (MOQS)
Interview method
Focus groups
Laboratory experiment
Field study and field experiment
Participatory ergonomics (PE)
Cognitive walk-through method (CWM)
Kansei Engineering
HITOP analysis TM
TOP-Modeler C
CIMOP System C
Anthropotechnology
Systems analysis tool (SAT)
Macroergonomic analysis of structure (MAS)
Macroergonomic analysis and design (MEAD)

*Source*: Stanton et al. (2004).

**Table 11 Standards for Ergonomics Literacy: Ergonomics and Technology**

An understanding of:

  Standard 1: characteristics and scope of ergonomics

  Standard 2: core concepts of ergonomics

  Standard 3: connections between ergonomics and other fields of study and relationships among technology, environment, industry, and society

  Standard 4: cultural, social, economic, and political effects of ergonomics

  Standard 5: role of society in the development and use of technology

  Standard 6: effects of technology on the environment

  Standard 7: attributes of ergonomics design

  Standard 8: role of ergonomics research, development, invention, and experimentation

Abilities to:

  Standard 9: apply the ergonomics design process

  Standard 10: assess the impact of products and systems on human health, well-being, system performance, and safety

*Source*: Karwowski (2007).

understanding of the nature, scope, attributes, and role of the HFE discipline in modern society, while two of them refer to the need for developing the abilities to apply the ergonomics design process and evaluate the impact of artifacts on human safety and well-being.

## 7  ERGONOMICS DESIGN

Ergonomics is the design-oriented discipline. However, as discussed by Karwowski (2005), ergonomists do not design systems; rather HFE professionals design the interactions between the artifact systems and humans. One of the fundamental problems involved in such a design is that typically there are multiple functional system–human compatibility requirements that must be satisfied at the same time. In order to address this issue, structured design methods for complex human–artifact systems are needed. In such a perspective, ergonomics design can be defined in general as mapping from the human capabilities and limitations to system (technology–environment) requirements and affordances (Figure 6), or, more specifically, from system–human compatibility needs to relevant human–system interactions.

Suh (1990, 2001) proposed a framework for axiomatic design which utilizes four different domains that reflect mapping between the identified needs ("what one wants to achieve") and the ways to achieve them ("how to satisfy the stated needs"). These domains include (1) customer requirements (customer needs or desired attributes), (2) the functional domain (functional requirements and constraints), (3) the physical domain (physical design parameters), and (4) the processes domain (processes and resources). Karwowski (2003) conceptualized the above domains for ergonomics design purposes as illustrated in
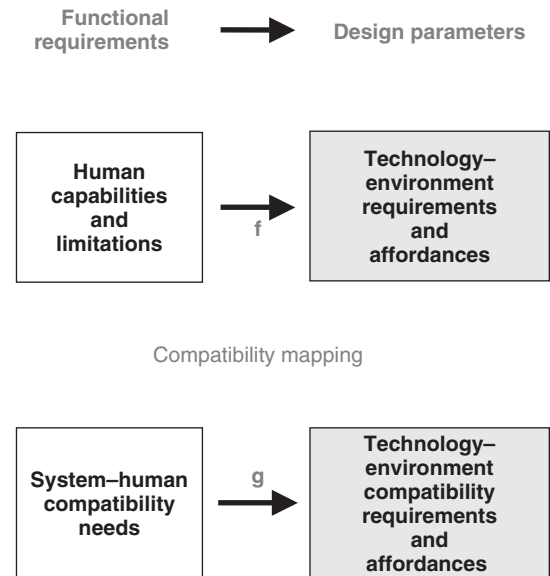


**Figure 6** Ergonomics design process: compatibility mapping (Karwowski 2005).

Figure 7 using the concept of compatibility requirements and compatibility mappings between the domains of (1) HFE requirements (goals in terms of human needs and system performance), (2) functional requirements and constraints expressed in terms of human capabilities and limitations, (3) the physical domain in terms of design of compatibility, expressed through the human–system interactions and specific work system design solutions, and (4) the processes domain, defined as management of compatibility (see Figure 8).

### 7.1  Axiomatic Design: Design Axioms

The axiomatic design process is described by the mapping process from functional requirements (FRs) to design parameters (DPs). The relationship between the two vectors FR and DP is as follows:

$$\{FR\} = [\mathbf{A}]\{DP\}$$

where $[\mathbf{A}]$ is the design matrix that characterizes the product design. The design matrix $[\mathbf{A}]$ for three functional domains (FRs) and three physical domains (DPs) is shown below:

$$[\mathbf{A}] = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}$$

The following two design axioms, proposed by Suh (1991), are the basis for the formal methodology of design: (1) the independence axiom and (2) the information axiom.
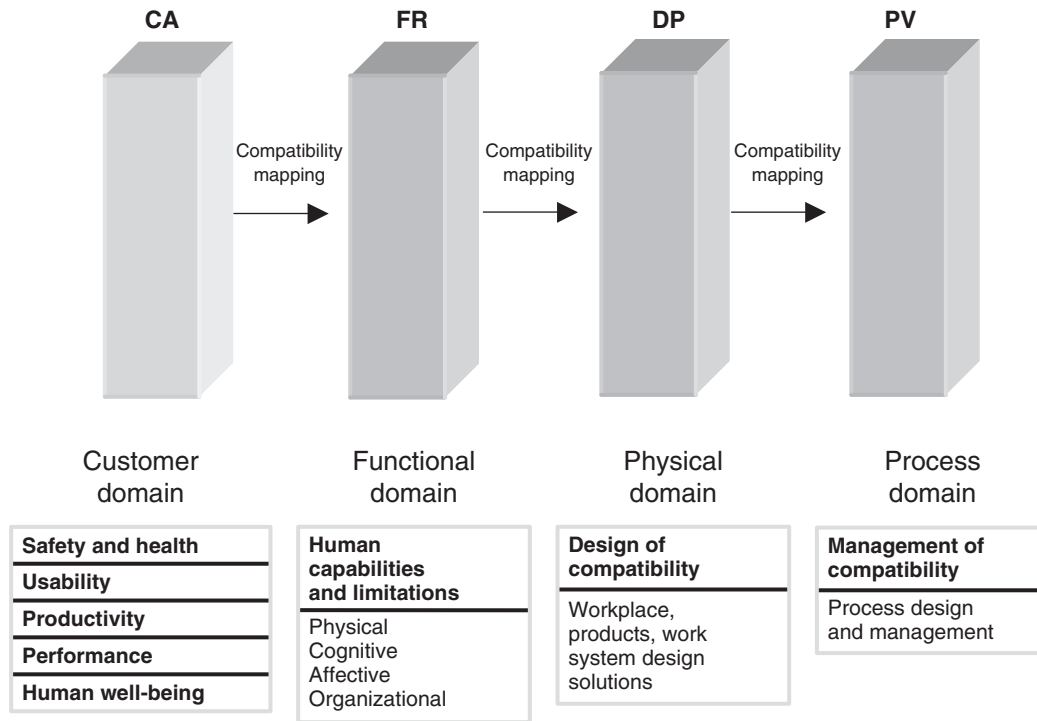
**Figure 7**   Four domains of design in ergonomics (Karwowski, 2003).
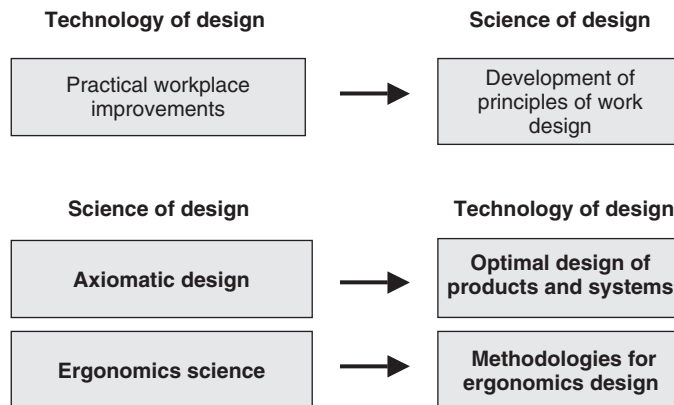


**Figure 8**   Axiomatic approach to ergonomics design (Karwowski, 2003).

### 7.1.1   Axiom 1: The Independence Axiom

This axiom stipulates a need for independence of the FRs, which are defined as the minimum set of independent requirements that characterize the design goals (defined by DPs).

### 7.1.2   Axiom 2: The Information Axiom

This axiom stipulates minimizing the information content of the design. Among those designs that satisfy the *independence axiom*, the design that has the smallest information content is the best design.

According to the second design axiom, the information content of the design should be minimized. The information content $I_i$ for a given functional requirement ($FR_i$) is defined in terms of the probability $P_i$ of satisfying $FR_i$:

$$I_i = \log_2(1/P_i) = -\log_2 P_i \text{ bits}$$

The information content will be additive when there are many functional requirements that must be satisfied simultaneously. In the general case of $m$ number of FRs, the information content for the entire system $I_{sys}$ is

$$I_{sys} = -\log_2 C_{\{m\}}$$

where $C_{\{m\}}$ is the joint probability that all $m$ FRs are satisfied.

The above axioms can be adapted for ergonomics design purposes as follows:

### 7.1.3 Axiom 1: The Independence Axiom

This axiom stipulates a need for independence of the functional compatibility requirements (FCRs), which are defined as the minimum set of independent compatibility requirements that characterize the design goals (defined by ergonomics design parameters, EDPs).

### 7.1.4 Axiom 2: The Human Incompatibility Axiom

This axiom stipulates a need to minimize the incompatibility content of the design. Among those designs that satisfy the independence axiom, the design that has the smallest incompatibility content is the best design.

As discussed by Karwowski (2001, 2003), in ergonomics design, the above axiom can be interpreted as follows. The human incompatibility content of the design $I_i$ for a given functional requirement (FR$_i$) is defined in terms of the compatibility $C_i$ index satisfying FR$_i$:

$$I_i = \log_2(1/C_i) = -\log_2 C_i \text{ ints}$$

where $I$ denotes the incompatibility content of a design.

### 7.2 Theory of Axiomatic Design in Ergonomics

As discussed by Karwowski 1991, 2001, 2003), a need to remove the system–human incompatibility (or ergonomics entropy) plays the central role in ergonomics design. In view of such discussion, the second axiomatic design axiom can be adopted for the purpose of ergonomics theory as follows.

The incompatibilty content of the design, $I_i$ for a given functional compatibility requirement (FCR$_i$), is defined in terms of the compatibility $C_i$ index that satisfies this FCR$_i$:

$$I_i = \log_2(1/C_i) = -\log_2 C_i \text{ [ints]}$$

where $I$ denotes the incompatibility content of a design, while the compatibility index $C_i$ [$0 < C < 1$] is defined depending on the specific design goals,that is, the applicable or relevant ergonomics design criterion used for system design or evaluation.

In order to minimize system–human incompatibility, one can (1) minimize exposure to the negative (undesirable) influence of a given design parameter on the system–human compatibility or (2) maximize the positive influence of the desirable design parameter (adaptability) on system–human compatibility. The first design scenario, that is, a need to minimize exposure to the negative (undesirable) influence of a given design parameter ($A_i$), typically occurs when $A_i$ exceeds some maximum exposure value of $R_i$, for example, when the compressive force on the human spine (lumbosacral joint) due to manual lifting of loads exceeds the accepted (maximum) reference value. It should be noted that if $A_i < R_i$, then $C$ can be set to 1, and the related incompatibility due to the considered design variable will be zero.

The second design scenario, that is, the need to maximize the positive influence (adaptability) of the desirable feature (design parameter $A_i$) on system human compatibility), typically occurs when $A_i$ is less than or below some desired or required value of $R_i$ (i.e., minimum reference value). For example, when the range of chair height adjustability is less than the recommended (reference) range of adjustability to accommodate 90% of the mixed (male/female) population. It should be noted that if $A_i > R_i$, then $C$ can be set to 1, and the related incompatibility due to the considered design variable will be zero. In both of the above described cases, the human–system incompatibility content can be assessed as discussed below.

1. *Ergonomics Design Criterion. Minimize exposure* when $A_i > R_i$.

The compatibility index $C_i$ is defined by the ratio $R_i/A_i$ where $R_i$ = maximum exposure (standard) for design parameter $i$ and $A_i$ = actual value of a given design parameter $i$:

$$C_i = R_i/A_i$$

and hence

$$I_i = -\log_2 C_i = -\log_2(R_i/A_i) = \log_2(A_i/R_i) \text{ ints}$$

Note that if $A_i < R_i$, then $C$ can be set to 1, and incompatibility content $I_i$ is zero.

2. *Ergonomics Design Criterion. Maximize adaptability* when $A_i < R_i$.

The compatibility index $C_i$ is defined by the ratio $A_i/R_i$, where $A_i$ = actual value of a given design parameter $i$ and $R_i$ = desired reference or required (ideal) design parameter standard $i$:

$$C_i = A_i/R_i$$

Hence

$$I_i = -\log_2 C_i = -\log_2(A_i/R_i) = \log_2(R_i/A_i) \text{ ints}$$

Note that if $A_i > R_i$, then $C$ can be set to 1 and incompatibility content $I_i$ is zero.

As discussed by Karwowski (2005), the proposed units of measurement for system–human incompatibility (ints) are parallel and numerically identical to the

measure of information (bits). The information content of the design in expressed in terms of the (ergonomics) incompatibility of design parameters with the optimal, ideal, or desired reference values, expressed in terms of ergonomics design parameters, such as range of table height or chair height adjustability, maximum acceptable load of lift, maximum compression on the spins, optimal number of choices, maximum number of hand repetitions per cycle time on a production line, minimum required decision time, and maximum heat load exposure per unit of time.

The general relationships between technology of design and science of design are illustrated in Figure 8. Furthermore, Figure 9 depicts such relationships for the HFE discipline. In the context of axiomatic design in ergonomics, the functional requirements are the human−system compatibility requirements, while the design parameters are the human−system interactions. Therefore, ergonomics design can be defined as mapping from the human−system compatibility requirements to the human−system interactions. More generally, HFE can be defined as the science of design, testing, evaluation, and management of human−system interactions according to the human−system compatibility requirements.

## 7.3  Axiomatic Design Approach in Ergonomics: Applications

Helander (1994, 1995) was first to provide a conceptualization of the second design axiom in ergonomics by considering selection of a chair based on the information content of specific chair design parameters. Recently, Karwowski (2003) introduced the concept of system incompatibility measurements and the measure on incompatibility for ergonomics design and evaluation. Furthermore, Karwowski (2003) has also illustrated an application of the first design axiom adapted to the needs of ergonomics design using an example of the design of the rear-light system utilized to provide information about application of brakes in a passenger car. The rear-light system is illustrated in Figure 10. In this highway safety-related example, the FRs of the rear-lighting (braking display) system were defined in terms
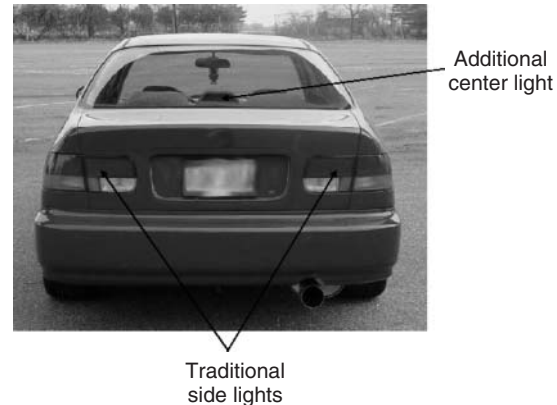


**Figure 10**  Illustration of redesigned rear-light system of an automobile.

of FRs and DPs as follows:

$$\begin{aligned}
\text{FR}_1 =\ & \text{Provide early warning to maximize lead} \\
& \text{response time (MLRT) (information} \\
& \text{about the car in front that is applying brakes)}
\end{aligned}$$

$$\text{FR}_2 = \text{Assure safe braking (ASB)}$$

The traditional (old) design solution is based on two DPs:

$$\text{DP}_1 = \text{Two rear brake lights on the sides (TRLS)}$$
$$\text{DP}_2 = \text{Efficient braking mechanism (EBM)}$$

The design matrix of the traditional rear-lighting system (TRLS) is as follows:

$$\begin{Bmatrix} \text{FR}_1 \\ \text{FR}_2 \end{Bmatrix} = \begin{pmatrix} \text{X} & 0 \\ \text{X} & \text{X} \end{pmatrix} \begin{Bmatrix} \text{DP}_1 \\ \text{DP}_2 \end{Bmatrix}$$

| MLRT | X | 0 | TRLS |
|------|---|---|------|
| ASB  | X | X | EBM  |
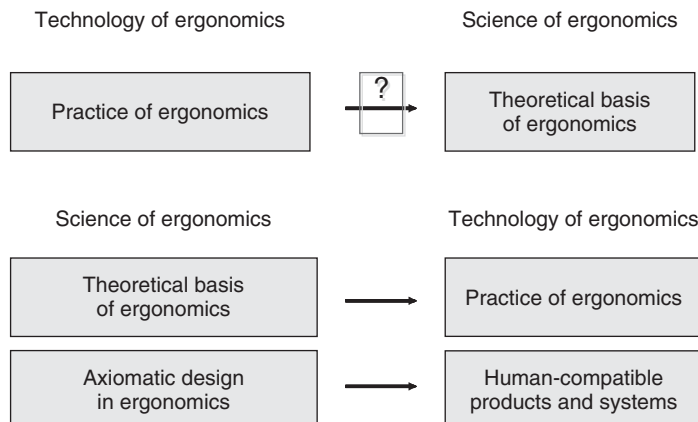


**Figure 9**   Science, technology, and design in ergonomics (Karwowski, 2003).

This rear-lighting warning system (old solution) can be classified as a *decoupled design* and is not an optimal design. The reason for such classification is that, even with the efficient braking mechanism, one cannot compensate for the lack of time in the driver's response to braking of the car in front due to a sudden traffic slowdown. In other words, this rear-lighting system does not provide early warning that would allow the driver to maximize his or her lead response time (MLRT) to braking.

The solution that was implemented two decades ago utilizes a new concept for the rear lighting of the braking system (NRLS). The new design is based on addition of the third braking light, positioned in the center and at a height that allows this light to be seen through the windshields of the car proceeding the car immediately in front. This new design solution has two DPs:

DP1 = A new rear-lighting system (NRLS)
DP2 = Efficient braking mechanism) (EBM) (the same as before)

The formal design classification of the new solution is an uncoupled design. The design matrix for this new design is as follows:

| MLRT | X | 0 | NRLS |
|------|---|---|------|
| ASB  | 0 | X | EBM  |

It should be noted that the original (traditional) rear-lighting system (TRLS) can be classified as decoupled design. This old design $[DP_{1,O}]$ does not compensate for the lack of early warning that would allow to maximize a driver's lead response time (MLRT) whenever braking is needed and, therefore, violates the second functional requirement ($FR_2$) of safe beaking. The design matrix for new system (NRLS) is an *uncoupled design* that satisfies the independence of functional requirements (independence axiom). This uncoupled design, $[DP_{1,N}]$, fulfills the requirement of maximizing lead response time (MLRT) whenever braking is needed and does not violate the $FR_2$ (safe braking requirement).

## 8 THEORETICAL ERGONOMICS: SYMVATOLOGY

It should be noted that the system–human interactions often represent complex phenomena with dynamic compatibility requirements. They are often nonlinear and can be unstable (chaotic) phenomena, the modeling of which requires a specialized approach. Karwowski (2001) indicated a need for symvatology as a corroborative science to ergonomics that can help in developing solid foundations for the ergonomics science. The proposed subdiscipline is called *symvatology*, or the science of the artifact–human (system) compatibility. Symvatology aims to discover laws of the artifact–human compatibility, proposes theories of the artifact–human compatibility, and develops a quantitative matrix for

measurement of such compatibility. Karwowski (2001) coined the term symvatology, by joining two Greek words: *symvatotis* (compatibility) and *logos* (logic, or reasoning about). Symvatology is the systematic study (which includes theory, analysis, design, implementation, and application) of interaction processes that define, transform, and control compatibility relationships between artifacts (systems) and people. An *artifact system* is defined as a set of all artifacts (meaning objects made by human work) as well as natural elements of the environment, and their interactions occurring in time and space afforded by nature. A human system is defined as the human (or humans) with all the characteristics (physical, perceptual, cognitive, emotional, etc.) which are relevant to an interaction with the artifact system.

To optimize both the human and system well-being and performance, system–human compatibility should be considered at all levels, including the physical, perceptual, cognitive, emotional, social, organizational, managerial, environmental, and political. This requires a way to measure the inputs and outputs that characterize the set of system–human interactions (Karwowski, 1991). The goal of quantifying artifact–human compatibility can only be realized if we understand its nature. Symvatology aims to observe, identify, describe, and perform empirical investigations and produce theoretical explanations of the natural phenomena of artifact–human compatibility. As such, symvatology should help to advance the progress of the ergonomics discipline by providing a methodology for the design for compatibility as well as the design of compatibility between artificial systems (technology) and humans. In the above perspective, the goal of ergonomics should be to optimize both the human and system well-being and their mutually dependent performance. As pointed out by Hancock (1997), it is not enough to assure the well-being of the human, as one must also optimize the well-being of a system (i.e., the artifacts-based technology and nature) to make the proper uses of life.

Due to the nature of the interactions, an artifact system is often a dynamic system with a high level of complexity, and it exhibits a nonlinear behavior. The *American Heritage Dictionary of English Language* (Morris, 1978) defines "complex" as consisting of interconnected or interwoven parts. Karwowski et al. (1988) proposed to represent the artifact–human system ($S$) as a construct which contains the human subsystem ($H$), an artifact subsystem ($A$), an environmental subsystem ($E$), and a set of interactions ($I$) occurring between different elements of these subsystems over time ($t$). In the above framework, compatibility is a dynamic, natural phenomenon that is affected by the artifact–human system structure, its inherent complexity, and its entropy or the level of incompatibility between the system's elements. Since the structure of system interactions ($I$) determines the complexity and related compatibility relationships in a given system, compatibility should be considered in relation to the system's complexity.

The system space, denoted here as an ordered set [(complexity, compatibility)], is defined by the four pairs as follows [(high, high), (high, low), (low, high), (low, low)]. Under the best scenario, that is, under the

most optimal state of system design, the artifact–human system exhibits high compatibility and low complexity levels. It should be noted that the transition from high to low level of system complexity does not necessarily lead to improved (higher) level of system compatibility. Also, it is often the case in most of the artifact–human systems that an improved (higher) system's compatibility can only be achieved at the expense of increasing the system's complexity.

As discussed by Karwowski et al. (1988), the lack of compatibility, or ergonomics incompatibility (EI), defined as degradation (disintegration) of the artifact–human system, is reflected in the system's measurable inefficiency and associated human losses. In order to express the innate relationship between the systems's complexity and compatibility, Karwowski et al. (1988, 1991) proposed the *complexity–incompatibility principle,* which can be stated as follows: *As the (artifact–human) system complexity increases, the incompatibility between the system elements, as expressed through their ergonomic interactions at all system levels, also increases, leading to greater ergonomic (nonreducible) entropy of the system and decreasing the potential for effective ergonomic intervention.* The above principle was illustrated by Karwowski (1995) using as an example the design of an office chair (see Figure 11). Karwowski (1992a) also discussed the complexity–compatibility paradigm in the context of organizational design. It should be noted that the above principle reflects the natural phenomena that others in the field have described in terms of difficulties

encountered in humans interacting with consumer products and technology in general. For example, according to Norman (1988), the paradox of technology is that added functionality to an artifact typically comes with the trade-off of increased complexity. These added complexities often lead to increase human difficulty and frustration when interacting with these artifacts. One of the reasons for the above is that technology which has more features also has less feedback. Moreover, Norman noted that the added complexity cannot be avoided when functions are added and can only be minimized with good design that follows natural mapping between the system elements (i.e., the control-display compatibility). Following Ashby's (1964) law of requisite variety, Karwowski (1995) proposed the corresponding law, called the "law of requisite complexity," which states that only design complexity can reduce system complexity. The above means that only the added complexity of the regulator (R = re/design), expressed by the system compatibility requirements (CR), can be used to reduce the ergonomics system entropy ($S$), that is, reduce overall artifact–human system incompatibility.

## 9   CONGRUENCE BETWEEN MANAGEMENT AND ERGONOMICS

Advanced technologies with which humans interact toady constitute complex systems that require a high level of integration from both the *design and management* perspectives. *Design integration* typically focuses on the interactions between hardware (computer-based
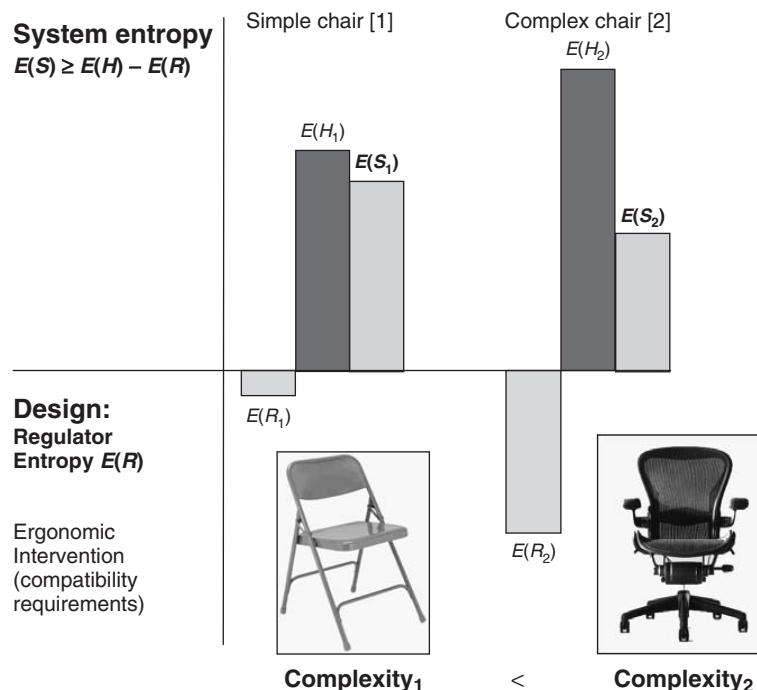


**Figure 11** System entropy determination: example of a chair design (after Karwowski, 1995).
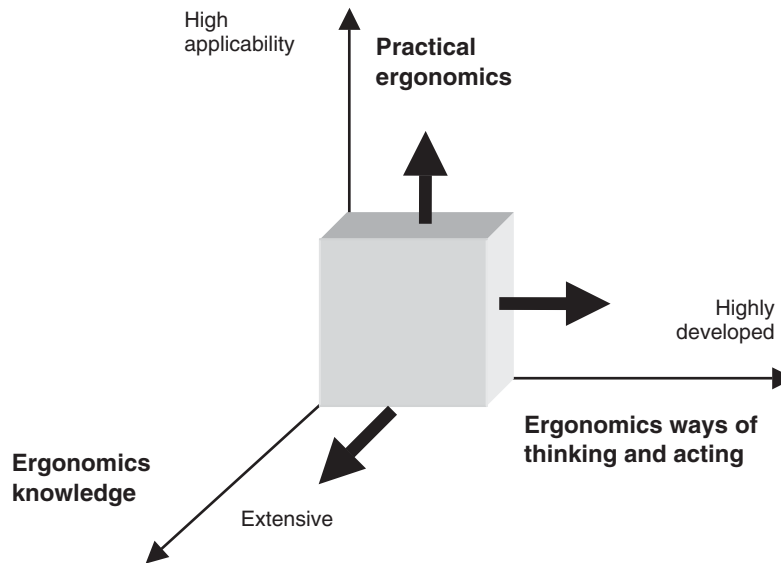
**Figure 12**  Desired goals for ergonomics literacy (Karwowski, 2003).

technology), organization (organizational structure), information system, and people (human skills, training, and expertise). *Management integration* refers to the interactions between various system elements across the process and product quality, workplace and work system design, occupational safety and health programs, and corporate environmental protection policies. As stated by Hamel (2007), "Probably for the first time since the Industrial Revolution, you cannot compete unless you are able to get the best out of people . . . ." Hamel also pointed out: "You cannot build a company that is fit for the future, unless you build a company that is fit for human beings." Unfortunately, the knowledge base of human factors and its principles of human-centered design have not yet been fully explored and applied in the area of business management. (See Figure 12.)

The scientific management originated with the work by Frederick W. Taylor (1911), who studied, among other problems, *how jobs were designed* and how workers could be trained to perform these jobs. The natural congruence between contemporary management and HFE can be described in the context of the respective definitions of these two disciplines. Management is defined today as a set of activities, including (1) planning and decision making, (2) organizing, (3) leading, and (4) controlling, directed at an organization's resources (human, financial, physical, and information) with the aim of achieving organizational goals in an efficient and effective manner (Griffin, 2001). The main elements of the management definition presented above and central to ergonomics are the following: (1) *organizing*, (2) *human resource planning*, and (3) *effective and efficient achievement of organizational goals*. In the description of these elements, the original terms proposed by the Griffin (2001) are applied in order to

ensure precision of the used concepts and terminology. *Organizing* is deciding which is the best way to group organizational elements. The *job design* is the basic building block of an organizational structure. Job design focuses on identification and determination of the tasks and activities for which the particular workers are responsible.

It should be noted that the basic ideas of management (i.e., planning and decision making, organizing, leading, and controlling) are also essential to HFE. An example of the mapping between the management knowledge (planning function) and human factors knowledge is shown in Figure 13. Specifically, common to management and ergonomics are the issues of job design and job analysis. Job design is widely considered to be the first building block of an organizational structure. Job analysis as a systematic analysis of jobs within an organization allows us to determine an individual's work-related responsibilities. The *human resource planning* is an integral part of the human resource management. The starting point for this business function is a *job analysis*, that is, a systematic analysis of the workplace in the organization. *Job analysis* consists of two parts: (1) *job description* and (2) *job specification*. Job description should include description of the task demands and the work environment conditions, such as work tools, materials, and machines needed to perform specific tasks. Job specification determines abilities, skills, and other worker characteristics necessary for *effective and efficient* tasks performance in a particular job.

The discipline of management also considers important human factors that play a role in *achieving organizational goals in an effective and efficient way*. Such factors include (1) work stress in the context of individual workers' behavior and (2) human resource management in the context of safety and health
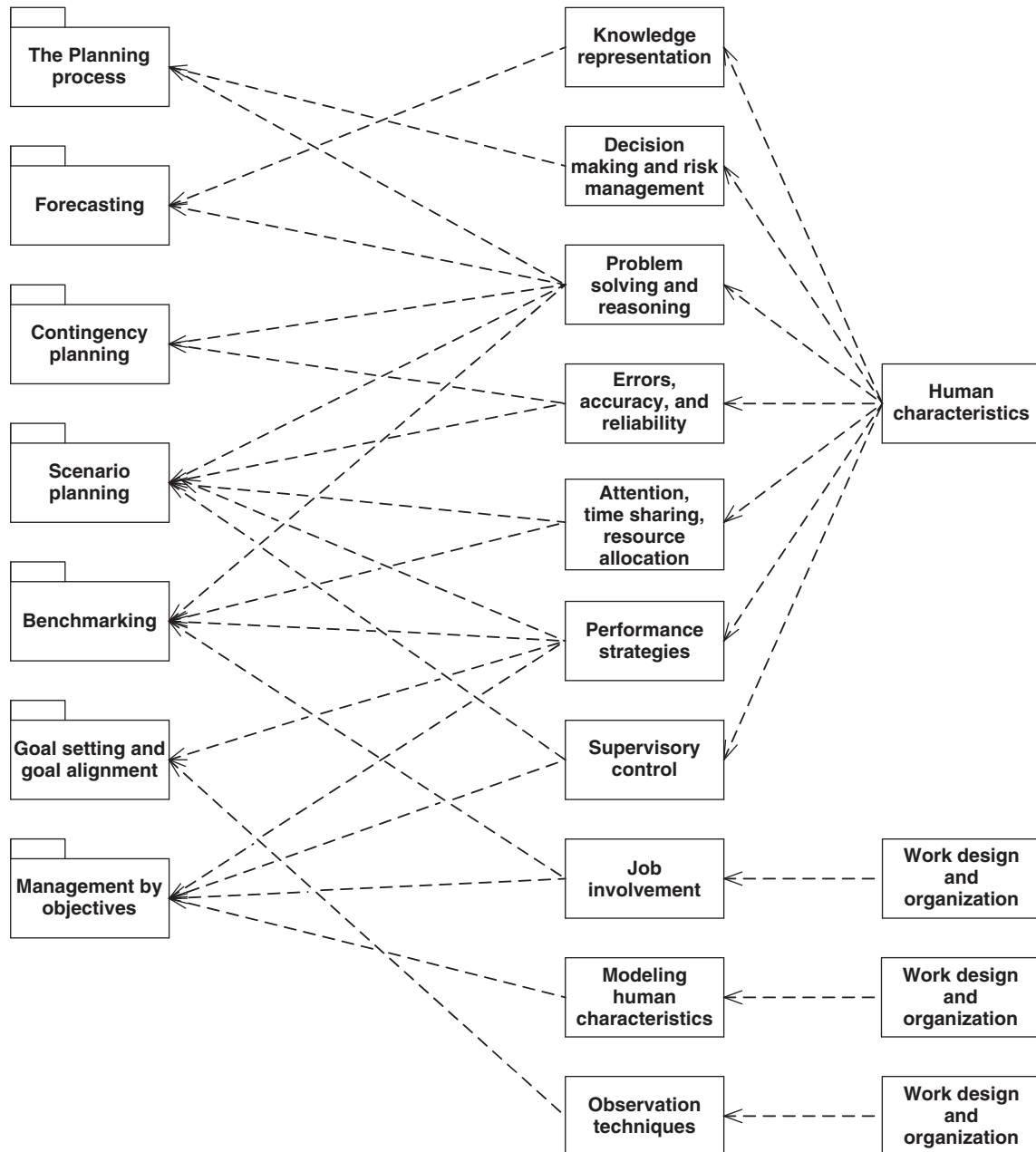
**Figure 13** Human factors knowledge mapping: planning processes (left side) related to organizational design as part of business management and relevant human characteristics (middle and right sides).

management. The work stress may be caused by the four categories of the organizational and individual factors: (1) decision related to the *task demands*; (2) *work environment demands*, including physical, perceptional, and cognitive task demands, as well as quality of the work environment, that is, adjustment of the tools and machines to the human characteristics and capabilities;

(3) *role demands* related to the relations with supervisor and co-workers; and (4) *interpersonal demands*, which can cause conflict between workers, for example, management style and group pressure. The human resource management includes provision of the safe work conditions and environment at each workstation, in the workplace, and in the entire organization.

It should also be noted that the elements of the management discipline described above, such as *job design, human resource planning* (job analysis and job specification), *work stress management,* and *safety and health management*, are essential components of the HFE subdiscipline often called industrial ergonomics. Industrial ergonomics, which investigates the human–system relationships at the individual workplace (workstation) level or at the work system level, embraces the knowledge that is also of central interest to management. From this point of view, industrial ergonomics in congruence with management is focusing on the organization and management at the workplace level (work system level) through the design and assessment (testing and evaluation) of job tasks, tools, machines, and work environments in order to adapt these to the capabilities and needs of the workers.

Another important subdiscipline of HFE with respect to the central focus of the management discipline is macroergonomics. According to Hendrick and Kleiner (2001), macroergonomics is concerned with the analysis, design, and evaluation of work systems. Work denotes any form of human effort or activity. System refers to *sociotechnical* systems, which range from a single individual to a complex multinational organization. A *work system* consists of people interacting with some form of (1) job design (work modules, tasks, knowledge, and skill requirements), (2) hardware (machines or tools) and/or software, (3) the internal environment (physical parameters and psychosocial factors), (4) the external environment (political, cultural, and economic factors), and (5) an organizational design (i.e., the work system's structure and processes used to accomplish desired functions).

The unique technology of HFE is the *human–system interface technology.* The human–system interface technology can be classified into five subparts, each with a related design focus (Hendrick, 1997; Hendrick & Kleiner, 2001):

1. *Human–machine* interface technology or hardware ergonomics
2. *Human–environment* interface technology or environmental ergonomics
3. *Human–software* interface technology or cognitive ergonomics
4. *Human–job* interface technology, or work design ergonomics
5. *Human–organization* interface technology or macroergonomics In this context, as disussed by (Hendrick and Kleiner, 2001), the HFE discipline discovers knowledge about human performance capabilities, limitations, and other human characteristics in order to develop human–system interface (HSI) technology, which includes the interface design principles, methods, and guidelines. Finally, the HFE profession applies the HSI technology to the design, analysis, test and evaluation, standardization, and control of systems.

## 10 HUMAN-CENTERED DESIGN OF SERVICE SYSTEMS

An important area of interest to the contemporary HFE discipline is the development and operation of service systems that employ today more than 60% of the workforce in the United States, Japan, and Germany (Salvendy and Karwowski, 2010). The major components in most service operations are people, infrastructure, and technology (Bitran and Pedrosa, 1998). Contemporary service systems can be characterized into four main dimensions (Fähnrich and Meiren, 2007):

- *Structure*: human, material, information, communication, technology, resources, and operating facilities
- *Processes*: process model, service provision
- *Outcomes*: product model, service content, consequences, quality, performance and standards
- *Markets*: requirement model, market requirements, and customer needs

Service system design extends the basic design concepts to include the experience that clients have with products and services. It also applies to the processes, strategies, and systems that are behind the experiences (Moritz, 2005). The key principles of customer-centered service system (CSS) design are characterized by the relationship between knowledge and technology. CSS involves the knowledge that is required to deliver the service, whether it is invested in the technology of the service or in the service provider (Hulshoff et al., 1998; McDermott et al., 2001).

Knowledge requirements in service systems design and modeling have been categorized into three main categories: knowledge based, knowledge embedded, and knowledge separated (McDermott et al., 2001). A knowledge-based service system such as teaching depends on customer knowledge to deliver the service. This knowledge may become embedded in a product that makes the services accessible to more people. An example of this is logistics providers, where the technology of package delivery is embedded in service system computers that schedule and route the delivery of packages. The delivery personnel contribute to critical components of both delivery and pickup. Their knowledge is crucial to satisfying customers and providing quality services. The CSS approach contributes to systems development processes rather than replaces them. Key principles of customer-centered service systems have been identified:

- *Clear Understanding of User and Task Requirements*. Key strengths of customer-centered service systems design are the spontaneous and active involvement of service users and the understanding of their task requirements. Involving end users will improve service system acceptance and increase commitment to the success of the new service.
- *Consistent Allocation of Functions between Users and Service System*. Allocation of
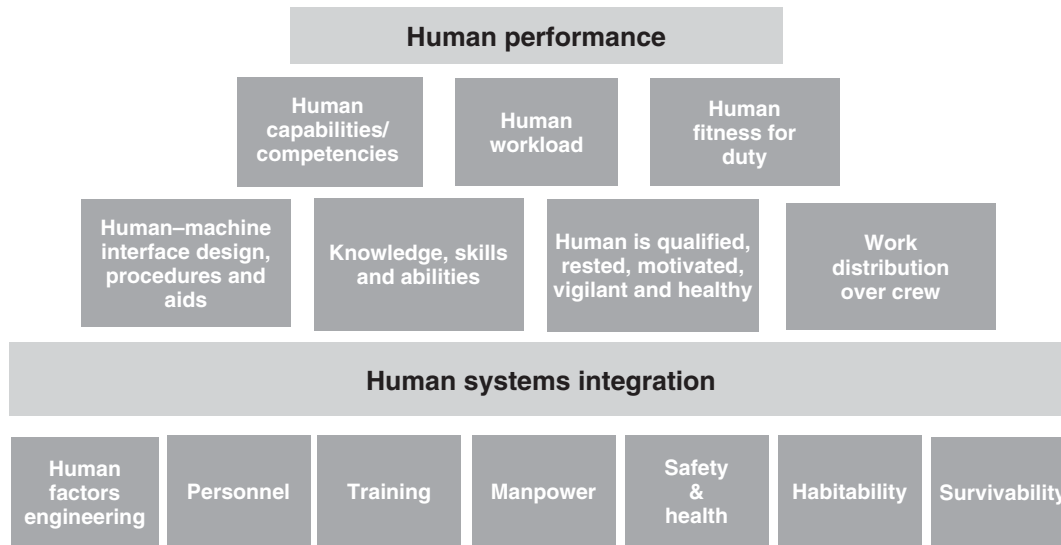
**Figure 14** Domains of human systems integration (adapted from Air Force, 2005).

functions should be based on full understanding of customer capabilities, limitations, and task demands.

- *Iterative Service System Design Approach*. Iterative service system design solutions include processing responses and feedback from service users after their use of proposed design solutions. Design solutions could range from simple paper prototypes to high-fidelity service systems mock-ups.

- *Multidisciplinary Design Teams*. Customer-centered service system design is a multitask collaborative process that involves multidisciplinary design teams. It is crucial that the service system design team comprise professionals and experts with suitable skills and interests in the proposed service system design. Such a team might include end users, service handlers (front-stage service system designers), managers, usability specialists, software engineers (back-stage service system designers), interaction designers, user experience architects, and training support professionals.

## 11 HUMAN–SYSTEMS INTEGRATION

The HFE knowledge is also being used for the purpose of human–systems integration (HSI), especially in the context of applying systems engineering to the design and development of large-scale, complex technological systems, such as those for the defense and space exploration industries (Malone and Carson, 2003; Handley and Smillie, 2008; Hardman et al., 2008; Folds et al., 2008). The knowledge management human domains have been identified internationally and are shown in Figure 14. These include human

factors engineering, manpower, personnel, training, safety and health hazards, habitability, and survivability. As discussed by Ahram and Karwowski (2009a, 2009b), these domains are the foundational human-centered domains of HSI and can be described as follows (Air Force, 2005, 2008, 2009):

**Manpower** Manpower addresses the number and type of personnel in the various occupational specialties required and potentially available to train, operate, maintain, and support the deployed system based on work and workload analyses. The manpower community promotes the pursuit of engineering designs that optimize the efficient and economic use of manpower, keeping human resource costs at affordable levels. Program managers and decision makers, who determine which manpower positions are required, must recognize the evolving demands on humans (cognitive, physical, and physiological) and consider the impact that technology can make on humans integrated into a system, both positive and negative.

**Personnel** The personnel domain considers the type of human knowledge, skills, abilities, experience levels, and human aptitudes (i.e., cognitive, physical, and sensory capabilities) required to operate, maintain, and support a system and the means to provide (recruit and retain) such people. System requirements drive personnel recruitment, testing, qualification, and selection. Personnel population characteristics can impact manpower and training as well as drive design requirements.

**Human Factors Engineering** Human factors engineering involves understanding and comprehensive integration of human capabilities (cognitive, physical, sensory, and team dynamics) into a system design, starting with conceptualization and continuing through system disposal. The primary concern for human factors engineering is to effectively integrate human–system interfaces to achieve optimal total system performance

(use, operation, maintenance, support, and sustainment). Human factors engineering, through comprehensive task analyses (including cognitive), helps define system functions and then allocates those functions to meet system requirements.

**Environment** Environment considers conditions within and around the system that affect the human's ability to function as part of the system. Steps taken to protect the total system (human, hardware, and software) from the environment as well as the environment (water, air, land, space, cyberspace, markets, organizations, and all living things and systems) from the systems design, development, manufacturing, operation, sustainment, and disposal activities are considered here. Environmental considerations may affect the concept of operations and requirements.

**Safety and Occupational Health** Safety promotes system design characteristics and procedures that minimize the potential for accidents or mishaps that cause death or injury to operators, maintainers, and support personnel as well as stakeholders and bystanders. The operation of the system itself is considered as well as prohibiting cascading failures in other systems. Using safety analyses and lessons learned from prior systems (if they exist), the safety community prompts design features to prevent safety hazards where possible and to manage safety hazards that cannot be avoided. The focus is on designs that have redundancy and, where an interface with humans exists, alerting the operators and users alike when problems arise and also help to avoid and recover from errors. Occupational health promotes system design features and procedures that minimize the risk of injury, acute or chronic illness, and disability and enhance job performance of personnel who operate, maintain, or support the system. The occupational health community seeks to prevent health hazards where possible and recommends personal protective equipment, protective enclosures, or mitigation measures where health hazards cannot be avoided. However, a balance must be found between providing too much information, thus increasing workload to unsafe levels, and mitigating minor concerns (i.e., providing too much information on faults such that managing this information becomes a task in of itself).

**Habitability** Habitability involves the characteristics of system living and working conditions such as lighting, ventilation, adequate space, vibration, noise, temperature control, availability of medical care, food and drink services, suitable sleeping quarters, sanitation, and personnel hygiene facilities. Such characteristics are necessary to sustain high levels of personnel morale, motivation, quality of life, safety, health, and comfort, contributing directly to personnel effectiveness and overall system performance. These habitability characteristics also directly impact personnel recruitment and retention.

**Survivability** Survivability addresses the characteristics of a system (e.g., life support, personal protective equipment, shielding, egress or ejection equipment, air bags, seat belts, electronic shielding) that reduce susceptibility of the total system to operational degradation or termination, to injury or loss of life, and to a partial or complete loss of the system or any of its components. These issues must be considered in the context of the full spectrum of anticipated operations and operational environments and for all people who will interact with the system (e.g., users/customers, operators, maintainers, or other support personnel). Adequate protection and escape systems must provide for personnel and system survivability when they are threatened with harm.

Malone and Carson (2003) stated the goal of the HSI paradigm as "to develop a system where the human and machine synergistically and interactively cooperate to conduct the mission." They state that the "low hanging fruit" of performance improvement lies in the human–machine interface block. The basic steps for the HSI approach can be summarized as follows (Karwowski and Ahram, 2009):

- *Human–Systems Integration Process*. Apply a standardized HSI approach that is integrated with systems processes.
- *Top-Down Requirements Analysis*. Conduct this type of analysis at the beginning and at appropriate points to decide which steps to take to optimize manpower and system performance.
- *Human–Systems Integration Strategy*. Incorporate HSI inputs into system processes throughout the life cycle, starting from the beginning of the concept and continuing through the operational life of the system.
- *Human–Systems Integration Plan*. Prepare and update this plan regularly to facilitate HSI activities.
- *Human–Systems Integration Risks*. Identify, prioritize, track, and mitigate factors that will adversely affect human performance.
- *Human–Systems Integration Metrics*. Implement practical metrics in specifications and operating procedures to evaluate progress continually.
- *Human Interfaces*. Assess the relationships between the individual and the equipment, between the individual and other individuals, and between the individual (or organization) and the organization to optimize physiological, cognitive, or sociotechnical operations.
- *Modeling*. Use simulation and modeling tools to evaluate trade-offs.

## 12 COMMITTEE ON HUMAN–SYSTEMS INTEGRATION OF THE NATIONAL RESEARCH COUNCIL

As described by the NRC (2010), the Committee on Human Factors was originally created in 1980 at the request of the U.S. Army, Navy, and Air Force to assist them in addressing various military issues. This committee was renamed in 2008 as the Committee on Human-Systems Integration (COHSI) and has expanded its scope of activities to include nonmilitary issues, such as human factors engineering, physical ergonomics, training, occupational health and safety, health care,

**Table 12  Membership of the International Ergonomics Association**

| Federated Societies | Society name | Initials | Website |
|---|---|---|---|
| Argentina | Argentinian Ergonomics Society | ADEA | www.adeargentina.org.ar |
| Australia | Human Factors and Ergonomics Society of Australia | HFESA | http://www.ergonomics.org.au |
| Austria | Austrian Ergonomics Society | OAE | www.imw.tuwien.ac.at/oeae |
| Belgium | Belgian Ergonomics Society | BES | http://www.emploi.belique.be |
| Brazil | Brazilian Ergonomics Society | ABERGO | www.abergo.org.br |
| Canada | Association of Canadian Ergonomists | ACE | www.ace-ergocanada.ca |
| Chile | Chilean Ergonomics Society | SOCHERGO | http://www.sochergo.cl/ |
| China | Chinese Ergonomics Society | ChES | n/a |
| Columbia | Colombian Ergonomics Society | SCE | http://www.sociedadcolombianadeergonomia.com/ |
| Croatia | Croatian Ergonomics Society | CrES | n/a |
| Czech Republic | Czech Ergonomics Society | CzES | http://www.bozpinfo.cz/ |
| Ecuador | Ecuador Ergonomics Society | AEERGO | n/a |
| Francophone Society | French Language Ergonomics Society | SELF | http://www.ergonomie-self.org/ |
| Germany | German Ergonomics Society | GFA | www.gfa-online.de |
| Greece | Hellenic Ergonomics Society | HES | www.ergonomics.gr |
| Hong Kong | Hong Kong Ergonomics Society | HKES | http://www.ergonomics.org.hk/ |
| Hungary | Hungarian Ergonomics Society | MES | http://www.met.ergonomiavilaga.hu/subsites/index_eng.htm |
| India | Indian Society of Ergonomics | ISE | http://www.ise.org.in/ |
| Indonesia | Indonesian Ergonomics Society | PEI | http://www.iesnet.org |
| Iran | Iranian Ergonomics Society | IES | www.modares.ac.ir/ies |
| Ireland | Irish Ergonomics Society | IrES | http://www.ergonomics.ie/IES.html |
| Israel | Israel Ergonomics Association | IEA | http://www.ergonomicsisrael.org |
| Italy | Italian Society of Ergonomics | SIA | www.societadiergonomia.it |
| Japan | Japan Ergonomics Society | JES | http://www.ergonomics.jp |
| Latvia | Latvian Ergonomics Society | | http://www.ergonomika.lv |
| Mexico | Mexican Ergonomics Society | SEM | http://www.semac.org.mx |
| Netherlands | Dutch Ergonomics Society | NVVE | www.ergonoom.nl |
| New Zealand | New Zealand Ergonomics Society | NZES | www.ergonomics.org.nz |
| Nordic countries | Nordic Ergonomics Society | NES | http://www.nordicergonomics.org/ |
| Philippines | Philippines Ergonomics Society | PHILERGO | n/a |
| Poland | Polish Ergonomics Society | PES | http://ergonomia-polska.com |
| Portugal | Portuguese Ergonomics Association | APERGO | n/a |
| Russia | Inter-Regional Ergonomics Association | IREA | n/a |
| Serbia | Ergonomics Society of Serbia | ESS | n/a |
| Singapore | Ergonomics Society of Singapore | ERGOSS | http://www.ergoss.org/ |
| Slovakia | Slovak Ergonomics Association | SEA | n/a |
| South Africa | Ergonomics Society of South Africa | ESSA | www.ergonomicssa.com |
| South Korea | Ergonomics Society of Korea | ESK | http://esk.or.kr |
| Spain | Spanish Ergonomics Association | AEE | http://www.ergonomos.es |
| Switzerland | Swiss Society for Ergonomics | SSE | http://www.swissergo.ch/de/index.php |
| Taiwan | Ergonomics Society of Taiwan | EST | www.est.org.tw |
| Thailand | Ergonomics Society of Thailand | EST | www.est.or.th |
| Tunisia | Tunisian Ergonomics Society | STE | http://www.st-ergonomie.org/ |
| Turkey | Turkish Ergonomics Society | TES | |

**Table 12** (*continued*)

| Federated Societies | Society name | Initials | Website |
|---|---|---|---|
| Ukraine | All-Ukrainian Ergonomics Association | AUEA | http://ergotech.org.ua |
| United Kingdom | Institute of Ergonomics and Human Factors | ES | http://www.ergonomics.org.uk/ |
| United States | Human Factors and Ergonomics Society | HFES | http://hfes.org |
| **Affiliated Societies** | | | |
| Japan | Human Ergology Society | HES | http://www.humanergology.com |
| Nigeria | Ergonomics Society of Nigeria | ESN | www.esnig.org |
| **IEA Networks** | | | |
| | South-East Asia Network of Ergonomics Societies | SEANES | http://www.seanes.org/index1_type.html |
| | Federation of European Ergonomics Societies | FEES | www.fees-network.org |
| | Union of Latin-American Ergonomics Societies | ULAERGO | www.ulaergo.net |

*Note*: n/a = not available in public domain.
*Source*: www.iea.cc.

product design, and macroergonomics. The main objective of the committee is *to provide new perspectives on theoretical and methodological issues concerning the relationship of individuals and organizations to technology and the environment; identify critical issues in the design, test, evaluation, and use of new human-centered technologies; and advise committee sponsors on the research needed to expand the scientific and technical bases for effectively designing new technology and training employees*. Currently, the meetings and activities of the COHSI are sponsored by the Agency for Healthcare Research and Quality, Federal Aviation Administration, the Human Factors and Ergonomics Society, the National Institute on Disability and Rehabilitation Research, Office of Naval Research, the U.S. Army Research Laboratory, and the U.S. Air Force Research Laboratory.

## 13  THE INTERNATIONAL ERGONOMICS ASSOCIATION (WW.IEA.CC)

Over the last 30 years, ergonomics as a scientific discipline and as a profession has been rapidly growing, expanding its scope and breadth of theoretical inquiries, methodological basis, and practical applications (Meister 1997, 1999; Chapanis, 1999; Stanton and Young, 1999; Kuorinka, 2000; Karwowski, 2001; IEA 2003). As a profession, the field of ergonomics has seen development of formal organizational structures (i.e., the national and cross-national ergonomics societies and networks) in support of HFE discipline and professionals internationally. As of 2010, the IEA consisted of 47 member (federated) societies plus 2 affiliated societies and 3 IEA networks, representing over 18,000 HFE members worldwide (see Table 12). The main goals of

the IEA are to elaborate and advance the science and practice of ergonomics at an international level and to improve the quality of life by expanding the scope of ergonomics applications and contributions to the global society. A list of current IEA technical committees is shown in Table 13.

Some past IEA activities have focused on development of programs and guidelines in order to facilitate the discipline and profession of ergonomics worldwide. Examples of such activities include an international directory of ergonomics programs, core competencies in ergonomics, criteria for IEA endorsement of certifying bodies in professional ergonomics, guidelines for a process of endorsing a certification body in professional ergonomics, guidelines on standards for accreditation of ergonomics education programs at tertiary (university) level, or ergonomics quality in design (EQUID) programs. More information about these programs can be found on the IEA websire (www.ie.cc). In addition to the above, the IEA endorses scientific journals in the field. A list of the core HFE journals is given in Table 14. A complete classification of the core and related HFE journals was proposed by Dul and Karwowski (2004).

The IEA has also developed several actions for stimulating development of HFE in industrially developing countries (IDCs). Such actions include the following elements:

- Cooperating with international agencies such as the ILO (International Labour Organisation), WHO (World Health Organisation), and professional scientific associations with which the IEA has signed formal agreements
- Working with major publishers of ergonomics journals and texts to extend their access to

**Table 13  IEA Technical Committees**

Activity Theories for Work Analysis and Design
Aerospace HFE
Affective Product Design
Aging
Agriculture
Anthropometry
Auditory Ergonomics
Building and Construction
Ergonomics for Children and Educational Environments
Ergonomics in Design
Ergonomics in Manufacturing
Gender and Work
Healthcare Ergonomics
Human Factors and Sustainable Development
Human Simulation and Virtual Environments
Mining
Musculoskeletal Disorders
Online Communities
Organizational Design and Management
Process Control
Psychophysiology in Ergonomics
Safety & Health
Slips, Trips and Falls
Transport
Visual Ergonomics
Work with Computing Systems (WWCS)

*Source*: www.iea.cc

- federated societies, with particular focus on developing countries
- Development of support programs for developing countries to promote ergonomics and extend ergonomics training programs

- Promotion of workshops and training programs in developing countries through the supply of educational kits and visiting ergonomists
- Extending regional ergonomics "networks" of countries to countries with no ergonomics programs located in their region
- Supporting non-IEA member countries considering application for affiliation to the IEA in conjunction with the IEA Development Committee

## 14  FUTURE HFE CHALLENGES

The contemporary HFE discipline exhibits rapidly expanding application areas, continuing improvements in research methodologies, and increased contributions to fundamental knowledge as well as important applications to the needs of the society at large. For example, the subfield of neuroergonomics focuses on the neural control and brain manifestations of the perceptual, physical, cognitive, emotional, and so on, interrelationships in human work activities (Parasuraman, 2003). As the science of the brain and work environment, *neuroergonomics* aims to explore the premise of design of work to match the neural capacities and limitations of people. The potential benefits of this emerging branch of HFE are improvements of medical therapies and applications of more sophisticated workplace design principles. The near future will also see development of the entirely new HFE domain that can be called nanoergonomics. Nanoergonomics will address the issues of humans interacting with the devices and machines of extremely small dimensions and in general with the nanotechnology.

Finally, it should be noted that developments in technology and the socioeconomic dilemmas of the twenty-first century pose significant challenges for HFE discipline and profession. According to the report on major predictions for science and technology in the

**Table 14  Core HFE Journals**

| | |
|---|---|
| Official IEA journal | *Ergonomics*[a] |
| IEA-endorsed Journals | *Applied Ergonomics*[a] |
| | *Human Factors and Ergonomics in Manufacturing and Service Industries*[a] |
| | *International Journal of Industrial Ergonomics*[a] |
| | *International Journal of Human-Computer Interaction*[a] *International Journal of Occupational Safety and Ergonomics* |
| | *Theoretical Issues in Ergonomics Science* |
| | *Ergonomia: An International Journal of Ergonomics and Human Factors* |
| Other core journals | *Human Factors*[a] |
| | *Le Travail Human*[a] |
| Non-ISI journals | *Asian Journal of Ergonomics* |
| | *Japanese Journal of Ergonomics* |
| | *Occupational Ergonomics* |
| | *Tijdschrift voor Ergonomie* |
| | *Zeitschrift für Arbeitswissenschaft* |
| | *Zentralblatt für Arbeirsmedizin, Arbeitsschurz und Ergonomie* |

*Source*: Dul and Karwowski (2004).
[a]ISI (Institute for Scientific Information) ranked journals.

twenty-first century published by the Japan Ministry of Education, Culture, Sports, Science and Technology MEXT (2006), several issues will affect the future of our civilization, including developments in genetics (creation of an artificial life, extensive outer space exploration); developments in cognitive sciences (human cognitive processes through artificial systems); a revolution in medicine (cell and organ regeneration, nanorobotics for diagnostics and therapy, superprosthesis, artificial photosynthesis of foods, elimination of human starvation and malnutrition, and safe genetic foods manipulation); full recycling of resources and reusable energy (biomass and nanotechnology); changes in human habitat (100% underground manufacturing, separation of human habitat from natural environments); clean-up of the negative effects of the twentieth century (natural sources of clean energy); communication, transport, and travel (automated transport systems, revolution in supersonic small aircraft and supersonic ocean travel, underwater ocean travel); and human safety (human error avoidance technology, control of the forces of nature, intelligent systems for safety in all forms of transport). The above issues will also affect the future direction in the development of human factors and ergonomics, as the discipline that focuses on the science, engineering, design, technology, and management of human-compatible systems.

# REFERENCES

Ahram, T. Z., and Karwowski, W. (2009a), "Human Systems Integrations Modeling," in *Proceedigns of the 53rd Annual Meeting of the Human Factors and Ergonomics Society* (*HFES*), San Antonio, TX, pp. 1849–1853

Ahram, T. Z., and Karwowski, W. (2009b), "Measuring Human Systems Integration Return on Investment," paper presented at the International Council on Systems Engineering — INCOSE Spring 09 Conference: Virginia Modeling, Analysis and Simulation Center (VMASC), Suffolk, VA.

Air Force (2005), AF SAB 2005, "System-of-Systems Engineering for Air Force Capability Development," Report SAB-TR-05-04, U.S. Air Force,Department of Defense, Washington, DC.

Air Force (2008). *U.S. Air Force Human Systems Integration Handbook*, Planning and Execution of Human Systems Integration, Directorate of Human Performance Integration, Human Performance Optimization Division, 711 HPW/HPO, TX.

Air Force (2009), "United States Air Force FY09 Human Systems Integration Management Plan," Air Force Human Systems Integration Office, Office of the Vice Chief of Staff, Falls Church, VA.

Ashby, W. R. (1964), *An Introduction to Cybernetics*, Methuen & Co., London.

Awad, E., and Ghaziri, H. M. (2004), *Knowledge Management*, Prentice-Hall, Upper Saddle River, NJ.

Baber, C. (1996), "Repertory Grid Theory and Its Application to Product Evaluation," in *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland, Eds., Taylor & Francis, London, pp. 157–166.

Bedny, G., and Karwowski, W. (2007), *A Systemic-Structural Theory of Activity: Applications to Human Performance and Work Design*, CRC Press, Boca Raton, FL.

Bisantz, A. M., and Burns, C. M. (2009), *Applications of Cognitive Work Analysis*, CRC Press, Boca Raton, FL.

Bitran, G., and Pedrosa, L. (1998), "1998" *European Management Journal*, Vol. 16, No. 2, pp. 169–189.

Bridger, R. S. (2006), *Introduction to Ergonomics*, CRC Press, Boca Raton, FL.

Carayon, P. (2011), *Handbook of Human Factors and Ergonomics in Health Care and Patient Safety*, 2nd ed., CRC Press, Boca Raton, FL.

Card, S., Moran, T., and Newell, A. (1983), *The Psychology of Human-Computer Interaction*, Erlbaum, Hillsdale, NJ.

Chaffin, D. B., Anderson, G. B. J., and Martin, B. J. (2006), *Occupational Biomechanics*, 3rd ed., Wiley, New York.

Chapanis, A. (1995), *Human Factors in System Engineering*, Wiley, New York.

Chapanis, A. (1999), *The Chapanis Chronicles: 50 Years of Human Factors Research, Education, and Design*, Aegean Publishing, Santa Barbara, CA.

Chasen, C. (2009), *Safety Managers Guide to Office Ergonomics*, New York, Wiley.

Chebbykin, O., Bedny, G., and Karwowski, W., Eds. (2008), *Ergonomics and Psychology: Developments in Theory and Practice*, CRC Press, Boca Raton, FL.

Committee on Technological Literacy, National Academy of Engineering, and National Research Council (2002), in *Technically Speaking: Why All Americans Need to Know More About Technology*, G. Pearson and A. T. Young, Eds., National Academies Press, Washington, DC.

Conrad, M. (1983), *Adaptability*, Plenum, New York.

Cook, N. J., and Durso, F. (2008), *Stories of Modern Technology Failures and Cognitive Engineering Successes*, CRC Press, Boca Raton, FL.

Dekker, S. (2006), *The Field Guide to Understanding Human Error*, Ashgate, Hampshire, United Kingdom.

Dekker, S. (2007), *Just Culture: Balancing Safety and Accountability*, Hampshire, Ashgate, United Kingdom.

Diaper, D., and Stanton, N. A. (2004), *The Handbook of Task Analysis for Human-Computer Interaction*, Lawrence Erlbaum Associates, Mahwah, NJ.

Dix, A., Finlay, J., Abowd, G., and Beale R. (1993), *Human Computer Interaction*, Prentice Hall, Englewood Cliffs, NJ.

Dul, J., and Karwowski, W. (2004), "An Assessment System for Rating Scientific Journals in the Field of Ergonomics and Human Factors," *Applied Ergonomics*, Vol. 35, No. 4, pp. 301–310.

DuI, J., and Weerdmeester, B. (1993), *Ergonomics for Beginners: A Quick Reference Guide*, Taylor & Francis, London.

Dul, J., and Weerdmeester, B. (2008), *Ergonomics for Beginners: A Quick Reference Guide*, CRC Press, Boca Raton, FL.

Dzissah, J., Karwowski, W., and Yang, Y. N. (2001), "Integration of Quality, Ergonomics, and Safety Management Systems", in *International Encyclopedia of Ergonomics and Human Factors*, W. Karwowski, Ed., Taylor & Francis, London, pp. 1129–1135.

Edholm, O. G., and Murrell, K. F. H. (1973), *The Ergonomics Society: A History 1949-1970*, Ergonomics Research Society, London.

Ergonomics Information Analysis Centre (EIAC) (2000), Ergonomics Abstracts, EIAC, School of Manufacturing and Mechanical Engineering, The University Of Birmingham, Birmingham, England.

Fähnrich, K.-P., and Meiren, T. (2007), "Service Engineering: State of the Art and Future Trends," in *Advances in Service Innovations*, D. Spath and K.-P. Fähnrich, Eds., Springer, Berlin, pp. 3–16.

Folds, D., Gardner, D., and Deal, S. (2008), "Building Up to the Human Systems Integration Demonstration," *INCOSE INSIGHT* Vol. 11, No. 2, pp. 15–18.

Goldman, S. L., Nagel, R. N., and Preiss, K. (1995), *Agile Competitors and Virtual Organizations*, Van Nostrand Reinhold, New York.

Gould, J. D., and Lewis C. (1983), "Designing for Usability: Key Principles and What Designers Think," in *Proceedings of the CHI, 83 Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, pp. 50–53.

Grandjean, E. (1986), *Fitting the Task to the Man*, Taylor & Francis, London.

Griffin, R. W. (2001), *Management*, 7th ed., Houghton Mifflin, Boston, MA.

Guerin, F., Laville, F., Daniellou, F., Duraffourg, J., and Kerguelen, A. (2007), *Understanding and Transforming Work: The Practice of Ergonomics*, Anact, Lyon, France.

Hamel, G. ( 2007), "Gary Hamel speaks about the Future of Management," video by BusinessWire, available: http://www.youtube.com/watch?v=1OE0q3I-xMc.

Hancock, P. (1997), *Essays on the Future of Human-Machine Systems*, BANTA Information Services Group, Eden Prairie, MN.

Handley, H. A., and Smillie, R. J. (2008, May), "Architecture Framework Human View: The NATO Approach," *Systems Engineering*, Vol. 11, No. 2, pp. 156–164.

Hardman, N., Colombi, J., Jacques, D., and Hill, R. (2008), "What Systems Engineers Need to Know About Human-Computer Interaction," *INCOSE INSIGHT*, Vol. 11 No. 2, pp. 19–23.

Harre, R. (1972), *The Philosophies of Science*, Oxford University Press, London.

Helander, M. G. (1994), "Axiomatic Design with Human Factors Design Variables," in *Proceedings of IEA 1994 Congress*, Vol. 4, pp. 68–70.

Helander, M. G. (1995), "Conceptualizing the Use of Axiomatic Design Procedures in Ergonomics", in *Proceedings of IEA World Conference*, Associacao Brasileira de Ergonomia, Rio de Janeiro, pp. 38–41.

Helander, M. G. (1997), "Forty Years of IEA: Some Reflections on the Evolution of Ergonomics," *Ergonomics*, Vol. 40, pp. 952–961.

Helander, M. G. (1997), "The Human Factors Profession," *Handbook of Human Factor and Ergonomics*, G. Salvendy, Ed., Wiley, New York, pp. 3–16.

Helander, M. G., and Lin, L. (2002), "Axiomatic Design in Ergonomics and Extension of Information Axiom," *Journal of Engineering Design*, Vol. 13, No. 4, pp. 321–339.

Helander, M. G., Landaur, T. K., and Prabhu, P. V., Eds. (1997), *Handbook of Human-Computer Interaction*, Elsevier, Amsterdam.

Hendrick, H. W. (1997), "Organization Design and Macroergonomics" in *Handbook of Human Factors and Ergonomics*, G. Salvendy, Ed., Wiley, New York, pp. 594–636.

Hendrick, H. W., and Kleiner, B. M. (2001), *Macroergonomics: An Introduction to Work Systems Design*, The Human Factors and Ergonomics Society, Santa Monica, CA.

Hendrick, H. W., and Kleiner, B. W. (2002a), *Macroergonomics: An Introduction to Work Systems Design*, The Human Factors and Ergonomics Society, Santa Monica, CA.

Hendrick, H. W., and Kleiner, B. M., Eds. (2002b), *Macroergonomics: Theory, Methods, and Applications*, Lawrence Erlbaum Associates, Mahwah, NJ.

Hollnagel, E., Ed. (2003), *Handbook of Cognitive Task Design*, Lawrence Erlbaum Associates, Mahwah, NJ.

Holman, D., Wall, T. D., Clegg, C. W., Sparrow, P., and Howard, A. (2003), *New Workplace: A Guide to the Human Impact of Modern Working Practices*, Wiley, Chichester.

Howell, W. C. (1986), "Trends in Ergonomics Human-Factors," *Contemporary Psychology*, Vol. 31, No. 6, pp. 457–458.

Hulshoff, H. E., Westhof, F.M.J., Kirchhoff, J. J., Kirchhoff, B. A., and Walsh, S. T. (1998), *New Services: Strategic Exploratory Survey of a Dynamic Phenomenon*, EIM Small Business Research and Consultancy, Zoetermeer, The Netherlands.

Human Factors and Ergonomics Society (HFES) (2003), *Directory and Yearbook*, HFES, Santa Monica, CA.

Human Factors and Ergonomics Society (HFES) (2010). *Directory and Yearbook*, HFES, Santa Monica, CA.

Iacocca Institute (1991), *21st Century Manufacturing Enterprise Strategy: An Industry-Led View*, Vols. 1 and 2, Iacocca Institute, Bethlehem, PA.

International Ergonomics Association (IEA) (2003), "IEA Triennial Report: 2000–2003," IEA Press, Santa Monica, CA.

Jamaldin, B., and Karwowski, W. (1997), "Quantification of Human-System Compatibility (HUSYC): An Application to Analysis of the Bhopal Accident," in *From Experience to Innovation: Proceedings of the 13th Triennial Congress of the International Ergonomics Association*, Vol. 3, P. Seppala, T. Luopajarvi, C-H. Nygard, and M. Mattila, Eds., Tampere, Finland, pp. 46–48.

Japan Ministry of Education, Culture, Sports, Science and Technology MEXT (2006), White Paper on Science and Technology, http://www.mext.go.jp/english/whitepaper/1302556.htm.

Jastrzebowski, W. B. (1857a), "An Outline of Ergonomics, or the Science of Work Based upon the Truths Drawn from the Science of Nature, Part I," *Nature and Industry*, Vol. 29, pp. 227–231.

Jastrzebowski, W. B. (1857b), "An Outline of Ergonomics, or the Science of Work Based upon the Truths Drawn from the Science of Nature, Part II," *Nature and Industry*, Vol. 30, pp. 236–244.

Jastrzebowski, W. B. (1857c), "An Outline of Ergonomics, or the Science of Work Based upon the Truths Drawn from the Science of Nature, Part III," *Nature and Industry*, Vol. 31, pp. 244–251.

Jastrzebowski, W. B. (1857d), "An Outline of Ergonomics, or the Science of Work Based upon the Truths Drawn from the Science of Nature, Part IV," *Nature and Industry*, Vol. 32, pp. 253–258.

Kaber, D., and Boy, G., Eds. (2010), *Advances in Cognitive Ergonomics*, CRC Press, Boca Raton, FL.

Karwowski, W. (1991), "Complexity, Fuzziness and Ergonomic Incompatibility Issues in the Control of Dynamic Work Environments," *Ergonomics*, Vol. 34, No. 6, pp. 671–686.

Karwowski, W. (1992a), "The Complexity-Compatibility Paradigm in the Context of Organizational Design of Human-Machine Systems," in *Human Factors in Organizational Design and Management*, V. O. Brown and H. Hendrick, Eds., Elsevier, Amsterdam, pp. 469–474.

Karwowski, W. (1992b), "The Human World of Fuzziness, Human Entropy, and the Need for General Fuzzy Systems Theory," *Journal of Japan Society for Fuzzy Theory and Systems*, Vol. 4, No. 5, pp. 591–609.

Karwowski, W. (1995), "A General Modeling Framework for the Human-Computer Interaction Based on the Principles of Ergonomic Compatibility Requirements and Human Entropy," in *Work with Display Units 94*, A. Grieco, G. Molteni, E. Occhipinti, and B. Piccoli, Eds., North-Holland, Amsterdam, pp. 473–478.

Karwowski, W. (1997), "Ancient Wisdom and Future Technology: The Old Tradition and the New Science of Human Factors/Ergonomics," in *Proceedings of the Human Factors and Ergonomics Society 4th Annual Meeting, Albuquerque, New Mexico*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 875–877.

Karwowski, W. (2000), "Symvatology: The Science of an Artifact-Human Compatibility," *Theoretical Issues in Ergonomics Science*, Vol. 1, No. 1, pp. 76–91.

Karwowski, W., Ed. (2001), *International Encyclopedia of Ergonomics and Human Factors*, Taylor & Francis, London.

Karwowski, W. (2003), "The Discipline of Ergonomics: Paradigms for Ergonomics Science, Engineering and Design," presidential address presented at the Congress of the international Ergonomics Association, Seoul, Korea, August 28, 2003.

Karwowski, W. (2005), "Ergonomics and Human Factors: The Paradigms for Science, Engineering, Design, Technology, and Management of Human-Compatible Systems," *Ergonomics*, Vol. 48, No. 5, pp. 436–463.

Karwowski, W., Ed. (2006), *Handbook of Standards and Guidelines in Human Factors and Ergonomics*, Lawrence Erlbaum Associates, Mahwah, NJ.

Karwowski, W. (2007), "Towards an HF/E-Literate Society," in *Human Factors and Ergonomics Society Bulletin*, Vol. 50, No. 2, pp. 1–2, available: http://www.hfes.org/web/BulletinPdf/0207bulletin.pdf.

Karwowski, W., and Ahram, T. Z. (2009), "Interactive Management of Human Factors Knowledge for Human Systems Integration Using Systems Modeling Language," *Information Systems Management*, Vol. 26, No. 3, pp. 262–274.

Karwowski, W., and Jamaldin, B. (1995), "The Science of Ergonomics: System Interactions, Entropy, and Ergonomic Compatibility Measures," in *Advances in Industrial Ergonomics and Safety*, Vol. VII, A. C. Bittner, Jr., and P. C. Champney, Eds., Taylor & Francis, London, pp. 121–126.

Karwowski, W., and Marras, W. S., Eds. (1999), *The Occupational Ergonomics Handbook*, CRC Press, Boca Raton, FL.

Karwowski, W., and Mital, A., Eds. (1986), *Applications of Fuzzy Set Theory in Human Factors*, Elsevier Science, Amsterdam.

Karwowski, W., and Rodrick, D. (2001), "Physical Tasks: Analysis, Design and Operation," in *Handbook of Industrial Engineering*, 3rd ed., G. Salvendy, Ed., Wiley, New York, pp. 1041–1110.

Karwowski, W., and Salvendy, G., Eds. (1994), *Organization and Management of Advanced Manufacturing*, Wiley, New York.

Karwowski, W., and Salvendy, G., Eds. (2010), *Advances in Human Factors, Ergonomics, and Safety in Manufacturing and Service Industries*, CRC Press, Taylor & Francis, London.

Karwowski, W., Marek, T., and Noworol, C. (1988, August), "Theoretical Basis of the Science of Ergonomics," in *Proceedings of the 10th Congress of the International Ergonomics Association*, A. S. Adams, R. R. Hall, B. J. McPhee, and M. S. Oxenburgh, Eds., Taylor & Francis, Sydney, Australia, pp. 756–758.

Karwowski, W., Marek, T., and Noworol, C. (1994a), "The Complexity-Incompatibility Principle in the Science of Ergonomics," in Advances in Industrial Ergonomics & Safety, Vol. VI, F. Aghazadeh, Ed., Taylor & Francis, London, pp. 37–40.

Karwowski, W., Salvendy, G., Badham, R., Brodner, P., Clegg, C., Hwang, L., Iwasawa, J., Kidd, P. T., Kobayashi, N., Koubek, R., Lamarsh, J., Nagamachi, M., Naniwada, M., Salzman, H., Seppälä, P., Schallock, B., Sheridan, T., and Warschat, J. (1994b), "Integrating People, Organization and Technology in Advance Manufacturing," *Human Factors and Ergonomics in Manufacturing*, Vol. 4, pp. 1–19.

Karwowski, W., Kantola, J., Rodrick, D., and Salvendy, G. (2002a), "Macroergonomics Aspects of Manufacturing," in *Macroergonomics: An Introduction to Work System Design*, H. W. Hendrick and B. M. Kleiner, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 223–248.

Karwowski, W., Rizzo, F., and Rodrick, D. (2002b), "Ergonomics in Information Systems," in *Encyclopedia of Information Systems*, H. Bidgoli Ed., Academic, San Diego, pp. 185–201.

Karwowski, W., Siemionow, W., and Gielo-Perczak, K. (2003), "Physical Neuroergonomics: The Human Brain in Control of Physical Work Activities," *Theoretical Issues in Ergonomics Science*, Vol. 4, Nos. 1–2, 175–199.

Karwowski, W., Salvendy, G., and Ahram, T. (2010), "Customer-Centered Design of Service Organizations," in *Introduction to Service Engineering*, G. Salvendy and W. Karwowski, Eds., Wiley, Hoboken, NJ, pp. 179–206.

Kroemer, K., and Grandjean, E. (1997), *Fitting the Task to the Human*, 5th ed., Taylor & Francis, London.

Kroemer, K., Kroemer, H., and Kroemer-Elbert, K. (1994), *Ergonomics: How to Design for Ease and Efficiency*, Prentice-Hall, Englewood Cliffs, NJ.

Kuorinka, I., Ed. (2000), *History of the Ergonomics Association: The First Quarter of Century*, International Ergonomics Association Press, Santa Monica, CA.

Lehto, M., and Buck, J., Eds. (2007), *Introduction to Human Factors and Ergonomics for Engineers*, CRC Press, Boca Raton, FL.

Luczak, H., Schilick, C., and Springer, J. (1999), "A Guide to Scientific of Ergonomics Knowledge," in *The Occupational Ergonomic Handbook*, W. Karwowski and W. S. Marras, Eds., CRC Press, London, pp 27–50.

Malone, T. B., and Carson, F. (2003), "HSI Top Down Requirements Analysis," *Naval Engineers Journal*, Vol. 115, No. 2, pp. 37–48.

Marek, T., Karwowski, W., and Rice, V., Eds. (2010), *Advances in Understanding Human Performance: Neuroergonomics, Human Factors Design, and Special Populations*, CRC Press, Taylor & Francis, London.

Marras, W. S. (2008), *The Working Back: A Systems View*, Wiley Interscience, New York.

Marras, W. S., and Karwowski, W., Eds. (2006a), *Fundamentals and Assessment Tools for Occupational Ergonomics*, CRC Press, Boca Raton, FL.

Marras, W. S., and Karwowski, W., Eds. (2006b), *Interventions, Controls, and Applications in Occupational Ergonomics*, CRC Press, Boca Raton, FL.

McDermott, C. M., Kang, H., and Walsh, S. (2001), "A Framework for Technology Management in Services," *IEEE Transactions in Engineering Management*, Vol. 48, No. 3, pp. 333–341.

Meister, D. (1987), "Systems Design, Development and Testing," in *Handbook of Human Factors*, G. Salvendy, Ed., Wiley, New York, pp. 17–42.

Meister, D. (1991), *The Psychology of System Design*, Elsevier, Amsterdam.

Meister, D. (1997), *The Practice of Ergonomics*, Board of Certification in Professional Ergonomics, Bellinghan, WA.

Meister, D. (1999), *The History of Human Factors and Ergonomics*, Lawrence Erlbaum, London.

Mitchem, C. (1994), *Thinking through Technology: The Path between Engineering and Philosophy*, University of Chicago Press, Chicago.

Moray, M. (1994), "The Future of Ergonomics_The Need for Interdisciplinary Integration," in *Proceedings of the IEA Congress*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1791–1793.

Moritz, S. (2005), *Service Design: Practical Access to an Evolving Field*, Köln International School of Design, Koln, Gemany.

Morris, W., Ed. (1978), *The American Heritage Dictionary of English Language*, Houghton Mifflin, Boston.

National Academy of Engineering (2004), *The Engineer of 2020: Visions of Engineering in the New Century*, National Academies Press, Washington, DC.

National Institute of Occupational Safety and Health (NIOSH)] (1991), "Application Manual for the Revised NIOSH Lifting Equation," U.S. Department of Health and Human Services, NIOSH, Cincinnati, OH.

National Research Council (2001), *Musculoskeletal Disorders and the Workplace: Low Back and Upper Extremities*, National Academies Press, Washington, DC.

National Research Council ( 2010), The Committee on Human-Systems Integration, http://www7.nationalacademies.org/bbcss/COHSI_AbouttheCommittee.html.

Nemeth, C. (2004), *Human Factors Methods for Design*, CRC Press, Boca Raton, FL.

Nielsen, J. (1997), "Usability Engineering," in *The Computer Science and Engineering Handbook*, A. B. Tucker, Jr., Ed., CRC Press, Boca Raton, FL, pp. 1440–1460.

Nielsen, J. (2000), *Designing Web Usability: The Practice of Simplicity*, New Readers, Indianapolis, IN.

Norman, D. (1988), *The Design of Everyday Things*, Doubleday, New York.

Norman, D. A. (1993), *Things That Make Us Smart*, Addison Wesley, Reading, MA.

Opperman, R. (1994), "Adaptively Supported Adaptability," *International Journal of Human-Computer Studies*, Vol. 40, pp. 455–472.

Parasuraman, R. (2003), "Neuroergonomics: Research and Practice," *Theoretical Issues in Ergonomics Science*, Vol. 4, Nos. 1–2, pp. 5–20.

Pearson, G., and Young, T., Eds. (2002), *Technically Speaking: Why All Americans Need to Know More About Technology*, National Academies Press, Washington, DC.

Pew, R. W., and Mavor, A. S. (2008), *Human-System Integration in the System Development Process: A New Look*, National Academies Press, Washington, DC.

Pheasant, S. (1986), *Bodyspace: Anthropometry, Ergonomics and Design*, Taylor & Francis, London.

Putz-Anderson, V., Ed. (1988), *Cumulative Trauma Disorders: A Manual For Musculoskeletal Diseases of The Upper I Limbs*, Taylor & Francis, London.

Reason J. (1997), *Managing the Risks of Organizational Accidents*. Ashgate, Aldershot, United Kingdom.

Reason, J. (1999), *Managing the Risk of Organizational Accidents*, Ashgate, Aldershot, United Kingdom.

Reason, J. (2008), *The Human Contribution*, Ashgate, Aldershot, United Kingdom.

Rouse, W., Kober, N., and Mavor, A., Eds. (1997), *The Case of Human Factors in Industry and Government: Report of a Workshop*, National Academies Press, Washington, DC.

Rouse, W. B. (2007), *People and Organizations: Explorations of Human-Centered Design*, Wiley_Interscience, New York.

Salas, E., Goodwin, G. F., and Burke C. S. (2008), *Team Effectiveness in Complex Organizations*, Psychology Press, New York.

Salvendy, G., Ed. (1997), *Handbook of Human Factor and Ergonomics*, Wiley, New York.

Salvendy, G., and Karwowski, W., Eds. (2010), *Introduction to Service Engineering*, Wiley, New York.

Sanders, M. M., and McCormick, E. J. (1993), *Human Factors in Engineering & Design*, 7th ed., McGraw-Hill, New York.

Schmorrow, D., and Stanney, K. (2008), *Augmented Cognition: A Practitioner's Guide*, Human Factors and Ergonomics Society, Santa Monica, CA.

Sears, A., and Jacko, J. A. (2007), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, CRC Press, Boca Raton, FL.

Sears, A., and Jacko, J. A., Eds. (2009), *Human-Computer Interaction: Development Process*, CRC Press, Boca Raton, FL.

Silver, B. L. (1998), *The Ascent of Science*, Oxford University Press, Oxford.

Stanton, N. A., and Young, M. (1999), *A Guide to Methodology in Ergonomics: Designing for Human Use*, Taylor & Francis, London.

Stanton, N., Hedge, A., Brookhuis, K., Salas, E., and~ Hendrick, H.W. (2004), *Handbook of Human Factors and Ergonomics Methods*, CRC Press, Boca Raton, FL.

Stokes, D. E. (1997), *Pasteur's Quadrant: Basic Science and Technological Innovation*, Brookings Institution Press, Washington, DC.

Suh, N. P. (1990), *The Principles of Design*, Oxford University Press, New York.

Suh, N. P. (2001), *Axiomatic Design: Advances and Applications*, Oxford University Press, New York.

Taylor, F. W. (1911), *The Principles of Scientific Management*, Harper Bros., New York.

Vicente, K. J. (1999), *Cognitive Work Analysis: Towards Safe, Productive, and Healthy Computer-Based Work*, Lawrence Erlbaum Associates, Mahwah, NJ.

Vicente, K. J. (2004), *The Human Factor*, Routledge, New York.

Waters, T. R., Putz-Anderson, V., Garg, A., and Fine, L. J. (1993), "Revised NIOSH Equation for The Design and Evaluation of Manuallifting Tasks," *Ergonomics*, Vol. 36, No. 7, pp. 749–776.

Weick, K. E., and Sutcliffe, K. M. (2007), *Managing the Unexpected: Resilient Performance in an Age of Uncertainty*, 2nd ed., Jossey-Bass, San Francisco, CA.

Wickens, C. D., and Carswell, C. M. (1997), "Information Processing," in Handbook of Human Factors, G. Salvendy, Ed., Wiley, New York, pp. 89–129.

Wilson, J. R., and Corlett, E. N., Eds. (1995), *Evaluation of Human Work: A Practical Ergonomics Methodology*, 2nd ed., Taylor & Francis, London.

Wogalter, M., Ed. (2006), *Handbook of Warnings*, Lawrence Erlbaum Associates, Mahwah, NJ.

Womack, J., Jones, D., and Roos, D. (1990), *The Machine That Changed the World*, Rawson Associates, New York.

Woodson, W. E., Tillman, B., and Tillman, P. (1992), *Human Factors Design Handbook*, 2nd ed., McGraw Hill, New York.

Wulf, W. A. (1998), *Tech Literacy: Letter to the White House*, National Academy of Engineering, available: www.nae.edu/nae/naehome.nse/weblinks/NAEW-4NHM87.

Zacharias, G. L., MacMillian, J., and Van Hemel, S. B. (2008), *Behavioral Modeling and Simulations: From Individuals to Societies*, National Academies Press, Washington, DC.

# CHAPTER 2

# HUMAN FACTORS ENGINEERING AND SYSTEMS DESIGN

**Sara J. Czaja and Sankaran N. Nair**
**University of Miami Miller School of Medicine**
**Miami, Florida**

## 1 INTRODUCTION

### 1.1 Human Factors Engineering and the Systems Approach in Today's Environments

#### 1.1.1 Overview

*Human factors* is generally defined as the "scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data, and other methods to design in order to optimize human well-being and overall system performance" (International Ergonomics Association, 2010). The focus of human factors is on the application of knowledge about human abilities, limitations, behavioral patterns, and other characteristics to the design of person–machine systems. By definition, a person–machine system is a system which involves an interaction between people and other system components, such as hardware, software, tasks, environments, and work structures. The system may be simple, such as a human interacting with a hand tool, or it may be complex, such as an aviation system or a physician interacting with a complex computer display that is providing information about the status of a patient. The general objectives of human factors are to maximize human and system efficiency, health, safety, comfort, and quality of life (Sanders and McCormick, 1993; Wickens et al. 2004). In terms of research, this involves studying human performance to develop design principles, guidelines, methodologies, and tools for the design of the human–system interface. Research relevant to the field of human factors can range from basic, such as understanding the impact of aging on reaction time, to very applied, such as the understanding if multimodal cues enhance visual search performance in dynamic environments such as air traffic control. In terms of practice, human factors involves the application of these principles, guidelines, and tools to the actual design and evaluation of real-world systems and system components or the design of training programs and instructional materials that support the performance of tasks or the use of technology/equipment (Hendrick and Kleiner, 2001). In all instances human factors is concerned with optimizing the interaction between the human and the other systems components.

Given the focus on human performance within the context of tasks and environments, systems theory and the systems approach are fundamental to human factors engineering. Generally, systems theory argues for a unified nature of reality and the belief that the components of a system are meaningful only in terms of the general goals of the entire system. A basic tenet among systems theorists is that all systems are synergistic and that the whole is greater than the sum of its parts. This is in contrast to a reductionist approach, which focuses on a particular system component or element in isolation. The *reductionist* approach has traditionally been the "popular" approach to system design, where the focus has been on the physical or technical components of a system, with little regard for the behavioral component. In recent years the increased incidence of human error in the medical, transportation, safety, energy, and nuclear power environments and the resultant horrific consequences as well as the limited success of many technical developments have demonstrated the shortcomings of this approach and the need for a systems prospective.

As noted by Gorman and colleagues (2010), if there is a polarity between high technologies and humans that use them, errors can arise, especially if the system is asked to respond in a novel or unanticipated situation. A cogent example cited by the authors is the poor coordination of the system-level response following Hurricane Katrina. Other examples include adverse patient outcomes, oil spills (e.g., *Exxon Valdez*), and transportation incidents such as Flight 3407, the commercial commuter plane that crashed in the Buffalo, New York, area in 2009.

Implicit in the belief in systems theory is adoption of the systems approach. Generally, the systems approach considers the interaction among all of the components of a system relative to system goals when evaluating particular phenomena. Systems methodology represents a set of methods and tools applicable to (1) the analysis of systems and system problems; (2) the design, development, and deployment of systems; and (3) the management of systems and change in systems (Banathy and Jenlink, 2004). As noted by Sage and Rouse (2009), today's systems are often large scale and complex, and simply integrating individual subsystems is insufficient and does not typically result in a system that performs optimally. Instead, systems methodologies must be employed throughout the entire life cycle of the system. Further, systems engineering must use a variety of methodologies and analytical methods as well as knowledge from a multitude of disciplines.

Applied to the field of human factors, the systems concept implies that human performance must be evaluated in terms of the context of the system and that the efficiency of a system is determined by optimizing the performance of the human and the physical/technical components of the system. Further, optimization of human and system efficiency requires consideration of all major system components throughout the design process. Unfortunately, there has been a long tradition in the design and implementation of systems that places the primary emphasis on the technology components of the system without equal consideration of the person component (Gorman et al., 2010). A basic tenet of human factors is that design efforts that do not consider the human element will not achieve the maximum level of performance. For this reason, a discussion of the role of human factors in system design and evaluation is central to a handbook on human factors engineering. This is especially true in today's era of computerization and automation where systems are becoming increasingly large and complex and involve multiple components and interrelationships.

In this chapter we discuss the role of human factors engineering in system design. The focus is on the approaches and methodologies used by human factors engineers to integrate knowledge regarding human performance into the design process. The topic of system design is vast and encompasses many areas of specialization within human factors. Thus, we introduce several concepts that are covered in depth in other chapters of the handbook. Prior to discussing the design process, a summary of changes in today's systems and a brief history of the systems approach are provided. Our overall intent in the chapter is to provide an overview of the system design process and to demonstrate the importance of human factors to systems design. Further, we introduce new approaches to system design that are being applied to complex, integrated systems.

## 1.1.2 Changes in Work and Organizational Systems

Work organizations and social environments have changed enormously over the past decade, and these changes will continue as technology and demographic/social patterns evolve. Technology by its nature is dynamic, and continual developments in technology are changing work processes, the content of jobs, where work is performed, and the delivery of education and training. These changes will continue as new technologies emerge and we continue to move toward a service sector economy. For example, telework, where work is performed outside of the workplace and oftentimes in the home, is increasing on both a full- and part-time basis. In addition, technology-mediated learning, or "e-learning," is emerging as the preferred method for training employees (Czaja and Sharit, 2009).

Systems and organizations are also changing dramatically due to the growth of new organizational structures, new management practices, and technology. Changes include a shift from vertically integrated business organizations to less vertically integrated, specialized firms. Another shift is to decentralized management and collaborative work arrangements and team work across distributed organizational systems. Because of the complexity of tasks involved in complex systems, multi-operator teams are often preferred as the skills and abilities of a team can exceed the capabilities and workload constraints of individual operators (Salas et al., 2008). In these cases effective collaboration among the group members is challenging and requires a balance between efficiency and participatory involvement of as many stakeholders. In this regard, technology, such group support systems, has helped make it possible for organizations to use very large and diverse groups to solve problems. However, collaborative technology systems do not address all of the issues facing large groups such as meeting scheduling and information overload (de Vreede et al., 2010). Further, in many work domains, such as air traffic management and safety-critical domains, group members with different roles and responsibilities are distributed physically. There is also a shift towards knowledge-based organizations where intellectual capital is an important organizational asset. Together, these changes in work structures and processes result in an increased demand for more highly skilled workers who have a broader scope of knowledge and skills in decision making and knowledge management.

Also, in many domains such as the military, health care, and communication, there is an increased concern with "systems of systems" where different systems originally designed for their own purposes are integrated to produce a new and complex large system. The challenge associated with systems of systems has given rise to the discipline of human systems integration (HSI), which is a comprehensive multidisciplinary management and technical approach for ensuring consideration of

the person in all stages of the system life cycle. HSI includes manpower, personnel, training, environment, safety, health, human factors engineering, habitability, and survivability. It is also concerned with the design process and the development of tools and methods that help to ensure that stakeholders and designers work together to ensure that the abilities, limitations, and needs of users are considered in all phases of the design cycle. HSI has been largely employed in military systems (Pew and Mavor, 2007; Liu et al., 2009).

The demographics of the population are also changing. As depicted in Figure 1, the number of older adults in the United States is dramatically increasing. Of particular significance is the increase in the number of people aged 85+ years. The aging of the population has vast implications for the design of systems. For example, increases in the number of older people coupled with a shrinking labor pool due to a decline in fertility rates will threaten economic growth, living standards, and pension and health benefit financing. To this end, current changes in pension policies favor extending working life, and many industries are looking to older workers to address the emerging problem of labor and skill shortages due to the large number of older employees who are leaving the workforce and the smaller pool of available workers. Many adults in their middle and older years are choosing to remain in the workforce longer or return to work because of concerns about retirement income, health care benefits, or a desire to remain productive and socially engaged. Together, these trends suggest an increase in the number of older workers in the upcoming decades (Figure 2). These trends are paralleled in other countries. In the European Union (EU), the aging of the workforce and supporting social structures are a major

concern, and a major goal for the countries in the EU is to increase the employment rate of people aged 55–64 years (Ilmarinen, 2009). Overall, these trends imply that organizations will need to focus on strategies to accommodate an increasingly older workforce. Thus there is a great need for understanding the capabilities, limitations, and preferences of older adults with respect to current jobs, work scheduling, and training. There are also many unanswered questions regarding the impact of aging and an older workforce on team functioning and processes. This is an important consideration given the current focus on collaborative work. The aging of the population also has implications for system design within other domains such as transportation and health care.

The number of women in the labor force has also been increasing steadily, which also has implications for job and workplace design. For example, many women are involved in caregiving for an older relative or friend. Current estimates indicate that about 22% of adults in the United States are engaged in some form of caregiving. Most informal caregivers currently work either full time or part time, and these caregivers (∼59–75%) are typically middle-aged women at the peak of their earning power (Family Caregiving Alliance, 2010). As noted by Schulz and Martire (2009), increases in both labor force participation rates of women and the number of people who need informal care raise important questions about how effectively the work and caregiver roles can be combined and what strategies can be used to optimize work and caregiving scenarios. Finally, due to the globalization of trade and commerce, many systems include people from a variety of ethnic and cultural backgrounds. As noted by Strauch (2010), ethnic and
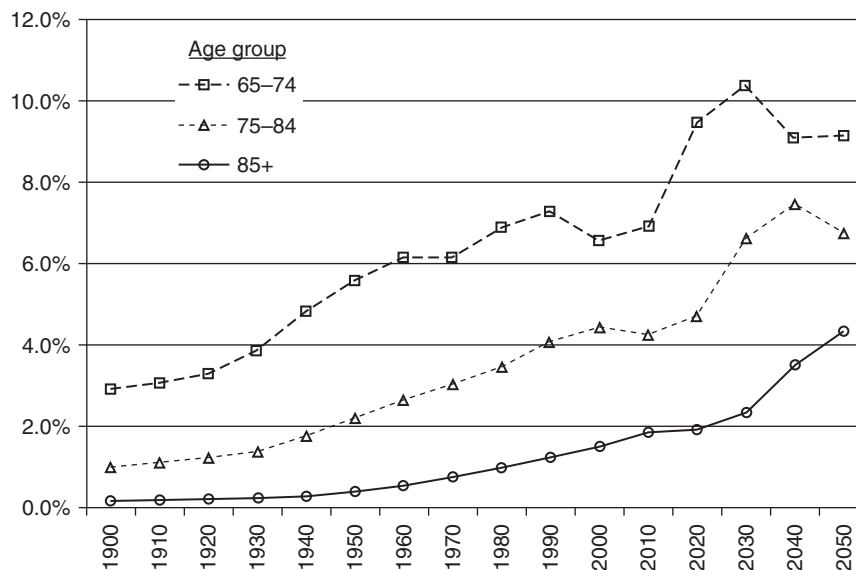


**Figure 1** U.S. historical and projected older adult population as percentage of the total population: 1900–2050 (Hobbs and Stoops, 2002; U.S. Census Bureau, 2008).
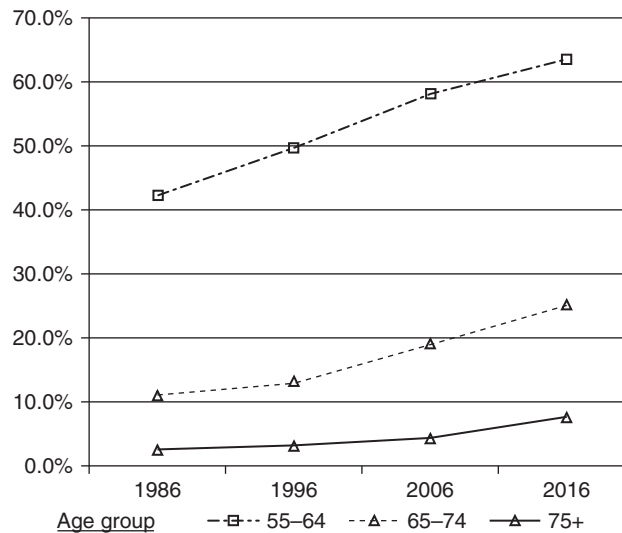
**Figure 2**   Projected labor force participation rates of older adults, 1986–2016 (Toossi, 2007).

cultural values vary with respect to work practices, communication, and family. Cultural/ethnic values are also dynamic and change over time. If ethnic/cultural factors are not considered in systems design and operations, there may be breakdowns in system team performance and overall system efficiency. To date, there is limited information on how cultural factors affect issues such as team work, communication, and the overall operations of systems. In general, all of the aforementioned issues underscore the need for a more human factors involvement in systems design.

## 1.2   Brief History of the Systems Approach and Human Factors Engineering

The systems concept was initially a philosophy associated with thinkers such as Hegel, who recognized that the whole is more than the sum of its parts. It was also a fundamental concept among Gestalt psychologists, who recognized the importance of "objectness" or wholeness to human perception. The idea of a general systems theory was developed by Bertalanffy in the late 1930s and developed further by Ashby in the 1940s (Banathy and Jenlink, 2004). The systems approach, which evolved from systems thinking, was developed initially in the biological sciences and refined by communication engineers in the 1940s. Adoption of this approach was bolstered during World War II when it was recognized that military systems were becoming too complex for humans to operate successfully. This discovery gave rise to the emergence of the field of human factors engineering and its emphasis on human–machine systems.

Sheridan (2002) classified the progress of human factors and the study of human–machine systems into three phases: phase A (knobs and dials), phase B (borrowed engineering models), and phase C (human–computer interaction). The initial time period, phase A, gave birth to the concept of human–machine systems. The focus of human factors engineers was primarily on aircraft (civilian and military) and weapon systems, with limited applications in the automotive and communication industries. Following World War II there was an appreciation of the need to continue to develop human factors. The initial focus of this effort was on the design of displays and controls and workstations for defense systems. In this era, human factors study was often equated with the study of knobs and dials. During phase B the field began to evolve beyond knobs and dials when human factors engineers recognized the applicability of system engineering models to the study of human performance. During the 1960s, systems theory became a dominant way of thinking within engineering, and human factors engineers began to use modeling techniques, such as control theory, to predict human–system performance. A number of investigators were concerned with developing models of human performance and applying these models to system design. At the same time, the application of human factors expanded beyond the military, and many companies began to establish human factors groups. The concept of the human–machine system also expanded as human factors engineers became involved with the design of consumer products and workplaces.

Phase C refers to the era of human–computer interaction. Advances in computing power and automation have changed the nature of human–machine systems dramatically, resulting in new challenges for human factors engineers and system designers. In many work domains the deployment of computers and automation has changed the nature of the demands placed on the worker. In essence, people are doing less physical work and are interacting mentally with computers and automated systems, with an emphasis on perceiving, attending, thinking, decision making, and problem solving (Rasmussen et al., 1994; Sheridan 2002; Proctor and Vu, 2010). The presence of computers and other forms

of communications technologies has become ubiquitous in most work systems. One of the most dramatic changes has been the development of the Internet, which allows increased access to vast amounts of information by a wide variety of users as well as greater interconnectivity than ever before across time zones and distances. Access to the Internet places greater demands on information processing, and information management and concerns about privacy and information security have become important issues within the fields of human factors and human–computer interaction (Proctor and Vu, 2010). Phase C is continuing to grow at a rapid pace and human factors engineers are confronted with many new types of technology and work systems, such as artificial intelligence agents, human supervisory control, and virtual reality. For example, robots are increasingly being introduced into military, space, aviation, and medical domains and research is being conducted on how to optimize human–robot teams. Issues being investigated include strategies for maximizing communication such as using gesture or gaze and how to optimally coordinate human–robot behavior. Ongoing research is also examining how theories and models of natural human interactions can be applied to robotic systems (e.g., Shah and Breazeal, 2010). Clearly, these types of systems present new challenges for system designers and human factors specialists.

To design today's work systems effectively, we need to apply knowledge regarding human information-processing capabilities to the design process. The need for this type of knowledge has created a greater emphasis on issues related to human cognition within the field of human factors and has led to the emergence of cognitive engineering (Woods, 1988). Cognitive engineering focuses on complex, cognitive thinking and knowledge-related aspects of human performance, whether carried out by humans or by machine agents (Wickens et al., 2004). It is closely aligned with the field of cognitive science and artificial intelligence. With the emphasis on team work, the concept of team cognition has emerged, which refers to the interaction between intraindividual and interindividual cognitive processes and applies the conceptual tools of cognitive science to a team or group as opposed to the individual. More recently, theories of macrocognition have been developed to guide complex collaborative processes or knowledge-based performance in nonroutinized, novel situations. It emphasizes expertise out of context and teams going beyond routine methods of performing and generating new performance processes to deal with novel situations (Fiore et al., 2010). Another new construct that has emerged is neuroergonomics, which involves the study of the mechanisms that underlie human information processing through methods used in cognitive neuroscience. These methods include neuroimaging techniques such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and event-related potentials (ERPs). These techniques have been applied to assessment workload in complex tasks and mental workload and vigilance (Parasuraman and Wilson, 2008; Proctor and Vu; 2010).

Further need for new approaches to system design comes from the changing nature of the design process. Developments in technology and automation have not only increased the complexities of the types of systems that are being designed but have also changed the design process itself and the way designers think, act, and communicate. System design is an extremely complex process that proceeds over relatively long time periods in an atmosphere of uncertainty (Meister, 2000; Sage and Rouse, 2009). The process is influenced by many factors, some of which are behavioral and some of which are physical, technical, and organizational (Table 1). As noted, design also involves interaction among many people with different types and levels of knowledge and diverse backgrounds. At the most basic level, this interaction involves engineers from many different specialties; however, in reality it also involves the users of the system being designed and organizational representatives. Further, system design often takes place under time constraints in turbulent economic and social markets. Design also involves the use of many different tools and technologies. For example, human performance models are often used to aid the design process. In this regard, there have been three major trends in the development of human performance models: manual control models, network models, and cognitive process models. Today, in many instances sophisticated models of human behavior are simulated in virtual environments to evaluate human system integration. In these instances a digital or numerical manikin is used to model human processes in an attempt to take into account factors, such as human behavior, that influence system reliability early in the design process (Lämkull et al., 2007; Fass and Leiber, 2009). These types of modeling techniques are being deployed in the aircraft and air traffic control systems as well as in the automotive and military industries. Currently, many models are complex and difficult to use without training. There is a strong need within the human factors community to improve the quality and usability of these models and to ensure that practitioners have the requisite skills to use these models (Pew, 2008).

Overall, it has become apparent that we cannot restrict the application of human factors to the design of specific jobs, workplaces, or human–machine interfaces; instead we must broaden our view of system design and consider broader sociotechnical issues. In other words, design of today's systems requires the adoption of a more *macroergonomic approach*, a top-down sociotechnical system approach to design that is concerned with the human–organizational interface and represents a broad perspective to systems design. Sociotechnical systems integrate people and social and technical elements to accomplish system objectives. Thus, people within these systems must demonstrate both social and technical skills and have an awareness of the broader environment to function effectively (Carayon, 2006). As illustrated throughout this chapter, a number of important trends are related to the organization and design of work systems that underscore the need for a macroergonomic approach, including (1) rapid

**Table 1  Design Process**

### *Elements*

1. Design specification
2. Design history (e.g., predecessor system data and analyses)
3. Design components transferred from a predecessor system
4. Design goals (technological and idiosyncratic)

### *Processes*

1. Analysis of design goals (performed by both designers and human factors ergonomics specialists)
2. Determination of design problem parameters (both)
3. Search for information to understand the design problem and parameters (both)
4. Behavioral analysis of functions and tasks (specialist only)
5. Transformation of behavioral information into physical surrogates (specialist only)
6. Development and evaluation of alternative solution to the design problem (both, mostly designers)
7. Selection of one design solution to be followed by detailed design (both, mostly designers)
8. Design of the human–machine interface, human–computer interface, human–robot interface (any may be primary)
9. Evaluation and testing of design outputs (both)
10. Determination of system status and development progress (both)

### *Factors Affecting Design*

1. Nature of the design problem and of the system, equipment, or product to be designed
2. Availability of needed relevant information
3. Strategies for solution of design problem (information-processing methods)
4. Idiosyncratic factors (designer/specialist intelligence, training, experience, skill, personality)
5. Multidisciplinary nature of the team
6. Environmental constraints and characteristics
7. Project organization and management

*Source*: Adapted from Meister (2000).

developments in technology, (2) demographic shifts, (3) changes in the value system of the workforce, (4) world competition, (5) an increased concern for safety and the resulting increase in ergonomics-based litigations, and (6) the failure of traditional microergonomics (Hendrick and Kleiner, 2001; Kleiner, 2008).

In sum, the nature of human–machine systems has changed drastically since the era of knobs and dials, presenting new challenges and opportunities for human factors engineers. We are faced not only with designing and evaluating new types of systems and a wider variety of systems (e.g., health care systems, living environments) but also with many different types of user populations. Many people with limited technical background and of varying ages are operating complex technology-based systems, which raises many new issues for system designers. For example, older workers may require different types of training or different work schedules to interact effectively with new technology, or operators with a limited technical background may require a different type of interface than those who are more experienced. Emergence of these types of issues reinforces the need to include human factors in system design. In the following section we present a general model of a system that will serve as background to a discussion of the system design process.

## 2  DEFINITION OF A SYSTEM

### 2.1  General System Characteristics

A *system* is an aggregation of elements organized in some structure (usually, hierarchical) to accomplish system goals and objectives. All systems have the following characteristics: interaction of elements, structure, purpose, and goals and inputs and outputs. A system is usually composed of humans and machines and has a definable structure and organization and external boundaries that separate it from elements outside the system. All the elements within a system interact and function to achieve system goals. Further, each system component has an effect on the other components. It is through the system inputs and outputs that the elements of a system interact and communicate. Systems also exist within an environment (physical and social), and the characteristics of this environment have an impact on the structure and the overall effectiveness of the system (Meister, 1989, 1991). For example, to be responsive to today's highly competitive and unstable environment, systems have to be flexible and dynamic. This creates the need for changes in organizational structures. Formal, hierarchical organizations do not effectively support distributed decision making and flexible processes.

Generally, all systems have the following components: (1) elements (personnel, equipment, procedures); (2) conversion processes (processes that result in changes in system states); (3) inputs or resources (personnel abilities, technical data); (4) outputs (e.g., number of units produced); (5) an environment (physical and social and organizational); (6) purpose and functions (the starting point in system development); (7) attributes (e.g., reliability); (8) components and programs; (9) management, agents, and decision makers; and (10) structure. These components must be considered in the design and evaluation of every system. For example, the nature of the system inputs has a significant impact on the ability of a system to produce the desired outputs. Inputs that are complex, ambiguous, or unanticipated may lead to errors or time delays in information processing, which in turn may lead to inaccurate or inappropriate responses. If there is conflicting or confusing information on a patient's chart, a physician might have difficulty diagnosing the illness and prescribing the appropriate course of treatment.

There are various ways in which systems are classified. Systems can be distinguished according to degree of automation, functions and tasks, feedback mechanisms, system class, hierarchical levels, and combinations of system elements (Meister, 1991). A basic distinction between open- and closed-loop systems is usually made on the basis of the nature of a system's feedback mechanisms. *Closed-loop systems* perform a process that requires continuous control and feedback for error correction. Feedback mechanisms exist that provide continuous information regarding the difference between the actual and the desired states of the system. In contrast, *open-loop systems* do not use feedback

for continuous control; when activated, no further control is executed. However, feedback can be used to improve future operations of the system (Sanders and McCormick, 1993). The distinction between open- and closed-loop systems is important, as they require different design strategies.

We are also able to describe different classes of systems. For example, we can distinguish at a very general level among educational systems, production systems, maintenance systems and health care systems, transportation systems, communication systems, and military systems. Within each of these systems we can also identify subsystems, such as the social system or the technical system. Complex systems generally contain a number of subsystems. Finally, we are able to distinguish systems according to components or elements. For example, we can distinguish among machine systems, human systems (biological systems), and human–machine systems and more recently human–robot systems and collaborative team or group systems.

## 2.2 Person–Machine Systems

A person–machine system is some combination of humans and machines that interact to achieve the goals of a system. These systems are characterized by elements that interact, structure, goals, conversion processes, inputs, and outputs. Further, they exist in an environment and have internal and external boundaries. A simple model of a human–machine system is presented in Figure 3. This general systems model applies to person–machine systems; inputs are received and processed and outputs are produced through the interaction of the system components. A more complex
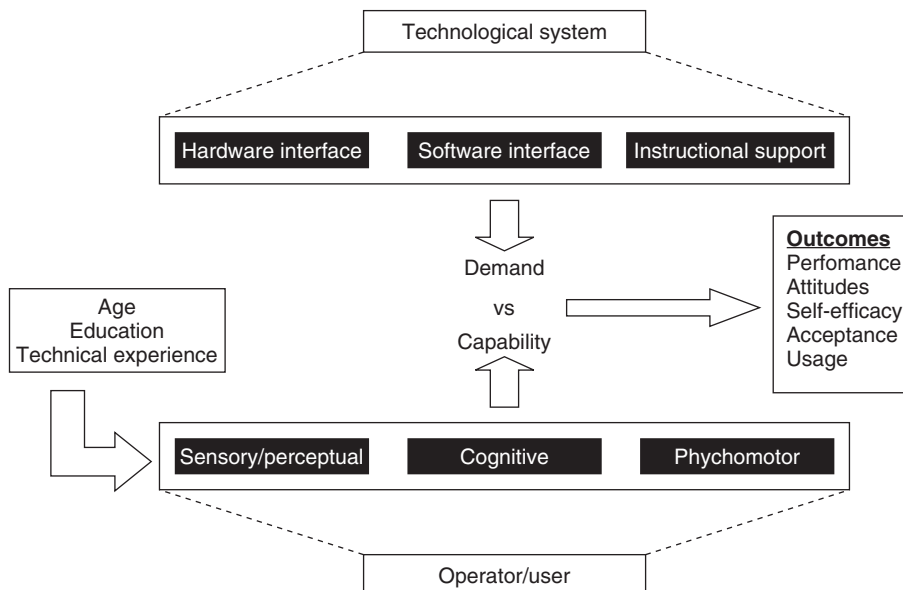


**Figure 3**  Example of human–machine system.
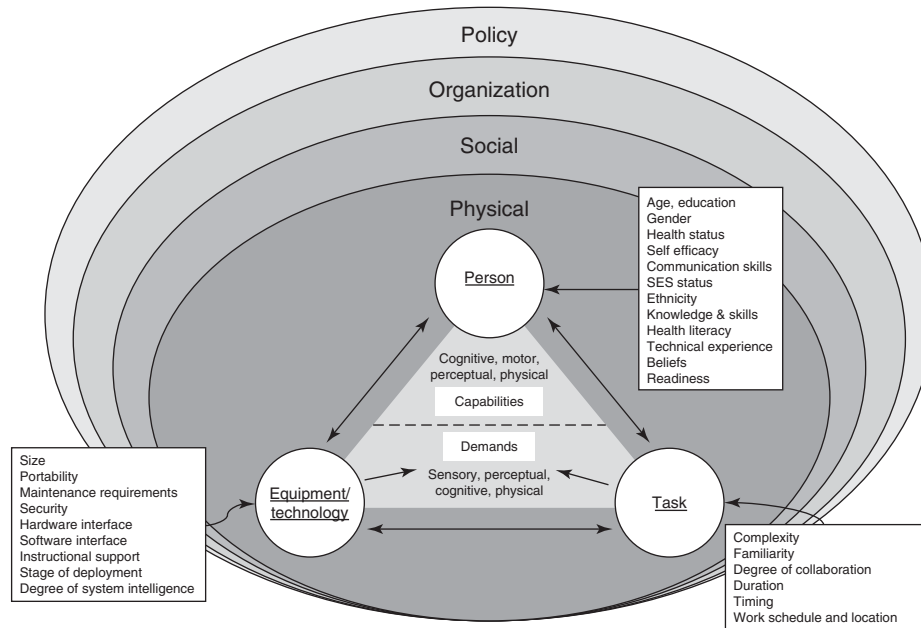
**Figure 4** Human factors model of person–task–equipment system.

model which integrates social and environmental components and is more representative of today's socio-technical systems is presented in Figure 4.

With the emergence of computer and automation technologies, the nature of person–machine systems has changed dramatically. For example, display technology has changed, and information can be presented in a wide variety of formats using multimedia approaches. Control functions have also changed, and humans can even speak commands. In addition, as noted earlier, with the advent of the Internet a vast amount of information on a wide variety of topics is available at an unprecedented rate and communication is taking on new forms with the advent of applications email and instant messaging. Perhaps more important, machines have become more intelligent and capable of performing tasks formerly restricted to humans. Prior to the development of intelligent machines, the model of the human–machine interface was formed around a control relationship in which the machine was under human control. In current human–machine systems (which involve some form of advanced technology), the machine is intelligent and capable of extending the capabilities of the human. Computer/automation systems can now perform routine, elementary tasks and complex computations, suggest ways to perform tasks, or engage in reasoning or decision making. In these instances, the human–machine interface can no longer be conceptualized in terms of a control relationship where the human controls the machine. A more accurate representation is a partnership where the human and the machine are engaged in two-way cognitive interaction. Also, in today's workplace human–computer interaction tasks often involve networks among groups of individuals.

For example, in aircraft piloting, the introduction of the flight management system (FMS) has dramatically changed the tasks of the pilot. The FMS is capable of providing the pilot with advice on navigation, weather patterns, airport traffic patterns, and other topics and is also capable of detecting and diagnosing abnormalities. The job of the pilot has become that of a process manager, and in essence the workspace of the pilot has become a desk; there is limited manual control of the flight system (Sheridan, 2002). Further the Next Generation Air Transportation System (NextGen) project is transforming the air transportation system in the United States through the incorporation of modern technologies. This will also have vast implications for pilots and air traffic controllers who will be assuming vast changes in job demands, roles, and responsibilities (http://www.jpdo.gov; Proctor and Vu, 2010). Rapidly advancing technologies such as image-guided navigation systems are being designed to support minimally invasive surgical procedures. Initially these systems, which represent a partial automation system for some aspects of a surgeon's task, were largely used in neurosurgery; however, they are increasingly being used in other surgical fields such as orthopedics. As discussed by Manzey and colleagues (2009), these tools are helpful for surgeons and have resulted in performance improvements. However, there are several human factors issues such as mental workload and training that need to be considered prior to their implementation. Other types of systems such as automotive systems are also incorporating new computer, communication, and control technologies that change the way that operators interact with these systems and raise new design concerns. With respect to automobiles, a number of issues related to driver

safety are emerging: For example, are maps and route information systems a decision aid or a distraction?

Similar issues are emerging in other domains. For example, flexible manufacturing systems represent some combination of automatic, computer-based, and human control. In these systems the operators largely assume the role of a supervisory controller and must plan and manage the manufacturing operation. Issues regarding function allocation are critical within these systems, as is the provision of adequate cognitive and technical support to the humans. Computers now offer the potential of assisting humans in the performance of cognitive activities, such as decision making, and a question arises as to what level of machine power should be deployed to assist human performance so that the overall performance of the system is maximized. This question has added complexity, as in most complex systems the problem is not restricted to one operator but to two or more operators who cooperate and have access to different databases. Today's automated systems are becoming even more complex with more decision elements, multiple controller set points, more rules, and more distributed objective functions and goals. Further, different parts of the system, both human and machine, may attempt to pursue different goals, and these goals may be in conflict. This is commonly referred to as the *mixed-initiative problem*, in which mixed human initiatives combine with mixed automation initiatives. Most systems of this type are supervised by teams of people in which the operator is part of a decision-making team of people who together with the automated system control the process (Sheridan, 2002). The mixed-initiative problem presents a particular challenge for system designers and human factors engineers.

Obviously, there are many different types of human–machine systems, and they vary greatly in size, structure, complexity, and so on. Although the emphasis in this chapter is on work systems where computerization is an integral system component, we should not restrict our conceptualization of systems to large, complex technological systems in production or process environments. We also need to consider other types of systems, such as a person using an appliance within a living environment, a physician interacting with a heart monitor in an intensive care unit, or an older person driving an automobile within a highway environment or using a telemedicine device within a home setting. In all cases, the overall performance of the system will be improved with the application of human factors engineering to system design.

New challenges for system design also arise from the evolution of virtual environments (VEs). Designers of these systems need to consider characteristics unique to VE systems, such as the design of navigational techniques, object selection and manipulation mechanisms, and the integration of visual, auditory, and haptic system outputs. Designers of these types of systems must enhance presence, immersion, and system comfort while minimizing consequences such as motion sickness. VE user interfaces are fundamentally different from traditional user interfaces with unique input–output devices, perspectives, and physiological interactions. As noted virtual human modeling is commonly used in the design of many systems to prevent changes late in the design process and enhance design efficiency.

Thus, in today's world, person–machine systems, which increasingly involve machine intelligence, can take many forms, depending on the technology involved and the function allocation between human and machine. Figure 5 presents the extremes of various degrees of automation and the complexity of various task scenarios. The lower left represents a system in which the human is left to perform completely predictable and, in most cases, "leftover tasks." In contrast, the upper right represents ideally intelligent automation where automated systems are deployed to maximal efficiency—a state
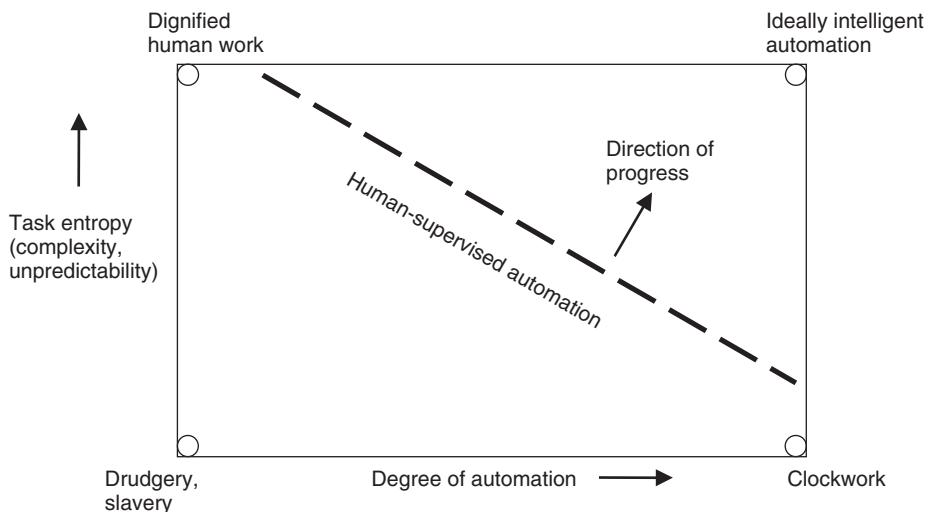


**Figure 5** Progress of human-supervised automation (Sheridan, 2002).

not attainable in the foreseeable future. The lower right also represents an effective use of full automation, and the upper left represents the most effective deployment of humans—working on undefined and unpredictable problems. As discussed by Sheridan (2002), few real situations occur at these extremes; most human-automated systems represent some trade-off of these options, which gradually progress toward the upper right—ideally, intelligent automation. Clearly, specification of the human–machine relationship is an important design decision. The relationship must be such that the abilities of both the human and machine components are maximized, as is cooperation among these components. Too often, technology is viewed as a panacea and implemented without sufficient attention to human and organizational issues.

The impact of the changing nature of person–machine systems on the system design process and current approaches to system design is discussed in a later section. However, before this topic is addressed, concepts of system and human reliability are introduced because these concepts are important to a discussion of system design and evaluation.

### 2.3 System Reliability

*System reliability* refers to the dependability of performance of the system, subsystem, or system component in carrying out its intended function for a specified period of time. Reliability is usually expressed as the probability of successful performance; therefore, for the probability estimate to be meaningful, the criteria for successful performance must be specified (Proctor and Van Zandt, 1994; Sheridan, 2008). The overall reliability of a system depends on the reliability of the individual components and how they are combined within a system. The reliability of a component is the probability that it does not fail and is defined as $r$, where $r = 1 − p$; $p$ represents the probability of failure.

Generally, components in a system are arranged in series, in parallel, or some combination of both. For total performance of the system to be satisfactory, if the components are arranged in series, they must all operate adequately. In this case, if the component failures are independent of each other, system reliability is the product of the reliability of the individual components. Further, as more components are added to a system, the reliability of the system decreases unless the reliability of these components is equal to 1.0. The reliability of the overall system can only be as great as that of the least reliable component.

In parallel systems, two or more components perform the same function such that successful performance of the system requires that only one component operate successfully. This is often referred to as *system redundancy*; the additional components provide redundancy to guard against system failure. For these types of systems, adding components in parallel increases the reliability of the system. If all of the components are equally reliable, system reliability is determined by calculating the probability that at least one component remains functional and considering the reliability of each of the parallel subsystems. Parallel redundancy is often provided for

human functions because the human component within a system is the least reliable.

### 2.4 Human Reliability

*Human reliability* is the probability that each human component of the system will perform successfully for an extended period of time and is defined as 1 minus the operator error probability (Proctor and Van Zandt, 1994). The study of human error has become an increasingly important research concern because it has become apparent that the control of human error is necessary for the successful operation of complex, integrated systems. The incidence of human error has risen dramatically over the past few years with many disastrous consequences. It has been estimated that human error is the primary cause of most major accidents and incidents in complex systems such as process control, aviation, and the health care environments (e.g., Wickens and Hollands, 2000; Morrow et al., 2006). In fact, human error has become a significant topic within the health care domain in efforts to improve human safety and decrease litigation and health insurance costs. The topic of human error is discussed in detail elsewhere in this handbook. It is discussed briefly in this chapter because the analysis of human error has important implications for system design. It is generally recognized that many errors that people make are the result of poor system design or organizational structure, and the error is usually only one in a lengthy and complex chain of breakdowns.

Due to the prevalence of human error and the enormous and often costly consequences, the study of human error has become an important focus within human factors engineering and in fact has emerged as a well-defined discipline. In recent years a number of techniques have emerged to study human error. Generally, these techniques fall into two categories, quantitative techniques and qualitative techniques. *Quantitative techniques* attempt to predict the likelihood of human error for the development of risk assessment for the entire system. These techniques can provide useful insights into human factors deficiencies in system design and thus can be used to identify areas where human factors knowledge needs to be incorporated. However, there are shortcomings associated with these techniques, such as limitations in providing precise estimates of human performance abilities, especially for cognitive processes (Wickens, 1992). Further, designers are not able to identify all of the contingencies of the work process.

*Qualitative techniques* emphasize the causal element of human error and attempt to develop an understanding of the causal events and factors contributing to human error. Clearly, when using these approaches, the circumstances under which human error is observed and the resultant causal explanation for error occurrence have important implications for system design. If the causal explanation stops at the level of the operator, remedial measures might encompass better training or supervision; for example, a common solution for back injuries is to provide operators with training on "how to lift," overlooking opportunities for other, perhaps more effective,

changes in the system, such as modifications in management, work procedures, work planning, or resources.

As noted, analyses of many major accident events indicate that the root cause of these events can be traced to latent failures and organizational errors. In other words, human errors and their resulting consequences usually result from inadequacies in system design. An example is the crash at Dryden Airport in Ontario. The analysis of this accident revealed that the accident was linked to organizational failings such as poor training, lack of management commitment to safety, and inadequate maintenance and regulatory procedures (Reason, 1995). These findings indicate that when analyzing human error it is important to look at the entire system and the organizational context in which the error occurred.

Several researchers have developed taxonomies for classifying human errors into categories. These taxonomies are useful, as they help identify the source of human error and strategies that might be effective in coping with error. Different taxonomies emphasize different aspects of human performance. For example, some taxonomies emphasize human actions, whereas others emphasize information-processing aspects of behavior. Rasmussen and colleagues (Rasmussen, 1982; Rasmussen et al., 1994) developed a taxonomy of human errors from analyses of human involvement in failures in complex processes. This schema is based on a decomposition of mental processes and states involved in erroneous behavior. For the analysis, the events of the causal chain are followed backward from the observed accidental event through mechanisms involved at each stage. The taxonomy is based on an analysis of the work system and considers the context in which the error occurred (e.g., workload, work procedures, shift requirements). This taxonomy has been applied to the analysis of work systems and has proven to be useful for understanding the nature of human involvement in accident events.

Reason (1990, 1995) has developed a similar scheme for examining the etiology of human error for the design and analysis of complex work systems. The model is based on a systems approach and describes a pathway for identifying the organizational causes of human error. The model includes two interrelated causal sequences for error events: (1) an active failure pathway where the failure originates in top management decisions and proceeds through error-producing conditions in various workplaces to unsafe acts committed by workers at the immediate human–machine interface and (2) a latent failure pathway that runs directly from the organizational processes to deficiencies in the system's defenses. The model can be used to assess organizational safety health in order to develop proactive measures for remediating system difficulties and as an investigation technique for identifying the root causes of system breakdowns.

The implications of error analysis for system design depend on the nature of the error as well as the nature of the system. Errors and accidents have multiple causes, and different types of errors require different remedial measures. For example, if an error involves deviations from normal procedures in a well-structured technical system, it is possible to derive a corrective action for a particular aspect of an interface or task element. This might involve redesign of equipment or of some work procedure to minimize the potential for the error occurrence. However, in complex dynamic work systems it is often difficult or undesirable to eliminate the incidence of human error completely. In these types of systems, there are many possible strategies for achieving system goals; thus, it is not possible to specify precise procedures for performing tasks. Instead, operators must be creative and flexible and engage in exploratory behavior in order to respond to the changing demands of the system. Further, designers are not able to anticipate the entire set of possible events; thus, it is difficult to build in mechanisms to cope with these events. This makes inevitable a certain amount of error.

Several researchers (Rouse and Morris, 1987; Rasmussen et al., 1994) advocate the design of *error-tolerant systems*, where the system tolerates the occurrence of errors but avoids the consequences; there is a means to control the impact of error on system performance. Design of these interfaces requires an understanding of the work domain and the acceptable boundaries of behavior and modeling the cognitive activity of operators dealing with incidents in a dynamic environment. A simple example of this type of design would be a computer system which holds a record of a file so that it is not lost permanently if an operator mistakenly deletes the files. A more sophisticated example would be an intelligent monitoring system which is capable of varying levels of intervention.

Rouse and Morris (1987) describe an error-tolerant system that provides three levels of support. Two levels involve feedback (current state and future state) and rely on an operator's ability to perceive his or her own errors and act appropriately. The third level involves *intelligent monitoring*, that is, online identification and error control. They propose an architecture for the development of this type of system that is based on an operator-centered design philosophy and involves incremental support and automation. Rasmussen and Vincente (1989) have developed a framework for an interface that supports recovery from human errors. The framework, called *ecological interface design*, is based on an analysis of the work system. This approach is described in more detail in a later section.

## 3 SYSTEM DESIGN PROCESS

### 3.1 Approaches to System Design

System design is usually depicted as a highly structured and formalized process characterized by stages in which various activities occur. These activities vary as a function of system requirements, but they generally involve planning, designing, testing, and evaluating. More details regarding these activities are given in a subsequent section. Generally, system design is characterized as a top-down process that proceeds, in an interactive fashion, from broad molar functions to progressively more molecular tasks and subtasks. It is also a time-driven

process and is constrained by cost, resources, and organizational and environmental requirements. The overall goal of system design is to develop an entity that is capable of transforming inputs into outputs to accomplish specified goals and objectives.

In recent years, within the realm of system design, a great deal of attention has been given to the design philosophy and the resulting design architecture as it has become apparent that new design approaches are required to design modern complex systems. The design and analysis of such systems cannot be based on design models developed for systems characterized by a stable environment and stable task procedures. Instead, the design approach is concerned with supplying resources to people who operate in a dynamic work space, engage in collaborative relationships, use a variety of technologies, and often need to adapt their behavioral patterns to changing environmental conditions. In other words, a structural perspective whereby we describe the behavior of the system in terms of cause-and-effect patterns and arrange system elements in cause-and-effect chains is no longer adequate.

### 3.1.1  Models of System Design

The traditional view of the system design process is that it is a linear sequence of activities where the output of each stage serves as input to the next stage. The stages generally proceed from the conceptual level to physical design through implementation and evaluation. Human factors inputs are generally considered in the design and evaluation stages (Eason, 1991). The general characteristics of this approach are that it represents a reductionist approach where various components are designed in isolation and made to fit together; it is dominated by technological considerations where humans are considered secondary components. The focus is on fitting the person to the system, and different components of the system are developed on the basis of narrow functional perspectives (Kidd, 1992; Liker and Majchrzak, 1994). Generally, this approach has dominated the design of overall work systems, such as manufacturing systems, as well as the design of the human−machine interface. For example, the emphasis in the design of human−computer systems has largely been on the individual level of the human−computer interaction without much attention to task and environmental factors that may affect performance. To date too much attention has been on the microergonomic aspects of design without sufficient attention to social and organizational issues (Hendrick and Kleiner, 2001; Kleiner, 2008). The implementation of computers of automation into most work systems, coupled with the enhanced capabilities of technological systems, has created a need for new approaches to system design. As discussed, there are many instances where technology has failed to achieve its potential, resulting in failures in system performance with adverse and often disastrous consequences. These events have demonstrated that the traditional design approach is no longer adequate. A brief overview of these approaches and some other design approaches will be presented to provide some examples of alternative approaches to system design and demonstrate

methodologies and concepts that can be applied to the design of current human−machine systems. This will be followed by a discussion of the specification application of human factors engineering to design activities.

### 3.1.2  Alternative Approaches to System Design

**Sociotechnical Systems Approach** The sociotechnical systems approach, which evolved from work conducted at the Tavistock Institute, represents a complete design process for the analysis, design, and implementation of systems. The approach is based on open systems theory and emphasizes the fit between social and technical systems and the environment. This approach includes methods for analyzing the environment, the social system, and the technical system. The overall design objective is the joint optimization of the social and technical systems (Pasmore, 1988). Some drawbacks associated with sociotechnical design are that the design principles are often vague and there is often an overemphasis on the social system without sufficient emphasis on the design of the technical system.

Clegg (2000) recently presented a set of sociotechnical principles to guide system design. The principles are intended for the design of new systems that involve new technologies and modern management practices. The principles are organized into three interrelated categories: metaprinciples, content principles, and process principles. *Metaprinciples* are intended to demonstrate a world view of design, *content principles* focus on more specific aspects of the content of the new designs, and *process principles* are concerned with the design process. The principles also provide a potential for evaluative purposes. They are based on a macroergonomic perspective.

The central focus of macroergonomics is on interfacing organizational design with the technology employed in the system to optimize human−system functioning. Macroergonomics considers the human−organization−environment−machine interface, as opposed to microergonomics, which focuses on the human−machine interface. Macroergonomics is considered to be the driving force for microergonomics. Macroergonomics concepts have been applied successfully to manufacturing, service, and health care organizations as well as to the design of computer-based information systems (Hendrick and Kleiner, 2001; Kleiner, 2008).

**Participatory Ergonomics** Participatory ergonomics is the application of ergonomic principles and concepts to the design process by people who are part of the work group and users of the system. These people are typically assisted by ergonomic experts who serve as trainers and resource centers. The overall goal of participatory design is to capitalize on the knowledge of users and to incorporate their needs and concerns into the design process. Methods, such as focus groups, quality circles, and inventories, have been developed to maximize the value of user participation. Participatory ergonomics has been applied to the design of jobs and workplaces and to the design of products. For example, the quality circle approach was adopted by a refrigerator

manufacturing company that needed a systemwide method for assessing the issues of aging workers. The assembly line for medium-sized refrigerators was chosen as an area for job redesign. The project redesign team involved workers from the line as well as other staff members. The team was instructed with respect to the principles of ergonomics and design for older workers. The solution, proposed by the team, for improving the assembly line resulted in improved performance and also allowed older workers to continue to perform the task (Imada et al., 1986). The design of current personal computer systems also typically involves user participation. Representative users participate in usability studies. In general, participatory ergonomics does not represent a design process because it does not consider broader system design issues but rather focuses on individual components. However, the benefits of user participation should not be overlooked and should be a fundamental aspect of system design.

**User-Centered Design** The user-centered design approach represents an approach where human factors are of central concern within the design process. It is based on an open-systems model and considers the human and technical subsystems within the context of the broader environment. User-centered approaches propose general specifications for system design, such as that the system must maximize user involvement at the task level and the system should be designed to support cooperative work and allow users to maintain control over operations (Liker and Majchrzak, 1994). Essentially, this design approach incorporates user requirements, user goals, and user tasks as early as possible into the design of a system, when the design is still relatively flexible and changes can be made at least cost.

Eason (1989) has developed a detailed process for user-centered design in which a system is developed in an evolutionary incremental fashion and development of the social system complements development of the technical system. Eason maintains that the technical system should follow the design of jobs and the design of the technical system must involve user participation and consider criteria for four factors: functionality, usability, user acceptance, and organizational acceptance. Once these criteria are identified, alternative design solutions are developed and evaluated. There are different philosophies with respect to the nature of user involvement. Eason emphasizes user involvement throughout the design process, whereas with other models the users are considered sources of data and the emphasis is on translating knowledge about users into practice. Advocates of the user participation approach argue that users should participate in the choice between alternatives because they have to live with the results. Advocates of the knowledge approach express concern about the ability of users to make informed judgments. Eason (1991) maintains that designers and users can form a partnership where both can play an effective role. A number of methods are used in user-centered design, including checklists and guidelines, observations, interviews, focus groups, and task analysis.

**Computer-Supported Design** The design of complex technical systems involves the interpretation and integration of vast amounts of technical information. Further, design activities are typically constrained by time and resources and involve the contributions of many persons with varying backgrounds and levels of technical expertise. In this regard, computer-based design support tools have emerged to aid designers and support the design of effective systems. These systems are capable of offering a variety of supports, including information retrieval, information management, and information transformation. The type of support warranted depends on the needs and expertise of the designer (Rouse, 1987). A common example of this type of support is a computer-aided design/computer-aided manufacturing (CAD/CAM) system.

There are many issues surrounding the development and deployment of computer-based design support tools, including specification of the appropriate level of support, determination of optimal ways to characterize the design problem and the type of knowledge most useful to designers, and the identification of factors that influence the acceptance of these tools. A discussion of these issues is beyond the scope of this chapter. Refer to Rouse and Boff (1987a,b) for an excellent review of this topic.

**Ecological Interface Design** Ecological interface design (EID) is a theoretical framework for designing human–computer interfaces for complex sociotechnical systems (Rasmussen et al., 1994; Vincente, 2002). The primary aim of EID is to support knowledge workers who are required to engage in adaptive problem solving in order to respond to novelty and change in system demands. EID is based on a cognitive systems engineering approach and involves an analysis of the work domain and the cognitive characteristics and behavior tendencies of the individual. Analysis of the work domain is based on an abstraction hierarchy (means–end analysis) (Rasmussen, 1986) and relates to the specification of information content. The skills–rules–knowledge taxonomy (Rasmussen, 1983) is used to derive inferences for how information should be presented. The aims of EID are to support the entire range of activities that confront operators, including familiar, unfamiliar, and unanticipated events, without contributing to the difficulty of the task.

EID has been applied to a variety of domains, such as process control, aviation, software engineering, and medicine, and has been shown to improve performance over that achieved by more traditional design approaches. However, there are still some challenges confronting the widespread use of EID in the industry. These challenges include the time and effort required to analyze the work domain, choice of the interface form, and the difficulty of integrating EID with the design of other components of a system (Vincente, 2002).

## 3.2 Incorporating Human Factors in System Design

One problem faced by human factors engineers in system design is convincing project managers, engineers, and designers of the value of incorporating human factors knowledge and expertise into the system design process. In many instances, human factors issues are

ignored or human factors activities are restricted to the evaluation stage. This is referred to as the "too little, too late" phenomenon (Lim et al., 1992). Restricting human factors inputs to the evaluation stage limits the utility and effectiveness of human factors contributions. Either the contributions are ignored because it would be too costly or time consuming to alter the design of the system ("too late") or minor alterations are made to the design to pay lip service to human issues ("too little"). In either case there is limited realization of human factors contributions. For human factors to be effective, human factors engineers need to be involved throughout the design process.

There are a variety of reasons why human factors engineers are not considered as equal partners in a design team. One reason is that other team members (e.g., designers, engineers) have misconceptions about the potential contributions of human factors and the importance of human issues. They perceive, for example, that humans are flexible and can adapt to system requirements or that accommodating human issues will compromise the technical system. Another reason is that sometimes human factors inputs are of limited value to designers (Meister, 1989; Chapanis, 1995). The inputs are either so specific that they apply to a particular design situation and not to the design process in question or vague and overly general. For example, a design guideline which specifies that "older people need larger characters on computer screens" is of little value. How does one define "larger characters"? Obviously, the type of input required depends on the nature of the design problem. Design of a kitchen to accommodate people in wheelchairs requires precise information, such as counter height dimensions or required turning space. In contrast, guidelines for designing intelligent interfaces need to be expressed at the cognitive task level, independent of a particular technology (Woods and Roth, 1988). Thus, one important task for human factors engineers is to ensure that design inputs are in a form that is usable and useful to designers. Williges and colleagues (1992) demonstrate how integrated empirical models can be used as quantitative design guidelines. Their approach involved integrating data from four sequential experiments and developing a model for the design of a telephone-based information system.

To ensure that human factors will be applied to system design systematically, we need to market the potential contributions of human factors to engineers, project managers, and designers. One approach is to use case studies, relevant to the design problem, that illustrate the benefits of human factors. Case studies of this nature can be found in technical journals (e.g., *Applied Ergonomics*, *Ergonomics in Design*) and technical reports. Another approach is to perform a cost–benefit analysis. Estimating the costs and benefits associated with human factors is difficult because it is difficult to isolate the contribution of human factors relative to other variables, baseline measures of performance are unavailable, or performance improvements are hard to quantify and link to system improvements. There are methods available to conduct this type of analysis.

## 3.3 Applications of Human Factors to System Design Process

System design can be conceptualized as a problem-solving process that involves the formulation of the problem, the generation of solutions to the problem, analysis of these alternatives, and selection of the most effective alternative (Rouse, 1985). There are various ways to classify the various stages in system design. Meister (1989), on the basis of a military framework, distinguishes among four phases:

1. *System Planning*. The need for the system is identified and system objectives are defined.
2. *Preliminary Design*. Alternative system concepts are identified, and prototypes are developed and tested.
3. *Detail Design*. Full-scale engineering is developed.
4. *Production and Testing*. The system is built and undergoes testing and evaluation.

To maximize system effectiveness, human factors engineers need to be involved in all phases of the process. In addition to human factors engineers, a representative sample of operators (users) should also be included.

The basic role of human factors in system design is the application of behavioral principles, data, and methods to the design process. Within this role, human factors get involved in a number of activities. These activities include specifying inputs for job, equipment and interface design, human performance criteria, operator selection and training, and inputs regarding testing and evaluation. The nature of these activities is discussed at a general level in the next section. Most of these issues are discussed in detail in subsequent chapters. The intent of this discussion is to highlight the nature of human factors involvement in the design process.

### 3.3.1 System Planning

During system planning, the need for the system is established and the goals and objectives and performance specifications of the system are identified. Performance specifications define what a system must do to meet its objectives and the constraints under which the system will operate. These specifications determine the system's performance requirements. Human factors should be a part of the system planning process. The major role of human factors engineers during this phase is to ensure that human issues are considered in the specification of design requirements and the statement of system goals and objectives. This includes understanding personnel requirements, general performance requirements, the intended users of the system, user needs, and the relationship of system objectives relative to these needs.

### 3.3.2 System Design

System design encompasses both preliminary design and detailed design. During this phase of the process, alternative design concepts are identified and tested

and a detailed model of the system is developed. To ensure adequate consideration of human issues during this phase, the involvement of human factors engineers is critical. The major human factors activities include (1) function allocation, (2) task analysis, (3) job design, (4) interface design, (5) design of support materials, and (6) workplace design. The primary role of the human factors engineer is to ensure joint optimization of the human and technical systems.

**Function Allocation** Function allocation is a critical step in work system design. This is especially true in today's work systems, as machines are becoming more and more capable of performing tasks once restricted to humans. A number of studies have shown (e.g., Morris et al., 1985; Sharit et al., 1987) that proper allocation of functions between humans and machines results in improvements in overall system performance.

Function allocation involves formulating a functional description of a system and subsequent allocation of functions among system components. A frequent approach to function allocation is to base allocation decisions on machine capabilities and to automate wherever possible. Although this approach may appear expedient, there are several drawbacks. In most systems not all tasks can be automated, and thus some tasks must be performed by humans. These tasks are typically "leftover" tasks. Allocating them to humans generally leads to problems of underload, inattention, and job dissatisfaction. A related problem is that automated systems fail and humans have to take over. This can be problematic if the humans are out of the loop or if their skills have become rusty due to disuse. In essence the machine-based allocation strategy is inadequate. As discussed previously, there are numerous examples of technocentered design. It has become clear that a better approach is complementary where functions are allocated so that human operators are complemented by technical systems. This approach involves identifying how to couple humans and machines to maximize system performance. In this regard, there is much research aimed at developing methods to guide function allocation decisions. These methods include lists (e.g., Fitts's list), computer simulation packages, and general guidelines for function allocation (e.g., Price, 1985).

The traditional static approach (humans are better at . . . ) to function allocation has been challenged and dynamic allocation approaches have been developed. With dynamic allocation, responsibility for a task at any particular instance is allocated to the component most capable at that point in time. Hou et al. (1993) developed a framework to allocate functions between humans and computers for inspection tasks. Their framework represents a dynamic allocation framework and provides *for* a quantitative evaluation of the allocation strategy chosen. Morris et al. (1985) investigated the use of a dynamic adaptive allocation approach within an aerial search environment. They found that the adaptive approach resulted in an overall improvement in system performance. Similar to this approach is the adaptive automation approach. This approach involves invoking some form of automation as a function of the person's momentary needs (e.g., transient increase in workload or fatigue). The intent of this approach is to optimize the control of human–machine systems in varying environments. To date, few studies have examined the benefit of this approach. However, several important issues have emerged in the design of these types of systems, such as what aspect of the task should be adapted and who should make the decision to implement or remove automation.

**Task Analysis** Task analysis is also a central activity in system design. Task analysis helps ensure that human performance requirements match operators' (users') needs and capabilities and that the system can be operated in a safe and efficient manner. The output of a task analysis is also essential to the design of the interface, workplaces, support materials, training programs, and test and evaluation procedures.

A task analysis is generally performed after function allocation decisions are made; however, sometimes the results of the task analysis alter function allocation decisions. A task analysis usually consists of two phases: a task description and a task analysis. A *task description* involves a detailed decomposition of functions into tasks which are further decomposed into subtasks or steps. A *task analysis* specifies the physical and cognitive demands associated with each of these subtasks.

A number of methods are available for conducting task analysis. Commonly used methods include flow process charts, critical task analysis, and hierarchical task analysis. Techniques for collecting task data include documentation review, surveys and questionnaires, interviews, observation, and verbal protocols.

As the demands of tasks have changed and become more cognitive in nature, methods have been developed for performing *cognitive task analysis*, which attempts to describe the knowledge and cognitive processes involved in human performance in particular task domains. The results of a cognitive task analysis are important to the design of interfaces for intelligent machines. A common approach used to carry out a cognitive task analysis is a goal–means decomposition. This approach involves an analysis of the work domain to identify the cognitive demands inherent in a particular situation and building a model that relates these cognitive demands to situational demands (Roth et al., 1992). Another approach involves the use of cognitive simulation.

**Job Design** The type of work that a person performs is largely a function of job design. Jobs involve more than tasks and include work content, distribution of work, and work roles. Essentially, a job represents a person's prescribed role within an organization. *Job design* involves determining how tasks will be grouped together, how work will be coordinated among individuals, and how people will be rewarded for their performance (Davis and Wacker, 1987). To design jobs effectively, consideration must be given to workload requirements and to the psychosocial aspects of work (people's needs and expectations). This consideration is especially important in automated work systems, where the skills and potential contributions of humans are often overlooked.

In terms of workload, the primary concern is that work requirements are commensurate with human abilities and individuals are not placed in situations of underload or overload, as both situations can lead to performance decrements, job dissatisfaction, and stress. Both the physical and mental demands of a task need to be considered. There are well-established methods for evaluating the physical demands of tasks and for determination of work and rest schedules. The concept of mental workload is more esoteric. This issue has received a great deal of attention in the literature, and a variety of methods have been developed to evaluate the mental demands associated with a task.

Consideration of operator characteristics is also an essential element of job design, as the workforce is becoming more heterogeneous. For example, older workers may need different work/rest schedules than younger workers or may be unsuited to certain types of tasks. Those who are physically challenged may also require different job specifications.

In terms of psychosocial considerations, a number of studies have identified critical job dimensions. Generally, these dimensions include task variety, task identity, feedback, autonomy, task significance, opportunity to use skills, and challenge. As far as possible, these characteristics should be designed into jobs. Davis and Wacker (1987) have developed a quality-of-working-life-criteria checklist which lists job dimensions important to the satisfaction of individual needs. These dimensions relate to the physical environment, institutional rights and privileges, job content, internal social relations, external social relations, and career path.

A number of approaches to job design have been identified. These include work simplification, job enrichment, job enlargement, job rotation, and teamwork design. The method chosen should depend on the actual design problem, work conditions, and individuals. However, it is generally accepted that the work simplification approach does not lead to optimal job design.

**Interface Design** Interface design involves specification of the nature of the human–machine interaction, that is, the means by which the human is connected to the machine. During this stage of design, the human factors specialist typically works closely with engineers and designers. The role of human factors is to provide the design team with information regarding the human performance implications of design alternatives. This generally involves three major activities: (1) gathering and interpreting human performance data, (2) conducting attribute evaluations of suggested designs, and (3) human performance testing (Sanders and McCormick, 1993). Human performance testing typically involves building mock-ups and prototypes and testing them with a sample of users. This type of testing can be expensive and time consuming. Recently, the development of rapid prototyping tools has made it possible to speed up and compress this process. These tools have been used primarily in the testing of computer interfaces; however, they can be applied to a variety of situations.

Interface design encompasses the design of both the physical and cognitive components of the interface and includes the design and layout of controls and displays, information content, and information representation. Physical components include factors such as type of control or input device, size and shape of controls, control location, and visual and auditory specifications (e.g., character size, character contrast, labeling, signal rate, signal frequency).

Cognitive components refer to the information-processing aspects of the interface (e.g., information content, information layout). As machines have become more intelligent, much of the focus of interface design has been on the cognitive aspects of the interface: Issues of concern include determination of the optimal level of machine support, identification of the type of information that users need, determination of how this information should be presented, and identification of methodologies to analyze work domains and cognitive activities. The central concern is developing interfaces that best support human task performance. In this regard, a number of approaches have evolved for interface design. Ecological interface design (Rasmussen and Vincente, 1989) is an example of a recent design method.

There are a variety of sources of data on the characteristics of human performance that can serve as inputs to the design process. These include handbooks, textbooks, standards [e.g., American National Standards Institute (ANSI)], and technical journals. There are also a variety of models of human performance, including cognitive models (e.g., GOMS; Card et al., 1983), control theory models, and engineering models. These models can be useful in terms of predicting the effects of design parameters on human performance outcomes. As discussed previously, it is the responsibility of the human factors engineer to make sure that information regarding human performance is in a form that is useful to designers. It is also important when using these data to consider the nature of the task, the task environment, and the user population.

**Design of Support Materials** This phase of the design process includes identifying and developing materials that facilitate the user's interaction with the system. These materials include job aids, instructional materials, and training devices and programs. All too often this phase of the design process is neglected or given little attention. A common example is the cumbersome manuals that accompany software packages or VCRs.

Support materials should not be used as a substitute for good design, however; the design of effective support materials is an important part of the system design process. Users typically need training and support to interact successfully with new technologies and complex systems. To maximize their effectiveness, human factors principles need to be applied to the design of instructional materials, job aids, and training programs. Guidelines are available for the design of instructional materials and job aids. Bailey (1982) provides a thorough discussion of these issues. A great deal has also been written on the design of training programs.

**Design of Work Environment** The design of the work environment is an important aspect of work system design. Systems exist within a context, and

the characteristics of this context affect overall system performance. The primary concern of workplace design is to ensure that the work environment supports the operator and activity performance and allows the worker to perform tasks in an efficient, comfortable, and safe manner. Important issues include workplace and equipment layout, furnishings, reach dimensions, clearance dimensions, visual dimensions, and the design of the ambient environment. There are numerous sources of information related to workplace design and evaluation that can be used to guide this process. These issues are also covered in detail in other chapters of this handbook.

### 3.4 Test and Evaluation

Test and evaluation are critical aspects of system design and usually take place throughout the system design process. Test and evaluation provide a means for continuous improvement during system development. Human factors inputs are essential to the testing and evaluation of systems. The primary role of human factors is to assess the impact of system design features on human performance outputs, including objective outputs such as speed and accuracy of performance and workload, and subjective outputs such as comfort and user satisfaction. Human factors specialists are also interested in ascertaining the impact of human performance on overall system performance. Issues related to the evaluation and assessment of system effectiveness are covered in detail in Chapters 41.

Because the evaluation of systems and system components involves measurement of human performance in operational terms (relative to the system or subsystem in question), human factors engineers face a number of challenges when evaluating systems. Generally, the standards of generalizability are higher for human factors research, as the research results must be extended to real-world systems (Kantowitz, 1992). At the same time, it is often difficult to achieve an appropriate level of control. Unfortunately, in many instances the utility of a test and the evaluation results are limited because of deficiencies in the test and evaluation procedures (Bitner, 1992).

In this regard, there are three key issues that need to be addressed when developing methods for evaluating system effectiveness: (1) subject representativeness, (2) variable representativeness, and (3) setting representativeness (Kantowitz, 1992). *Subject representativeness* refers to the extent to which subjects tested in the research study represent the population to which the research results apply. In most cases, the sample involved in system evaluation should represent the population of interest on relevant characteristics. *Variable representativeness* refers to the extent that the study variables are representative of the research question. It is important to select variables that capture the essential issues being assessed in the research study. *Setting representativeness* is the degree of congruence between the test situation in which the research is performed and the target situation in which the research must be applied. The important issue is the comparability of the

psychological processes captured in these situations, not necessarily physical fidelity.

A variety of techniques are available for conducting human factors research, including experimental methods, observational methods, surveys and questionnaires, and audits. There is no single preferred method; each has its associated strengths and weaknesses. The method one chooses depends on the nature of the research question. It is generally desirable to use several methods in conjunction.

## 4 CONCLUSIONS

System design and development represent an important area of application for human factors engineers. System performance will be improved by consideration of behavioral issues. Although much has been written on system design, our knowledge of this topic is far from complete. The changing nature and complexity of systems coupled with the increased diversity of end users present new challenges for human factors specialists and afford many research opportunities. The goals of this chapter were to summarize some of the current issues in system design and to illustrate the important role of human factors engineers within the system design process. Further, the chapter provides a framework for many of the topics addressed in this handbook.

## REFERENCES

Bailey, R. W. (1982), *Human Performance Engineering: A Guide for System Design*, Prentice-Hall, Englewood Cliffs, NJ.

Banathy, B. H., and Jenlink, P. M. (2004), "Systems Inquiry and Its Application in Education," in *Handbook of Research on Educational Communications and Technology*, D. H. Jonassen, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 37–55.

Bitner, A. C. (1992), "Robust Testing and Evaluation of Systems," *Human Factors*, Vol. 34, pp. 477–484.

Carayon, P. (2006), "Human Factors of Complex Sociotechnical Systems," *Applied Ergonomics*, Vol. 37, pp. 525–534.

Card, S. K., Moran, T. P., and Newell, A. (1983), *The Psychology of Human–Computer Interaction*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Chapanis, A. (1995). "Ergonomics in Product Development: A Personal View," *Ergonomics*, Vol. 38, pp. 1625–1638.

Clegg, C. W. (2000), "Sociotechnical Principles for System Design," *Applied Ergonomics*, Vol. 31, pp. 463–477.

Czaja, S. J., and Sharit, J. (2009), "Preparing Organizations and Older Workers for Current and Future Employment: Training and Retraining Issues," in *Aging and Work: Assessment and Implications for the Future*, S. J. Czaja and J. Sharit, Eds., Johns Hopkins University Press, Baltimore, MD, pp. 259–278.

Davis, L. E., and Wacker, G. J. (1987), "Job Design," in *Handbook of Human Factors*, G. Salvendy, Ed., Wiley, New York, pp. 431–452.

De Vreede, G., Briggs, R.O., and Reiter-Palmon, R. (2010), "Exploring Asynchronous Brainstorming in Large Groups: A Field Comparison of Serial and Parallel Subgroups," *Human Factors*, Vol. 52, No. 2, pp. 189–202.

Eason, K. D. (1989), *Information Technology and Organizational Change*, Taylor & Francis, London.

Eason, K. D. (1991), "Ergonomic Perspectives on Advances in Human–Computer Interaction," *Ergonomics*, Vol. 34, pp. 721–741.

Fass, D., and Leiber, R. (2009), "Rationale for Human Modelling in Human in the Loop Systems Design," paper presented at IEEE SysCon 2009—3rd Annual IEEE International Systems Conference, Vancouver, Canada, March 23–26, 2009.

Fiore, S. M, Rosen, M. A., Smith-Jentsch, K., Salas, E., Letsky, M., and Warner, N. (2010), "Toward an Understanding of Macrocognition in Teams: Predicting Processes in Complex Collaborative Contexts," *Human Factors*, Vol. 52, pp. 203–224.

Gorman, J. C., Cooke, N. J., and Salas, E. (2010), "Preface to the Special Issue on Collaboration, Coordination, and Adaptation in Complex Sociotechnical Settings," *Human Factors*, Vol. 52, Vol. 2, pp. 143–146.

Hendrick, H. W., and Kleiner, B. M. (2001), *Macro Ergonomics: An Introduction to Work System Design*, Human Factors and Ergonomics Society, Santa Monica, CA.

Hobbs, F., and Stoops, N. (2003), "Demographic Trends in the 20th Century," U.S. Census Bureau, Census 2000 Special Reports, Series CENSR-4, U.S. Government Printing Office, Washington, DC, 2002, available: http://www.census.gov/prod/2002pubs/censr-4.pdf, accessed November 10, 2010.

Hou, T., Lin, L., and Drury, C. G. (1993), "An Empirical Study of Hybrid Inspection Systems and Allocation of Inspection Function," *International Journal of Human Factors in Manufacturing*, Vol. 3, pp. 351–367.

Ilmarinen, J. (2009), "Work Ability_A Comprehensive Concept for Occupational Health Research and Prevention," *Scandinavian Journal of Work Environment and Health*, Vol. 35, pp. 1–5.

Imada, A. S., Nora, K., and Nagamachi, M. (1986), "Participatory Ergonomics: Methods for Improving Individual and Organizational Effectiveness," in *Human Factors in Organizational Design and Management*, Vol. II, O. Brown, Jr., and H. W. Hendrick, Eds., North-Holland, New York, pp. 403–406.

International Ergonomics Association (IEA) (2010), *What Is Ergonomics*, available: http://www.iea.cc/01_what/What is Ergonomics.html, accessed November 12, 2010.

Kantowitz, B. H. (1992), "Selecting Measures for Human Factors Research," *Human Factors*, Vol. 34, pp. 387–398.

Kidd, P. T. (1992), "Interdisciplinary Design of Skill–Based Computer-Aid Technologies: Interfacing in Depth," *International Journal of Human Factors in Manufacturing*, Vol. 2, No. 3, pp. 209–228.

Kleiner, B. M. (2008), "Macroegonomics: Work System Analysis and Design," *Human Factors*, Vol. 50, No. 3, pp. 461–467.

Lämkull, D., Hanson, L., and Örtengren, R. (2007), "The Influence of Virtual Human Model Appearance on Visual Ergonomics Posture Evaluation," *Applied Ergonomics*, Vol. 38, pp. 713–722.

Liker, J. K., and Majchrzak, A. (1994), "Designing the Human Infrastructure for Technology," in *Organization and Management of Advanced Manufacturing*, W. Karwowski and G. Salvendy, Eds., Wiley, New York, pp. 121–164.

Lim, K. Y., Long, J. B., and Hancock, N. (1992), "Integrating Human Factors with the Jackson System Development Method: An Illustrated Overview," *Ergonomics*, Vol. 35, pp. 1135–1161.

Liu, K., Valerdi R., and Rhodes, D. H. (2009), "Economics of Human Systems Integration: The Pratt & Whitney F119 Engine," paper presented at the Conference on Systems Engineering Research, Loughborough University, United Kingdom, April 2009.

Manzey, D., et al. (2009), "Image-Guided Navigation: The Surgeon's Perspective on Performance Consequences and Human Factors Issues," *International Journal of Medical Robotics*, Vol. 5, pp. 297–308.

Meister, D. (1989), *Conceptual Aspects of Human Factors*, John Hopkins University Press, Baltimore, MD.

Meister, D. (1991), *Psychology of System Design*, Elsevier, New York.

Meister, D. (2000), "Cognitive Processes in System Design," *Theoretical Issues in Ergonomics Science*, Vol. 2, pp. 113–138.

Morris, N. M., Rouse, W. B., and Ward, S. L. (1985), "Experimental Evaluation of Adaptive Task Allocation in an Aerial Search Environment," in *Proceedings of the 2nd IFAC/IFIP/IFORS/IEA Conference: Analysis, Design and Evaluation of Man–Machine Systems*, G. Mancini, G. Johannsen, and L. Martensson, Eds., Varese, Italy, pp. 67–72, Pergamon: Varese, Italy.

Morrow, D. G., North, R. A. & Wickens, C. D. (2006) **Reducing and mitigating human error in medicine** In R. Nickerson (Ed.), Annual Review of Human Factors and Ergonomics, vol. 1 (pp. 254–296). Santa Monica, CA: Human Factors & Ergonomics Society.

Parasuraman, R., and Wilson, G. (2008), "Putting the Brain to Work: Neuroergonomics Past, Present, and Future," *Human Factors*, Vol. 50, pp. 468–474.

Pasmore, W. A. (1988), *Designing Effective Organizations: The Sociotechnical Systems Perspective*, Wiley, New York.

Pew, R. W. (2008, June), "More Than 50 Years of History and Accomplishments in Human Performance Model Development," *Human Factors: Journal of the Human Factors and Ergonomics Society*, Vol. 50, No. 3, pp. 489–496.

Pew, R. W., and Mavor, A. S. (2007), *Human-System Integration in the System Development Process: A New Look*, National Academy Press, Washington, DC, 2007.

Price, H. E. (1985), "The Allocation of Functions in Systems," *Human Factors*, Vol. 2, pp. 33–45.

Proctor, R., and van Zandt, T. (1994), *Human Factors in Simple and Complex Systems*, Allyn & Bacon, Needham Heights, MA.

Proctor, R. W., and Vu, K.-P. L. (2010), "Cumulative Knowledge and Progress in Human Factors," *Annual Review of Psychology*, Vol. 61, pp 623–651.

Rasmussen, J. (1982), "Human Errors: A Taxonomy for Describing Human Malfunctioning in Industrial Installations," *Journal of Occupational Accidents*, Vol. 4, pp. 311–333.

Rasmussen, J. (1986), *Information Processing and Human–Machine Interaction: An Approach to Cognitive Engineering*, North-Holland, Amsterdam.

Rasmussen, J., and Vincente, K. J. (1989), "Coping with Human Errors through System Design: Implications for Ecological Interface Design," *International Journal of Man–Machine Studies*, Vol. 31, pp. 517–534.

Rasmussen, J., Pejtersen, A. M., and Goodstein, L. P. (1994), *Cognitive Systems Engineering*, Wiley, New York.

Reason, J. (1990), *Human Error*, Cambridge University Press, New York.

Reason, J. (1995), "A Systems Approach to Organizational Error," *Ergonomics*, Vol. 38, pp. 1708–1721.

Roth, E. M., Woods, D. D., and Pople, H. E., Jr. (1992), "Cognitive Simulation as a Tool for Cognitive Task Analysis," *Ergonomics*, Vol. 35, pp. 1163–1198.

Rouse, W. B. (1985), "On the Value of Information in System Design: A Framework for Understanding and Aiding Designers," *Information Processing and Management*, Vol. 22, pp. 217–228.

Rouse, W. B. (1987), "Designers, Decision Making and Decision Support," in *System Design: Behavioral Perspectives on Designers, Tools and Organizations*, W. B. Rouse and K. R. Boff, Eds., North-Holland, New York, pp. 275–284.

Rouse, W. B., and Boff, K. R. (1987a), "Workshop Themes and Issues: The Psychology of Systems Design," in *System Design: Behavioral Perspectives on Designers, Tools and Organizations*, W. B. Rouse and K. R. Boff, Eds., North-Holland, New York, pp. 7–18.

Rouse, W. B., and Boff, K. R. (1987b), *System Design: Behavioral Perspectives on Designers, Tools and Organizations*, North-Holland, New York.

Rouse, W. B., and Morris, N. M. (1987), "Conceptual Design of Error Tolerant Interface for Complex Engineering Systems," *Automatica*, Vol. 23, pp. 231–235.

Sage, A. P., and Rouse, W. B. (2009), "An Introduction to Systems Engineering and Systems Management," in *Handbook of Systems Engineering and Management*, 2nd ed., A. P. Sage and W. B. Rouse, Eds., Wiley, Hoboken, NJ, pp. 1–63.

Salas, E., Cooke, N. J., and Rosen, M. A. (2008), "On Teams, Teamwork, Team Performance: Discoveries and Developments," *Human Factors*, Vol. 50, No. 3, pp 540–547.

Sanders, M. S., and McCormick, E. J. (1993), *Human Factors in Engineering and Design*, 7th ed., McGraw-Hill, New York.

Schulz, R., and Martire, L. M. (2009), "Caregiving and Employment," in *Aging and Work: Assessment and Implications for the Future*, S. J. Czaja and J. Sharit, Eds., Johns Hopkins University Press, Baltimore, MD, pp. 35–50.

Shah, J., and Breazeal, C. (2010), "An Empirical Analysis of Team Coordination Behaviors and Action Planning with Application to Human-Robot Teaming," *Human Factors*, Vol. 52, pp. 234–245.

Sharit, J., Chang, J. C., and Salvendy, G. (1987), "Technical and Human Aspects of Computer-Aided Manufacturing,"

in *Handbook of Human Factors*, G. Salvendy, Ed., Wiley, New York, pp. 1694–1724.

Sheridan, T. B. (2002), *Humans and Automation: System Design and Research Issues*, Human Factors and Ergonomics Society, Santa Monica, CA.

Sheridan, T. B. (2008), "Risk, Human Error, and System Resilience: Fundamental Ideas," *Human Factors*, Vol. 50, No. 3, pp. 418–426.

Strauch, B. (2010), "Can Cultural Differences Lead to Accidents? Team Cultural Differences and Sociotechnical System Operations," *Human Factors*, Vol. 52, pp. 246–263.

Toossi, M. (2007), "Employment Outlook: 2006–16: Labor Force Projections to 2016: More Workers in Their Golden Years," *Monthly Labor Review*, Vol. 130, No. 11, pp. 33–52.

U.S. Census Bureau, Population Division (2008), "Projections of the Population by Age and Sex for the United States: 2010–2050," NP2008-T12, available: http://www.census.gov/population/www/projections/summarytables.html, accessed November 10, 2010.

Vincente, K. J. (2002), "Ecological Interface Design: Progress and Challenges," *Human Factors*, Vol. 44, No. 1, pp. 62–78.

Wickens, C. D. (1992), *Engineering Psychology and Human Performance*, 2nd ed., Harper-Collins, New York.

Wickens, C. D., and Hollands, J. G. (2000), *Engineering Psychology and Human Performance*, 3rd ed., Prentice-Hall, Upper Saddle River, NJ.

Wickens, C. D., Lee, J., Liu, Y. D., and Gordon-Becker, S. (2004), *Introduction to Human Factors Engineering*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ.

Williges, R. C., Williges, B. H., and Han, S. (1992), "Developing Quantitative Guidelines Using Integrated Data from Sequential Experiments," *Human Factors*, Vol. 34, pp. 399–408.

Woods, D. D. (1988), "Commentary: Cognitive Engineering in Complex and Dynamic Worlds," in *Complex Engineering in Complex Dynamic Worlds*, E. Hollnagel, G. Mancini, and D. D. Woods, Eds., Academic, New York, pp. 115–129.

Woods, D. D., and Roth, E. M. (1988), "Cognitive Systems Engineering," In *Handbook of Human–Computer Interaction*, M. E. Helander, Ed., North-Holland, New York, pp. 3–43.

# HUMAN FACTORS FUNDAMENTALS

# CHAPTER 3

# SENSATION AND PERCEPTION

**Robert W. Proctor and Janet D. Proctor**
**Purdue University**
**West Lafayette, Indiana**

*Whether performing a simple skill like standing and looking, or something more exotic . . . , your perceptual experiences are fundamentally multisensory.*

L. D. Rosenblum (2010)

## 1 INTRODUCTION

Human–machine interaction, as all other interactions of persons with their environment, involves a continuous exchange of information between the operator(s) and the machine. The operator provides input to the machine, which acts on this input and displays information back to the operator regarding its status and the consequences of the input. The operator must process this information, decide what, if any, controlling actions are needed, and then provide new input to the machine. One important facet of this exchange of information between the machine and the operator is the displaying of information from the machine as input to the operator. All such information must enter through the operator's senses and be organized and recognized accurately to ensure correct communication of the displayed information. Thus, an understanding of how people sense and perceive is essential for display design. An effective display is consistent with the characteristics and limitations of the human sensory and perceptual systems. These systems are also involved intimately in both the control of human interactions with the environment and the actions taken to operate machines. However, because the selection and control of action are the topics of Chapter 4, in this chapter we focus primarily on the nature of sensation and perception. Similarly, because other chapters focus on the applied topics of motion and vibration (Chapter 22), noise (Chapter 23), illumination (Chapter 24), and displays (Chapter 42), we concentrate primarily on the nature of sensory and perceptual processes and the general implications for human factors and ergonomics.

Many classifications of sensory systems exist, but most commonly, distinctions are made between five sensory modalities: vision, audition, olfaction, gustation, and somasthesis. The vestibular system, which provides the sense of balance, is also of importance in many areas of human factors and ergonomics. Although the peripheral aspects of these sensory systems are distinct, the senses interact extensively in creating our perceptual experiences, as implied by Rosenblum's (2010) quote with which the chapter begins.

All sensory systems extract information about four characteristics of stimulation: (1) the sensory modality and submodalities (e.g., touch as opposed to pain), (2) the stimulus intensity, (3) the duration of the stimulation, and (4) its location (Gardner and Martin, 2000). Each system has receptors that are sensitive to some aspect of the physical environment. These receptors are responsible for sensory transduction, or the conversion of physical stimulus energy into electrochemical energy in the nervous system. Their properties are a major factor determining the sensitivity to stimulation. After sensory transduction, the sensory information for each sense is encoded in the activity of neurons and travels to
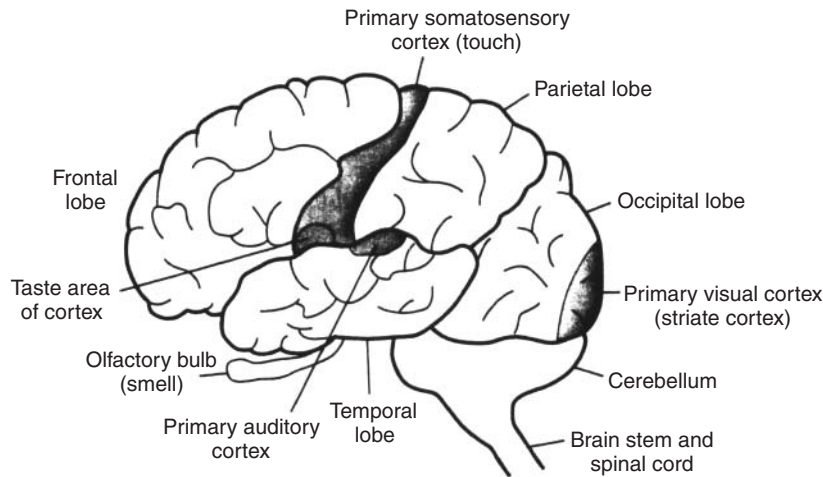
**Figure 1** Primary sensory receiving areas (visual, auditory, and somatosensory) of the cerebral cortex and other important landmarks and areas. (From Schiffman, 1996.)

the brain via specialized, structured pathways consisting of highly interconnected networks of neurons. For most modalities, two or more pathways operate in parallel to analyze and convey different types of information from the sensory signal. The pathways project to primary receiving areas in the cerebral cortex (see Figure 1), in most cases after passing through relay areas in the thalamus. From the primary receiving area, the pathways then project to many other areas within the brain.

Each neuron in the sensory pathways is composed of a cell body, dendrites at the input side, and an axon with branches at the output side. Most neurons fire in an all-or-none manner, sending spike, or action, potentials down the axon away from the cell body. The rate at which a neuron fires varies as a function of the input that the neuron is receiving from other neurons (or directly from sensory receptors) at its dendrites. Most neurons exhibit a baseline firing rate in the absence of stimulation, usually on the order of 5–10 spikes/s, and information is signaled by deviations above or below this baseline rate. The speed of transmission of a spike along the fiber varies across different types of neurons, ranging from 20 to 100 m/s. Immediately after an action potential occurs, the neuron is in a refractory state in which another action potential cannot be generated. This sets an upper limit on the firing rate of about 1000 spikes/s.

Transmission between neurons occurs at small gaps, called *synapses*, between the axonal endings of one neuron and the dendrites of another. Communication at the synapse takes place by means of transmitter substances that have an excitatory effect of increasing the firing rate of the neuron or an inhibitory effect of decreasing the firing rate. Because as many as several hundred neurons may have synapses with the dendrites of a specific neuron, whether the firing rate will increase or decrease is a function of the sum of the excitatory and inhibitory inputs that the neuron is receiving. Which specific neurons provide excitatory and inhibitory inputs

will determine the patterns of stimulation to which the neuron will be sensitive (i.e., to which the firing rate will increase or decrease from baseline). The patterns may be rather general (e.g., an increase in illumination) or quite specific (e.g., a pair of lines at a particular angle moving in a particular direction). In general, increases in stimulus intensity result in increased firing rates for individual neurons and in a larger population of neurons that respond to the stimulus. Thus, intensity is coded by firing rate (as well as possibly relative timing of spike potentials) and population codes.

The study of sensation and perception involves not only the anatomy and physiology of the sensory systems but also behavioral measures of perception. Psychophysical data obtained from tasks in which observers are asked to detect, discriminate, rate, or recognize stimuli provide information about how the properties of the sensory systems relate to what is perceived. Behavioral measures also provide considerable information about the functions of the higher level brain processes involved in perception. The sensory information must be interpreted and organized by these higher level processes, which include mental representations, decision making, and inference. Thus, perceptual experiments provide evidence about how the input from the various senses is organized into a coherent percept.

## 2 METHODS FOR INVESTIGATING SENSATION AND PERCEPTION

Many methods have been, and can be, used to obtain data relevant to understanding sensation and perception (see, e.g., Scharff, 2003). The most basic distinction is between methods that involve anatomy and physiology as opposed to methods that involve behavioral responses. Because the former are not of much direct use in human factors and ergonomics, we do not cover them in as much detail as we do the latter.

## 2.1 Anatomical and Physiological Methods

A variety of specific techniques exist for analyzing and mapping out the pathways associated with sensation and perception. These include injecting tracer substances into the neurons, classifying neurons in terms of the size of their cell bodies and characteristics of their dendritic trees, and lesioning areas of the brain (see Wandell, 1995). Such techniques have provided a relatively detailed understanding of the sensory pathways.

One particular technique that has produced a wealth of information about the functional properties of specific neurons in the sensory pathways and their associated regions in the brain is *single-cell recording*. Such recording is typically performed on a monkey, cat, or other nonhuman species; an electrode is inserted that is sufficiently small to record only the activity of a single neuron. The responsivity of this neuron to various features of stimulation can be examined to gain some understanding of the neuron's role in the sensory system. By systematic examination of the responsivities of neurons in a given region, it has been possible to determine much about the way that sensory input is coded. In our discussion of sensory systems, we will have the opportunity to refer to the results of single-cell recordings.

Neuropsychological and psychophysiological investigations of humans have been used increasingly in recent years to evaluate issues pertaining to information processing. Neuropsychological studies typically examine patients who have some specific neurological disorder associated with lesions in particular parts of the brain. Several striking phenomena have been observed that enhance our understanding of higher level vision (Farah, 2000). One example is visual neglect, in which a person with a lesion in the right cerebral hemisphere, often in a region called the *right posterior parietal lobe*, fails to detect or respond to stimuli in the left visual field (Mort et al., 2003). This is in contrast to people with damage to regions of the temporal lobe, who have difficulty recognizing stimuli (Milner and Goodale, 1995). These and other results have provided evidence that a dorsal system, also called the *parietal pathway*, determines where something is (and how to act on it), whereas a ventral system, also called the *temporal pathway*, determines what that something is (Merigan and Maunsell, 1993).

A widely used psychophysiological method involves the measurement of *event-related potentials* (ERPs) (Handy, 2005). To record ERPs, electrodes attached to a person's scalp measure voltage variations in the electroencephalogram (EEG), which reflects the summed electrical activity of neuron populations as recorded at various sites on the scalp. An ERP is those changes that involve the brain's response to a particular event, usually onset of a stimulus. ERPs provide good temporal resolution, but the spatial resolution is not very high. Those ERP components occurring within 100 ms after onset of a stimulus are sensory components that reflect transmission of sensory information to, and its arrival at, the sensory cortex. The latencies for these components differ across sensory modalities. Later components reflect other aspects of information processing. For example, a negative component called

*mismatch negativity* is evident in the ERP about 200 ms after presentation of a stimulus event other than the one that is most likely. It is present regardless of whether the stimulus is in an attended stream of stimuli or an unattended stream, suggesting that it reflects an automatic detection of physical deviance. The latency of a positive component called the *P*300 is thought to reflect stimulus evaluation time, that is, the time to update the perceiver's current model of the physical environment.

During the past 15 years, use of functional neuroimaging techniques such as *functional magnetic resonance imaging* (fMRI) and *positron emission tomography* (PET), which provide insight into the spatial organization of brain functions, has become widespread (Kanwisher and Duncan, 2004; Hall, 2010). Both fMRI and PET provide images of the neural activity in different areas of the brain by measuring the volume of blood flow, which increases as the activity in an area increases. They have good spatial resolution, but the temporal resolution is not as good as that of ERPs. By comparing measurements taken during a control period to those taken while certain stimuli are present or tasks performed, the brain-imaging techniques can be used to identify which areas of the brain are involved in the processing of different types of stimuli and tasks.

Electrophysiological and functional imaging methods, as well as other psychophysiological techniques, provide tools that can be used to address many issues of concern in human factors. Among other things, these methods can be used to determine whether a particular experimental phenomenon has its locus in processes associated with sensation and perception or with those involving subsequent response selection and execution. Because of this diagnosticity, it has been suggested that psychophysiological measures may be applied to provide precise measurement of dynamic changes in mental workload (e.g., Wilson, 2002) and to other problems in human factors (Kramer and Weber, 2000). The term *neuroergonomics* is used to refer to a neuroscience approach to ergonomics (e.g., Lees et al., 2010).

## 2.2 Psychophysical Methods

The more direct concern in human factors and ergonomics is with behavioral measures, because our interest is primarily with what people can and cannot perceive and with evaluating specific perceptual issues in applied settings. Because many of the methods used for obtaining behavioral measures can be applied to evaluating aspects of displays and other human factors concerns, we cover them in some detail. The reader is referred to textbooks on psychophysical methods by Gescheider (1997) and Kingdom and Prins (2010) and to chapters by Schiffman (2003) and Rose (2006) for more thorough coverage.

### 2.2.1 Psychophysical Measures of Sensitivity

**Classical Threshold Methods** The goal of one class of psychophysical methods is to obtain some estimate of sensitivity to detecting either the presence of some stimulation or differences between stimuli. The classical methods were based on the concept of a threshold, with an *absolute threshold* representing the minimum amount

**Table 1  Determination of Sensory Threshold by Method of Limits Using Alternating Ascending (A) and Descending (D) Series**

| Stimulus Intensity (Arbitrary Units) | A | D | A | D | A | D | A | D | A | D |
|---|---|---|---|---|---|---|---|---|---|---|
| 15 |  |  |  |  |  | Y |  |  |  |  |
| 14 |  |  |  | Y |  | Y |  |  |  | Y |
| 13 |  |  |  | Y |  | Y |  |  |  | Y |
| 12 |  | Y |  | Y |  | Y |  |  |  | Y |
| 11 |  | Y |  | Y | Y | Y |  | Y |  | Y |
| 10 | Y | Y |  | Y | N | Y | Y | Y |  | Y |
| 9 | N | Y | Y | N | N | N | N | Y | Y | N |
| 8 | N | N | N |  | N |  | N | Y | N |  |
| 7 | N |  | N |  | N |  | N | N | N |  |
| 6 | N |  | N |  | N |  | N |  | N |  |
| 5 | N |  |  |  | N |  | N |  | N |  |
| 4 | N |  |  |  | N |  |  |  | N |  |
| 3 | N |  |  |  |  |  |  |  | N |  |
| 2 | N |  |  |  |  |  |  |  | N |  |
| 1 | N |  |  |  |  |  |  |  |  |  |
| Transition points[a] | 9.5 | 8.5 | 8.5 | 9.5 | 10.5 | 9.5 | 9.5 | 7.5 | 8.5 | 9.5 |

[a]Mean threshold value = 9.1.

of stimulation necessary for an observer to tell that a stimulus was presented on a trial, and a *difference threshold* representing the minimal amount of difference in stimulation along some dimension required to tell that a comparison stimulus differs from a standard stimulus. Fechner (1860) developed several techniques for finding absolute thresholds, with the methods of limits and constant stimuli being among the most widely used.

To find a threshold using the *method of limits*, equally spaced stimulus values along the dimension of interest (e.g., magnitude of stimulation) that bracket the threshold are selected (see Table 1). In alternating series, the stimuli are presented in ascending or descending order, beginning each time from a different, randomly chosen starting value below or above the threshold. For the ascending order, the first response typically would be, "No, I do not detect the stimulus." The procedure is repeated, incrementing the stimulus value each time, until the observer's response changes to "yes," and the average of that stimulus value and the last one to which a "no" response was given is taken as the threshold for that series. A descending series is conducted in the same manner, but from a stimulus above threshold, until the response changes from yes to no. The thresholds for the individual series are then averaged to produce the final threshold estimate. A particularly efficient variation of the method of limits is the *staircase* method (Cornsweet, 1962). For this method, rather than having distinct ascending and descending series started from randomly selected values below and above threshold, only a single continuous series is conducted in which the direction of the stimulus sequence—ascending or descending—is reversed when the observer's response changes. The threshold is then taken to be the average of the stimulus

values at which these transitions occur. The staircase method has the virtue of bracketing the threshold closely, thus minimizing the number of stimulus presentations that is needed to obtain a certain number of response transitions on which to base the threshold estimate.

The *method of constant stimuli* differs from the method of limits primarily in that the different stimulus values are presented randomly, with each stimulus value presented many different times. The basic data in this case are the percentage of yes responses for each stimulus value. These typically plot as an S-shaped psychophysical function (see Figure 2). The threshold is taken to be the estimated stimulus value for which the percentage of yes responses would have been 50%.

Both the methods of limits and constant stimuli can be extended to difference thresholds in a straightforward manner (see Gescheider, 1997). The most common extension is to use stimulus values for the comparison stimulus that range from being distinctly less than that of the standard stimulus to being distinctly greater. For the method of limits, ascending and descending series are conducted in which the observer responds "less," then "equal," and then "greater" as the magnitude of the comparison increases, or vice versa as it decreases, The average stimulus value for which the responses shift from less to equal is the lower threshold, and from equal to greater is the upper threshold, The difference between these two values is called the *interval of uncertainty*, and the difference threshold is found by dividing the interval of uncertainty by 2. The midpoint of this interval is the point of subjective equality, and the difference between this point and the true value of the standard stimulus reflects the constant error, or the influence of any factors that cause the observer to overestimate or underestimate
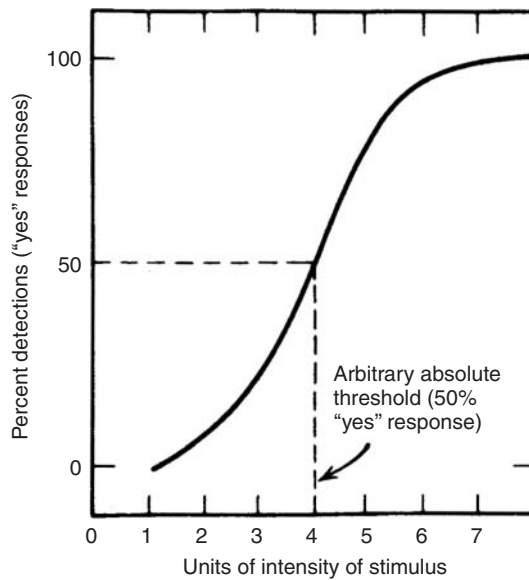
**Figure 2** Typical S-shaped psychophysical function obtained with the method of constant stimuli. The absolute threshold is the stimulus intensity estimated to be detected 50% of the time. (From Schiffman, 1996.)

systematically the value of the comparison in relation to that of the standard.

When the method of constant stimuli is used to obtain difference thresholds, the order in which the standard and comparison are presented is varied, and the observer judges which stimulus is greater than the other. The basic data then are the percentages of "greater" responses for each value of the comparison stimulus. The stimulus value corresponding to the 50th percentile is taken as the point of subjective equality. The difference between that stimulus value and the one corresponding to the 25th percentile is taken as the lower difference threshold, and the difference between the subjectively equal value and the stimulus value corresponding to the 75th percentile is the upper threshold: The two values are averaged to get a single estimate of the difference threshold.

Although threshold methods are often used to investigate basic sensory processes, variants can be used to investigate applied problems as well. Shang and Bishop (2000) argued that the concept of visual threshold is of value for measuring and monitoring landscape attributes. They measured three types of different thresholds—detection, recognition, and visual impact (changes in visual quality as a consequence of landscape modification)—for two types of objects, a transmission tower and an oil refinery tank, as a function of size, contrast, and landscape type. Shang and Bishop were able to obtain thresholds of high reliability and concluded that a visual variable that combined the effects of contrast and size, which they called *contrast weighted visual size*, was the best predictor of all three thresholds.

***Signal Detection Methods*** Although many variants of the classical methods are still used, they are not

as popular as they once were. The primary reason is that the threshold measures confound perceptual sensitivity, which they are intended to measure, with response criterion or bias (e.g., willingness to say yes), which they are not intended to measure. The threshold estimates can also be influenced by numerous other extraneous factors, although the impact of most of these factors can be minimized with appropriate control procedures. Alternatives to the classical methods, signal detection methods, have come to be preferred in many situations because they contain the means for separating sensitivity and response bias. Authoritative references for signal detection methods and theory include Green and Swets (1966), Macmillan and Creelman (2005), and Wickens (2001). Macmillan (2002) provides a briefer introduction to its principles and assumptions.

The typical signal detection experiment differs from the typical threshold experiment in that only a single stimulus value is presented for a series of trials, and the observer must discriminate trials on which the stimulus was not presented (noise trials) from trials on which it was (signal-plus-noise, or signal, trials). Thus, the signal detection experiment is much like a true–false test in that it is objective; the accuracy of the observer's responses with respect to the state of the world can be determined. If the observer says yes most of the time on signal trials and no most of the time on noise trials, we know that the observer was able to discriminate between the two states of the world. If, on the other hand, the proportion of yes responses is equal on signal and noise trials, we know that the observer could not discriminate between them. Similarly, we can determine whether the observer has a bias to say one response or the other by considering the relative frequencies of yes and no responses regardless of the state of the world. If half of the trials included the signal and half did not, yet the observer said yes 70% of the time, we know that the observer had a bias to say yes.

Signal detection methods allow two basic measures to be computed, one corresponding to discriminability (or *sensitivity)* and the other to *response bias*. Thus, the key advantage of the signal detection methods is that they allow the extraction of a pure measure of perceptual sensitivity separate from any response bias that exists, rather than combining the two in a single measure, as in the threshold techniques. There are many alternative measures of sensitivity (Swets, 1986) and bias (Macmillan and Creelman, 1990), based on a variety of psychophysical models and assumptions. We will base our discussion around signal detection theory and the two most widely used measures of sensitivity and bias, $d'$ and $\beta$. Sorkin (1999) describes how signal detection measures can be calculated using spreadsheet application programs such as Excel.

Signal detection theory assumes that the sensory effect of a signal or noise presentation on any given trial can be characterized as a point along a continuum of evidence indicating that the signal was in fact presented. Across trials, the evidence will vary, such that for either type of trial it will sometimes be higher (or lower) than at other times. For computation of $d'$ and $\beta$, it is assumed that the resulting distribution of
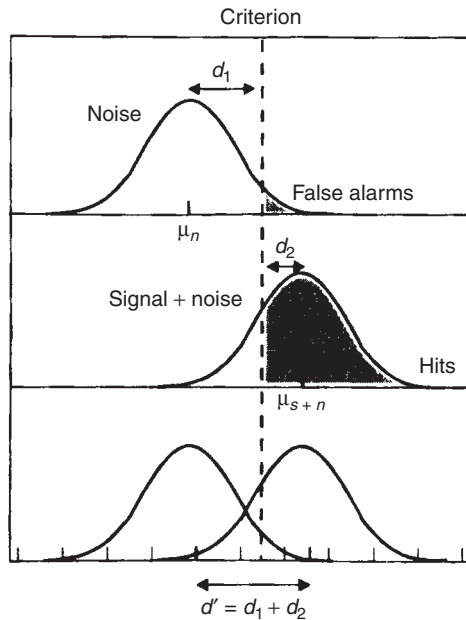
**Figure 3** Equal-variance, normal probability distributions for noise and signal-plus-noise distributions on sensory continuum, with depiction of proportion of false alarms, proportion of hits, and computation of $d'$. Bottom panel shows both distributions on a single continuum. (From Proctor and Van Zandt, 2008.)

values is normal (i.e., bell shaped and symmetric), or Gaussian, for both the signal and noise trials and that the variances for the two distributions are equal (see Figure 3). To the extent that the signal is discriminable from the noise, the distribution for the signal trials should be shifted to the right (i.e., higher on the continuum of evidence values) relative to that for the noise trials. The measure $d'$ is therefore the distance between the means of the signal and noise distributions, in standard deviation units. That is,

$$d' = \frac{\mu_s - \mu_n}{\sigma}$$

where $\mu_s$ is the mean of the signal distribution, $\mu_n$ is the mean of the noise distribution, and $\sigma$ is the standard deviation of both distributions. The assumption is that the observer will respond yes whenever the evidence value on any trial exceeds a criterion. The measure of $\beta$, which is expressed by the formula

$$\beta = \frac{f_s(C)}{f_n(C)}$$

where $C$ is the criterion and $f_s$ and $f_n$ are the heights of the signal and noise distributions, respectively, is the likelihood ratio for the two distributions at the criterion. It indicates the placement of this criterion with respect

to the distributions and thus reflects the relative bias to respond yes or no.

Computation of $d'$ and $\beta$ is relatively straightforward. The placement of the distributions with respect to the criterion can be determined as follows. The *hit rate* is the proportion of signal trials on which the observer correctly said yes; this can be depicted graphically by placing the criterion with respect to the signal distribution so that the proportion of the distribution exceeding it corresponds to the hit rate. The *false-alarm rate* is the proportion of noise trials on which the observer incorrectly said yes. This corresponds to the proportion of the noise distribution that exceeds the criterion; when the noise distribution is placed so that the proportion exceeding the criterion is the false-alarm rate, relative positions of the signal and noise distributions are depicted. Sensitivity, as measured by $d'$, is the difference between the means of the signal and noise distributions, and this difference can be found by separately calculating the distance of the criterion from each of the respective means and then combining those two distances. Computationally, this involves converting the false-alarm rate and hit rate into standard normal $z$ scores. If the criterion is located between the two means, $d'$ is the sum of the two $z$ scores. If the criterion is located outside that range, the smaller of the two $z$ scores must be subtracted from the larger to obtain $d'$. The likelihood ratio measure of bias, $\beta$, can be found from the hit and false-alarm rates by using a $z$ table that specifies the height of the distribution for each $z$ value. When $\beta$ is 1.0, no bias exists to give one or the other response. A value of $\beta$ greater than 1.0 indicates a bias to respond no, whereas a bias less than 1.0 indicates a bias to respond yes.

Although $\beta$ has been used most often as the measure of bias to accompany $d'$, several investigations have indicated that an alternative bias measure, $C$, is better (Snodgrass and Corwin, 1988; Macmillan and Creelman, 1990; Corwin, 1994), where $C$ is a measure of criterion location rather than likelihood ratio. Specifically,

$$C = -0.5[z(H) + z(F)]$$

where $H$ is the hit rate and $F$ the false-alarm rate. Here $C$ is superior to $\beta$ on several grounds, including that it is less affected by the level of accuracy than is $\beta$ and will yield a meaningful measure of bias when accuracy is near chance.

For a given $d'$, the possible combinations of hit rates and false-alarm rates that the observer could produce through adopting different criteria can be depicted in a *receiver operating characteristic* (ROC) curve (see Figure 4). The farther an ROC is from the diagonal that extends from hit and false-alarm rates of 0–1, which represents chance performance (i.e., $d'$ of 0), the greater the sensitivity. The procedure described above yields only a single point on the ROC, but in many cases it is advantageous to examine performance under several criteria settings, so that the form of the complete ROC is evident (Swets, 1986). One advantage is that the estimate of sensitivity will be more reliable when it is based on several points along the ROC than when it is based on
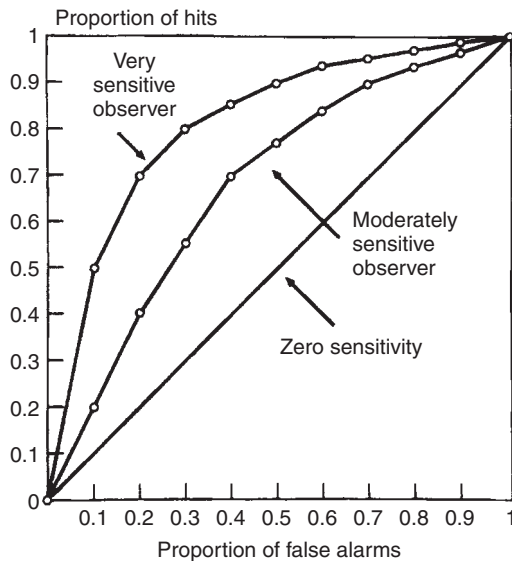
**Figure 4** ROC curves showing the possible hit and false-alarm rates for different sensitivities. (From Proctor and Van Zandt, 2008.)

only one. Another is that the empirical ROC can be compared to the ROC implied by the psychophysical model that underlies a particular measure of sensitivity to determine whether serious deviations occur. For example, when enough points are obtained to estimate complete ROC curves, it is possible to evaluate the assumptions of equal-variance, normal distributions on which the measures of $d'$ and $\beta$ are based. When plotted on $z$-score coordinates, the ROC curve will be linear with a slope of 1.0 if both assumptions are supported; deviations from a slope of 1.0 mean that one distribution is more variable than the other, whereas systematic deviations from linearity indicate that assumption of normality is violated. If either of these deviations is present, alternative measures of sensitivity and bias that do not rely on the assumptions of normality and equal variance should be used.

For cases in which a complete ROC curve is desired, several procedures exist for varying response criteria. The relative payoff structure may be varied across blocks of trials to make one or the other response more preferable; similarly, instructions may be varied regarding how the observer is to respond when uncertain. Another way to vary response criteria is to manipulate the relative probabilities of the signal-and-noise trials; the response criterion should be conservative when signal trials are rare and become increasingly more liberal as the signal trials become increasingly more likely. One of the most efficient techniques is to use rating scales (e.g., from 1, meaning very sure that the signal was not present, to 5, meaning very sure that it was present) rather than yes–no responses. The ratings are then treated as a series of criteria, ranging from high to low, and hit and false-alarm rates are calculated with respect to each. Eng (2006) provides an online program

for plotting an ROC curve and calculating summary statistics.

Signal detection methods are powerful tools for investigating basic and applied problems pertaining not only to sensation and perception but also to many other areas in which an observer's response must be based on probabilistic information, such as distinguishing normal from abnormal X rays (Manning and Leach, 2002) or detecting whether severe weather will occur within the next hour (Harvey, 2003). Although most work on signal detection theory has involved discriminations along a single psychological continuum, it has been extended also to situations in which multidimensional stimuli are presumed to produce values on multiple psychological continua such as color and shape (e.g., Macmillan, 2002). Such analyses have the benefit of allowing evaluation of whether the stimulus dimensions are processed in perceptually separable and independent manners and whether the decisions for each dimension are also separable. As these examples illustrate, signal detection methods can be extremely effective when used with discretion.

### 2.2.2 Psychophysical Scaling

Another concern in psychophysics is to construct scales for the relation between physical intensity and perceived magnitude (see Marks and Gescheider, 2002, for a review). One way to build such scales is to do so from discriminative responses to stimuli that differ only slightly. Fechner (1860) established procedures for constructing psychophysical scales from difference thresholds. Later, Thurstone (1927) proposed a method for constructing a scale from paired comparison procedures in which each stimulus is compared to all others. Thurstonian scaling methods can even be used for complex stimuli for which physical values are not known. Work on scaling in this tradition continues to this day in what is called Fechnarian multidimensional scaling (Dzhafarov and Colonius, 2005), which "borrows from Fechner the fundamental idea of computing subjective dissimilarities among stimuli from the observers' ability to tell apart very similar stimuli" (p. 3).

An alternative way to construct scales is to use direct methods that require some type of magnitude judgment (see, e.g., Bolanowski and Gescheider, 1991, for an overview). Stevens (1975) established methods for obtaining direct magnitude judgments. The technique of *magnitude estimation* is the most widely used. With this procedure, the observer is either presented a standard stimulus and told that its sensation is a particular numerical value (modulus) or allowed to choose his or her own modulus. Stimuli of different magnitudes are then presented randomly, and the observer is to assign values to them proportional to their perceived magnitudes. These values then provide a direct scale relating physical magnitude to perceived magnitude. A technique called *magnitude production* can also be used, in which the observer is instructed to adjust the value of a stimulus to be a particular magnitude. Variations of magnitude estimation and production have been used to measure such things as emotional stress (Holmes and Rahe, 1967) and pleasantness of voice quality for normal

speakers and persons with a range of vocal pathology (Eadie and Doyle, 2002). Furthermore, Walker (2002) provided evidence that magnitude estimation can be used as a design tool in the development of data sonifications, that is, representations of data by sound. Within physical ergonomics, Borg's (1998) RPE (ratings for perceived exertion) and CR-10 (category ratio, 10 categories) scales are used to measure magnitude of perceived exertion and discomfort, respectively, in situations where physical activity is required.

Baird and Berglund (1989) coined the term *environmental psychophysics* for the application of psychophysical methods such as magnitude estimation to applied problems of the type examined by Berglund (1991) that are associated with odorous air pollution and community noise. As Berglund puts it: "The method of ratio scaling developed by S. S. Stevens (1956) is a contribution to environmental science that ranks as good and important as most methods from physics or chemistry" (p. 141). When any measurement technique developed for laboratory research is applied to problems outside the laboratory, special measurement issues may arise. In the case of environmental psychophysics, the environmental stimulus of concern typically is complex and multisensory, diffuse, and naturally varying, presented against an uncontrollable background. The most serious measurement problem is that often it is not possible to obtain repeated measurements from a given observer under different magnitude concentrations, necessitating that a scale be derived from judgments of different observers at different points in time.

Because differences exist in the way that people assign magnitude numbers to stimuli, each person's scale must be calibrated properly. Berglund and her colleagues have developed what they call the *master scaling procedure* to accomplish this purpose. The procedure has observers make magnitude judgments for several values of a referent stimulus as well as for the environmental stimulus. Each observer's power function for the referent stimulus is transformed to a single master function (this is much like converting different normal distributions to the standard normal distribution for comparison). The appropriate transformation for each observer is then applied to her or his magnitude judgment for the environmental stimulus so that all such judgments are in terms of the master scale.

### 2.2.3  Other Techniques

Many other techniques have been used to investigate issues in sensation and perception. Most important are methods that use response times either instead of or in conjunction with response accuracy (see Welford, 1980; Van Zandt, 2002). Reaction time methods have a history of use approximately as long as that of the classical psychophysical methods, dating back to Donders (1868), but their use has been particularly widespread since about 1950. Simple reaction times require the observer to respond as quickly as possible with a single response (e.g., a keypress) whenever a stimulus event occurs. Alternative hypotheses of various factors that affect detection of the stimulus and the decision to respond, such as the locus of influence of visual masking

and whether the detection of two signals presented simultaneously can be conceived of as an independent race, can be evaluated using simple reaction times.

Decision processes play an even larger role in go–no go tasks, where responses must be made to some stimuli but not to others, and in choice–reaction tasks, where there is more than one possible stimulus and more than one possible response and the stimulus must be identified if the correct response is to be made. Methods such as the *additive factors logic* (Sternberg, 1969) can be used to isolate perceptual and decisional factors. This logic proposes that two variables whose effects are additive affect different processing stages, but two variables whose effects are interactive affect the same processing stage. Variables that interact with marker variables whose effects can be assumed to be in perceptual processes, but not with marker variables whose effects are on response selection or programming, can be assigned a perceptual locus. Analyses based on the distributions of reaction times have gained in popularity in recent years. Van Zandt (2002) provides MATLAB code for performing such analyses.

## 3  SENSORY SYSTEMS AND BASIC PERCEPTUAL PHENOMENA

The ways in which the sensory systems encode information have implications not only for the structure and function of the sensory pathways but ultimately also for the nature of human perception. They also place restrictions on the design of displays. Displays must be designed to satisfy known properties of sensory encoding (e.g., visual information that would be legible if presented in central vision will not be legible if the display were presented in the visual periphery), but they do not need to exceed the capabilities of sensory encoding. The sensory information that is encoded also must be represented in the nervous system. The nature of this representation also has profound implications for perception.

### 3.1  Vision

#### 3.1.1  Visual System

The sensory receptors in the eye are sensitive to energy within a limited range of the electromagnetic spectrum. One way of characterizing such energy is as continuous waves of different wavelengths. The visible spectrum ranges from wavelengths of approximately 370 nm (billionths of a meter) to 730 nm. Any energy outside this range, such as ultraviolet rays, will not be detected because they have no effect on the receptors. Light can also be characterized in terms of small units of energy called *photons*. Describing light in terms of wavelength is important for some aspects of perception, such as color vision, whereas for others it is more useful to treat it in terms of photons. As with any system in which light energy is used to create a representation of the physical world, the light must be focused and a clear image created. In the case of the eye, the image is focused on the photoreceptors located on the retina, which lines the back wall of the eye.
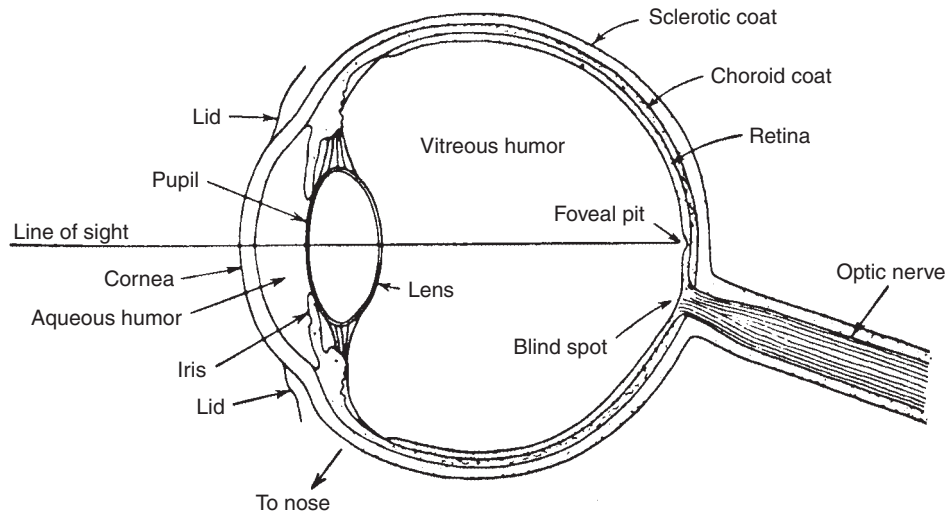
**Figure 5** Structure of the human eye. (From Schiffman, 1996.)

***Focusing System*** Light enters the eye (see Figure 5) through the cornea, a transparent layer that acts as a lens of fixed optical power and provides the majority of the focusing. The remainder of the focusing is accomplished by the crystalline lens, whose power varies automatically as a function of the distance from the observer of the object that is being fixated. Beyond a distance of approximately 6 m, the far point, the lens is relatively flat; for distances closer than the far point, muscles attached to the lens cause it to become progressively more spherical the closer the fixated object is to the observer, thus increasing its refractive power. The reason why this process, called *accommodation*, is needed is that without an increase in optical power for close objects their images would be focused at a point beyond the retina and the retinal image would be out of focus. Accommodation is effective for distances as close as 20 cm (the near point), but the extent of accommodation, and the speed at which it occurs, decreases with increasing age, with the near point receding to approximately 100 cm by age 60. This decrease in accommodative capability, called *presbyopia*, can be corrected with reading glasses. Other imperfections of the lens system—*myopia*, where the focal point is in front of the receptors; *hyperopia*, where the focal point is behind the receptors; and *astigmatism*, where certain orientations are out of focus while others are not—also typically are treated with glasses.

Between the cornea and the lens, the light passes through the pupil, an opening in the center of the iris that can vary in size from 2 to 8 mm. The pupil size is large when the light level is low, to maximize the amount of light that gets into the eye, and small when the light level is high, to minimize the imperfections in imaging that arise when light passes through the extreme periphery of the lens system. One additional consequence of these changes in image quality as a function of pupil size is that the *depth of field*, or the distance in front of or behind a fixated object at which

the images of other objects will be in focus also, will be greatest when the pupil size is 2 mm and decrease as pupil size increases, at least up to intermediate diameters (Marcos et al., 1999). In other words, under conditions of low illumination, accommodation must be more precise and work that requires high acuity, such as reading, can be fatiguing (Randle, 1988). When required to accommodate to near stimuli, adults show accommodative pupil restrictions that increase the depth of field. This tendency for restriction in pupil size is much weaker for children (Gislén et al., 2008; the children in their study were 9–10 years of age), most likely due to the superior accommodative range of their lenses.

If the eyes fixate on an object at a distance of approximately 6 m or farther, the lines of sight are parallel. As the object is moved progressively closer, the eyes turn inward and the lines of sight converge. Thus, the degree of *vergence* of the eyes varies systematically as a function of the distance of the object being fixated. The near point for vergence is approximately 5 cm, and if an object closer than that is fixated, the images at the two eyes will not be fused and a double image will be seen.

The natural resting states for accommodation and vergence, called *dark focus* and *dark vergence*, respectively, are intermediate to the near and far points (Leibowitz and Owens, 1975; Andre, 2003; Jaschinski et al., 2007). One view for which there is considerable support is that dark focus and vergence provide zero reference points about which accommodative and vergence effort varies (Ebenholtz, 1992). A practical implication of this is that less eye fatigue will occur if a person working at a visual display screen for long periods of time is positioned at a distance that corresponds approximately to the dark focus and vergence points. As with most other human characteristics of concern in human factors and ergonomics, considerable individual differences in dark focus and vergence exist. People with far dark-vergence postures tend to position themselves

farther away from the display screen than will those with closer postures (Heuer et al., 1989), and they also show more visual fatigue when required to perform close visual work (Jaschinski-Kruza, 1991).

**Retina** If the focusing system is working properly, the image will be focused on the retina, which lines the back wall of the eye. Objects in the left visual field will be imaged on the right hemi-retina and objects in the right visual field on the left hemi-retina; objects above the point of fixation will be imaged on the lower half of the retina, and vice versa for objects below fixation. The retina contains the photoreceptors that transduce the light energy into a neural signal; their spatial arrangement limits our ability to perceive spatial pattern (see Figure 6). There also are two layers of neurons, and their associated blood vessels, that process the retinal image before information about it is sent along the optic nerve to the brain. These neural layers are in the light path between the lens and the photoreceptors and thus degrade to some extent the clarity of the image at the receptors.

There are two major types of photoreceptors, *rods* and *cones*, with three subtypes of cones. All photoreceptors contain light-sensitive photopigments in their outer segments that operate in basically the same manner. Photons of light are absorbed by the photopigment when they strike it, starting a reaction that leads to the generation of a neural signal. As light is absorbed, the photopigment becomes insensitive and is said to be *bleached*. It must go through a process of regeneration before it is functional again. Because the rod and cone photopigments differ in their absolute sensitivities to light energy, as well as in their differential sensitivities to light across the visual spectrum, the rods and cones have different roles in perception.

Rods are involved primarily in vision under very low levels of illumination, what is called *scotopic vision*. All rods contain the same photopigment, rhodopsin, which is highly sensitive to light. Its spectral sensitivity function shows it to be maximally sensitive to light around 500 nm and to a lesser degree to other wavelengths. One consequence of there being only one rod photopigment is that we cannot perceive color under scotopic conditions. The reason for this is easy to understand. The rods will respond relatively more to stimulation of 500 nm than they will to 560-nm stimulation of equal intensity. However, if the intensity of the 560-nm stimulus is increased, a point would be reached at which the rods responded equally to the two stimuli. In other words, with one photopigment, there is no basis for distinguishing among the wavelength differences associated with color differences from intensity differences.

Cones are responsible for vision in daylight, or what is known as *photopic vision*. Cone photopigments are less sensitive to light than rhodopsin, and hence cones are operative at levels of illumination at which the rod photopigment has been effectively fully bleached. Also, because there are three types of cones, each containing a different photopigment, cones provide color vision. As explained previously, there must be more than one photopigment type if differences in the wavelength of stimulation are to be distinguished from differences in intensity. The spectral sensitivity functions for each of the three cone photopigments span broad ranges of the visual spectrum, but their peak sensitivities are located at different wavelengths. The peak sensitivities are approximately 440 nm for the short-wavelength ("blue") cones, 540 nm for the middle-wavelength ("green") cones, and 565 nm for the long-wavelength ("red") cones. Monochromatic light of a particular wavelength will produce a pattern of activity for the three cone types that is unique from the patterns produced by other wavelengths, allowing each to be distinguished perceptually.

The retina contains two landmarks that are important for visual perception. The first of these is the optic disk, which is located on the nasal side of the retina. This is the region where the optic nerve, composed of the nerve fibers from the neurons in the retina, exits the eye. The significant point is that there are no photoreceptors in this region, which is why it is sometimes called the *blind spot*. We do not normally notice the blind spot because (1) the blind spot for one of the eyes corresponds to part of the normal visual field for the other eye and (2) with monocular viewing, the perceptual system fills it in with fabricated images based on visual attributes from nearby regions of the visual field (Araragi et al., 2009). How this filling in occurs has been the subject of considerable investigation, with evidence from physiological studies and computational modeling suggesting that the filling in is induced by neurons in the primary visual cortex through slow conductive paths of horizontal connections in the primary visual cortex and fast feed-forward/feedback paths by way of the visual association cortex (Matsumoto and Komatsu, 2005; Satoh and Usui, 2008). If the image of an object falls only partly on the blind spot, the filling in from the surrounding region will cause the object to appear complete. However, if the image of an object falls entirely within the blind spot, this filling in will cause the object to not be perceived.

The second landmark is the *fovea*, which is a small indentation about the size of a pinhead on which the image of an object at the point of fixation will fall. The fovea is the region of the retina in which visual acuity is highest. Its physical appearance is due primarily to the fact that the neural layers are pulled away, thus allowing the light a straight path to the receptors. Moreover, the fovea contains only cones, which are densely packed in this region.

As shown in Figure 6, the photoreceptors synapse with bipolar cells, which in turn synapse with ganglion cells; the latter cells are the output neurons of the retina, with their axons making up the optic nerve. In addition, horizontal cells and amacrine cells provide interconnections across the retina. The number of ganglion cells is much less than the number of photoreceptors, so considerable convergence of the activity of individual receptors occurs. The neural signals generated by the rods and cones are maintained in distinct pathways until reaching the ganglion cells (Kolb, 1994). In the fovea, each cone has input into more than one ganglion cell. However, convergence is the rule outside the fovea, being an increasing function of distance from

Rods                    Cones



Horizontal
cell

Bipolar
cell

Amacrine
cell

Ganglion
cell

Light
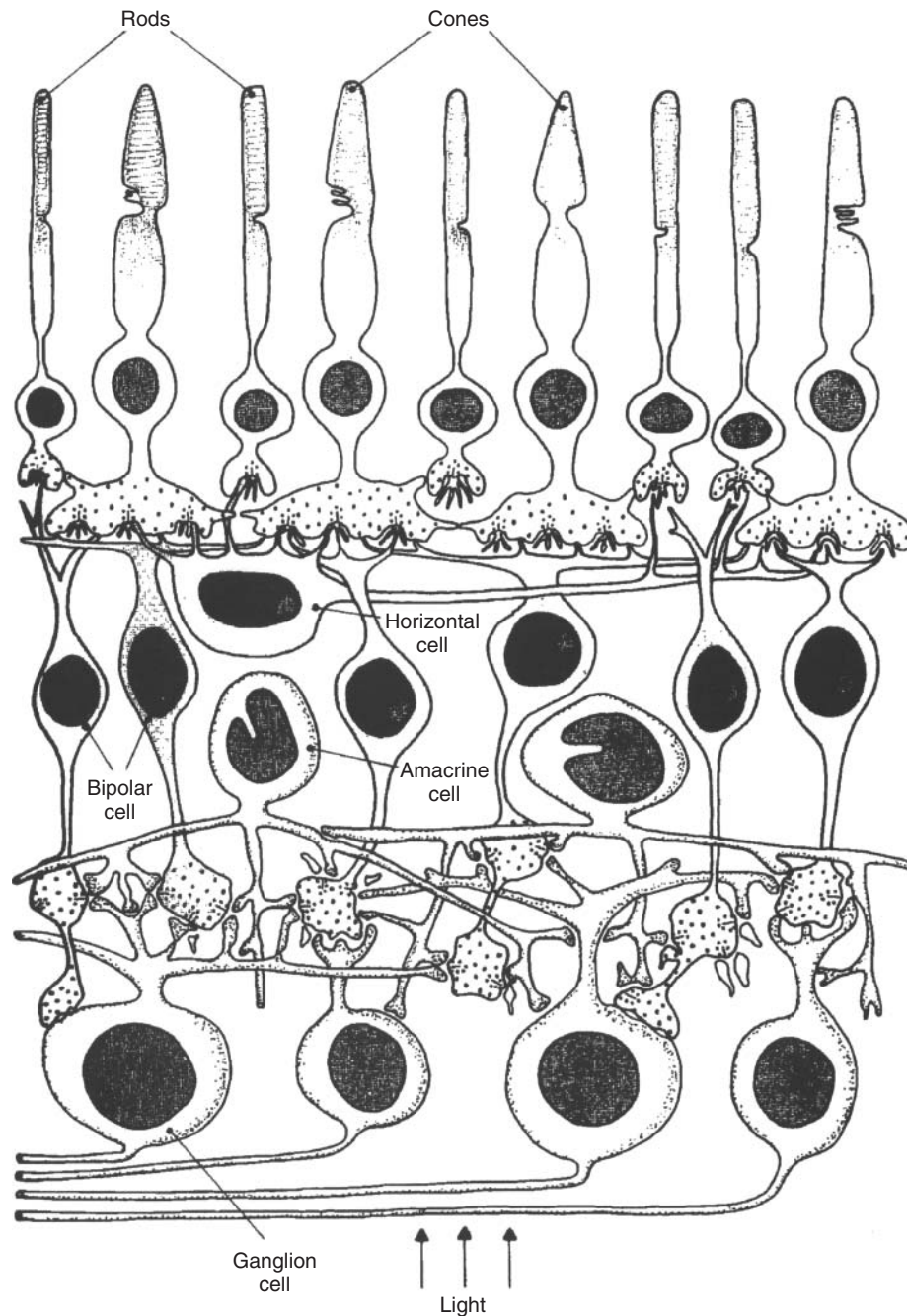
**Figure 6**   Neural structures and interconnections of vertebrate retina. (From Dowling and Boycott, 1966.)

the fovea. Overall, the average convergence is 120:1 for rods as compared to 6:1 for cones. The degree of convergence has two opposing perceptual consequences. Where there is little or no convergence, as in the neurons carrying signals from the fovea, the pattern of stimulation at the retina is maintained effectively complete, thus maximizing spatial detail. When there is considerable convergence, as for the rods, the activity of many photoreceptors in the region is pooled together, optimizing sensitivity to light at the cost of detail. Thus,

the wiring of the photoreceptors is consistent with the fact that the rods operate when light energy is at a premium but the cones operate when it is not.

The ganglion cells show several interesting properties pertinent to perception. When single-cell recording techniques are used to measure their receptive fields (i.e., the regions on the retina that when stimulated produce a response in the cell), these fields are found to have a circular, center-surround relation for most cells. If light presented in a circular, center region causes an increase in the firing rate of the neuron, light presented in a surrounding ring region will cause a decrease in the firing rate, or vice versa. What this means is that the ganglion cells are tuned to respond primarily to discontinuities in the light pattern within their receptive fields. If the light energy across the entire receptive field is increased, there will be little if any effect on the firing rate. In short, the information extracted and signaled by these neurons is based principally on contrast, which is important for perceiving objects in the visual scene, and not on absolute intensity, which will vary as a function of the amount of illumination. Not surprisingly, the average receptive field size is larger for ganglion cells receiving their input from rods than for those receiving it solely from cones and increases with increasing distance from the fovea.

Although most ganglion cells have the center-surround receptive field organization, two pathways can be distinguished on the basis of other properties. The ganglion cells in the *parvocellular pathway* have small cell bodies and relatively dense dendritic fields. Many of these ganglion cells, called midget cells, receive their input from the fovea. They have relatively small receptive fields, show a sustained response as long as stimulation is present in the receptive field, and have a relatively slow speed of transmission. The ganglion cells in the *magnocellular pathway* have larger cell bodies and sparse dendritic trees. They have their receptive fields at locations across the retina, have relatively large receptive fields, show a transient response to stimulation that dissipates if the stimulus remains on, have a fast speed of transmission, and are sensitive to motion. Because of these unique characteristics and the fact that these channels are kept separated later in the visual pathways, it has been thought that they contribute distinct information to perception. The parvocellular pathway is presumed to be responsible for pattern perception and the magnocellular pathway for high-temporal-frequency information, such as in motion perception and perception of flicker. The view that different aspects of the sensory stimulus are analyzed in specialized neural pathways has received considerable support in recent years.

**Visual Pathways**  The optic nerve from each eye splits at what is called the *optic chiasma* (see Figure 7). The fibers conveying information from the nasal halves of the retinas cross over and go to the opposite sides of the brain, whereas the fibers conveying information from the temporal halves do not cross over. Functionally, the significance of this is that for both eyes input from the right visual field is sent to the left half of the brain and

input from the left visual field is sent to the right half. A relatively small subset of the fibers (approximately 10%) splits off from the main tract and the fibers go to structures in the brain stem, the tectum, and then the pulvinar nucleus of the thalamus. This *tectopulvinar pathway* is involved in localization of objects and the control of eye movements.

Approximately 90% of the fibers continue on the primary *geniculostriate pathway*, where the first synapse is at the lateral geniculate nucleus (LGN). The distinction between the parvocellular and magnocellular pathways is maintained here. The LGN is composed of six layers, four parvocellular and two magnocellular, each of which receives input from only a single eye. Hence, at this level the input from the two eyes has yet to be combined. Each layer is laid out in a retinotopic map that provides a spatial representation of the retina. In other respects, the receptive field structure of the LGN neurons is similar to that of the ganglion cells. The LGN also receives input from the visual cortex, both directly and indirectly, by way of a thalamic structure called the reticular nucleus that surrounds the LGN (Briggs and Usrey, 2011). This feedback likely modulates the activity in the LGN, allowing the communication between the visual cortex and LGN to be bidirectional.

From the LGN, the fibers go to the primary visual cortex, which is located in the posterior cortex. This region is also called the *striate cortex* (because of its stripes), *area 17*, or *area V1*. The visual cortex consists of six layers. The fibers from the LGN have their synapses in the fourth layer from the outside, with the parvocellular neurons sending their input to one layer $(4C\beta)$ and the magnocellular neurons to another $(4C\alpha)$, and they also have collateral projections to different parts of layer 6. The neurons in these layers then send their output to other layers. In layer 4 the neurons have circular-surround receptive fields, but in other layers, they have more complex patterns of sensitivity. Also, whereas layer 4 neurons receive input from one or the other eye, in other layers most neurons respond to some extent to stimulation at either eye.

A distinction can be made between simple cells and complex cells (e.g., Hubel and Wiesel, 1977). The responses of simple cells to shapes can be determined from their responses to small spots of light (e.g., if the receptive field for the neuron is plotted using spots of light, the neuron will be most sensitive to a stimulus shape that corresponds with that receptive field), whereas those for complex cells cannot be. Simple cells have center-surround receptive fields, but they are more linear than circular; this means that they are orientation selective and will respond optimally to bars in an orientation that corresponds with that of the receptive field. Complex cells have similar linear receptive fields and so are also orientation selective, but they are movement sensitive as well. These cells respond optimally not only when the bar is at the appropriate orientation but also when it is moving. Some cells, which receive input from the magnocellular pathway, are also directionally sensitive: they respond optimally to movement in a particular direction. Certain cells, called *hypercomplex cells*, are sensitive to the length of
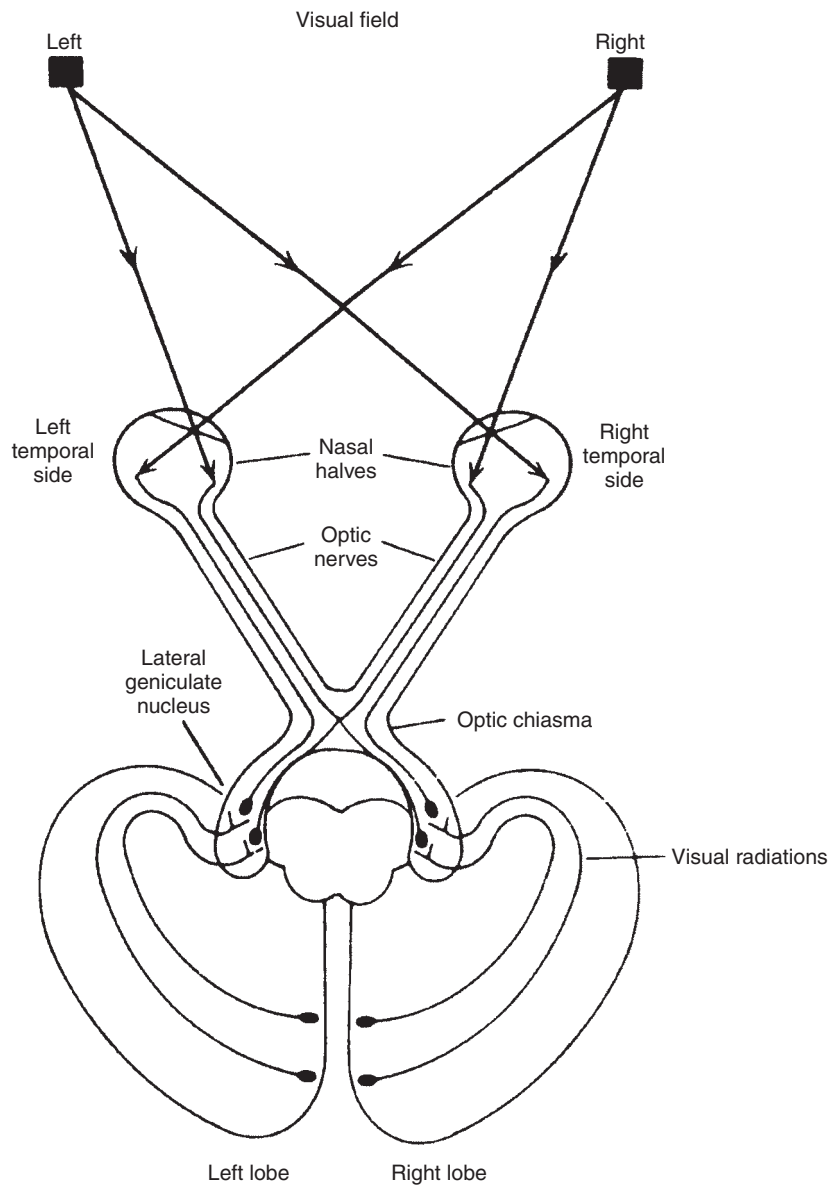
**Figure 7** Human visual system showing projection of visual fields through the system. (From Schiffman, 1996.)

the bar so that they will not respond if the bar is too long. Some neurons in the visual cortex are also sensitive to disparities in the images at each eye and to motion velocity. In short, the neurons of the visual cortex analyze the sensory input for basic features that provide the information on which higher level processes operate.

The cortex is composed of columns and hypercolumns arranged in a spatiotopic manner. Within a single column, all of the cells except for those in layer 4 will have the same preferred orientation. The next column will respond to stimulation at the same location on the retina but will have a preferred orientation

that is approximately $10°$ different than that of the first column. As we proceed through a group of about 20 columns, called a *hypercolumn*, the preferred orientation will rotate $180°$. The next hypercolumn will show the same arrangement, but for stimulation at a location on the retina that is adjacent to that of the first. A relatively larger portion of the neural machinery in the visual cortex is devoted to the fovea, which is to be expected because it is the region for which detail is being represented.

Two cortical areas, V2 (the first visual association area) and MT (medial temporal), receive input from

V1 (Briggs and Usrey, 2010). Part of the V2 neurons also project to MT, but another part goes to cortical area V4, making up the dorsal (where) and ventral (what) streams, respectively. Barber (1990) found that visual abilities pertaining to detection, identification, acquisition, and tracking of aircraft in air defense simulations clustered into subgroups consistent with the hypothesis that different aspects of visual perception are mediated by these distinct subsystems in the brain.

### 3.1.2 Basic Visual Perception

***Brightness***   Brightness is that aspect of visual perception that corresponds most closely to the intensity of stimulation. To specify the effective intensity of a stimulus, photometric measures, which are calibrated to reflect human spectral sensitivity, should be used. A photometer can be used to measure either *illuminance*, that is, the amount of light falling on a surface, or *luminance*, that is, the amount of light generated by a surface. To measure illuminance, an illuminance probe is attached to the photometer and placed on the illuminated surface. The resulting measure of illuminance is in lumens per square meter ($lm/m^2$) or lux (lx). To measure luminance, a lens with a small aperture is attached to the photometer and focused onto the surface from a distance. The resulting measure of luminance is in candelas per square meter ($cd/m^2$). Although measures with a photometer are suitable with most displays, the relatively new technology of laser-based video projectors requires an alternative method, because each pixel produces light for too brief of a time for accurate photometric measurement (Doucet et al., 2010).

Judgments of brightness are related to intensity by the power function

$$B = aI^{0.33}$$

where $B$ is brightness, $a$ is a constant, and $I$ is the physical intensity. However, brightness is not determined by intensity alone but also by several other factors. For example, at brief exposures on the order of 100 ms or less and for small stimuli, temporal and spatial summation occurs. That is, a stimulus of the same physical intensity will look brighter if its exposure duration is increased or if its size is increased.

One of the most striking influences on brightness perception and sensitivity to light is the level of *dark adaptation*. When a person first enters a dark room, he or she is relatively insensitive to light energy. However, with time, dark adaptation occurs and sensitivity increases drastically. The time course of dark adaptation is approximately 30 min. Over the first couple of minutes, the absolute threshold for light decreases and then levels off. However, after approximately 8 min in the dark, it begins decreasing again, approaching a maximum around the 30-min point. After 30 min in the dark, lights can be seen that were of too low intensity to be visible initially, and stimuli that appeared dim now seem much brighter. Dark adaptation reflects primarily regeneration into a maximally light-sensitive state of the cone photopigments and then the rod photopigment. Jackson et al. (1999) reported a dramatic slowing in the rod-mediated component of dark adaptation that may contribute to increased night vision problems experienced by older adults. After becoming dark adapted, vision may be impaired momentarily when the person returns to photopic viewing conditions. Providing gradually changing light intensity in regions where light intensity would normally change abruptly, such as at the entrances and exits of tunnels, may help minimize such impairment (e.g., Oyama, 1987).

The brightness of a monochromatic stimulus of constant intensity will vary as a function of its wavelength because the photopigments are differentially sensitive to light of different wavelengths. The scotopic spectral sensitivity function is shifted toward the short-wavelength end of the spectrum, relative to the photopic function. Consequently, if two stimuli, one short wavelength and one long, appear equally bright at photopic levels, the short-wavelength stimulus will appear brighter at scotopic levels, a phenomenon called the *Purkinje shift*. Little light adaptation will occur when high levels of long-wavelength light are present because the sensitivity of the rods to long-wavelength light is low. Thus, it is customary to use red light sources to provide high illumination for situations in which a person needs to remain dark adapted.

It is common practice to distinguish between brightness and lightness as two different aspects of perception (Blakeslee et al., 2008): *Judgments of brightness* are of the apparent luminance of a stimulus, whereas *judgments of lightness* are of perceived achromatic color along a black-to-white dimension (i.e., apparent reflectance). Both the brightness and lightness of an object are greatly influenced by the surrounding context. *Lightness contrast* is a phenomenon where the intensity of a surrounding area influences the lightness of a stimulus. The effects can be quite dramatic, with a stimulus of intermediate reflectance ranging in appearance from white to dark gray or black as the reflectance of the surround is increased from low to high. The more common phenomenon of *lightness constancy* occurs when the level of illumination is increased across the entire visual field. In this case, the absolute amount of light reflected to the eye by an object may be quite different, but the percept remains constant. Basically, lightness follows a constant-ratio rule (Wallach, 1972): Lightness will remain the same if the ratio of light energy for a stimulus relative to its surround remains constant. Lightness constancy holds for a broad range of ratios and across a variety of situations, with brightness constancy obtained under a more restricted set of viewing conditions (Jacobsen and Gilchrist, 1988; Arend, 1993).

Although low-level mechanisms early in the sensory system probably contribute at least in part to constancy and contrast, more complex higher level brain mechanisms apparently do as well. Particularly compelling are demonstrations showing that the lightness and brightness of an object can vary greatly simply as a function of organizational and depth cues. Agostini and Proffitt (1993) demonstrated lightness contrast as a function of whether a target gray circle was organized perceptually with black or white circles, even though the inducing circles were not in close proximity to the target. Gilchrist
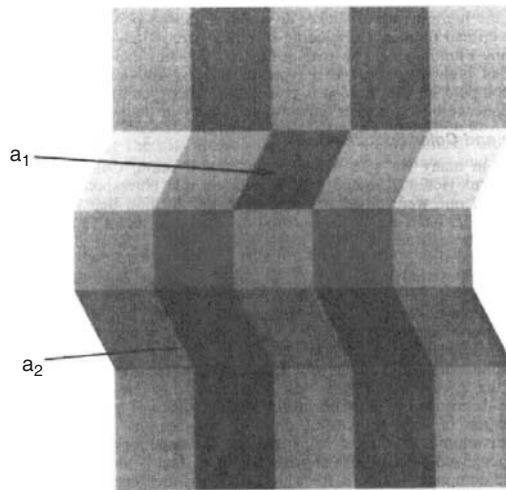
**Figure 8** Effect of perceived shading on brightness judgments. The patches $a_1$ and $a_2$ are the same shade of gray, but $a_1$ appears much darker than $a_2$. (Reprinted with permission from Adelson, 1993. Copyright 1993 by the American Association for the Advancement of Science.)

(1977) used depth cues to cause a piece of white paper to be perceived incorrectly as in a back chamber that was highly illuminated or correctly as in a front chamber that was dimly illuminated. When perceived as in the front chamber, the paper was seen as white; however, when perceived as in the back chamber, the paper appeared to be almost black. Adelson (1993) showed that such effects are not restricted to lightness but also occur for brightness judgments. For example, when instructed to adjust the luminance of square $a_1$ in Figure 8 to equal that of square $a_2$, observers set the luminance of $a_1$ to be 70% higher than that of $a_2$. Thus, even relatively "sensory" judgments such as brightness are affected by higher order organizational factors.

***Visual Acuity and Sensitivity to Spatial Frequency*** *Visual acuity* refers to the ability to perceive detail. Acuity is highest in the fovea, and it decreases with increasing eccentricities due in part to the progressively greater convergence of activity from the sensory receptors that occurs in the peripheral retina. Distinctions can be made between different types of acuity. *Identification acuity* is the most commonly measured, using a Snellen eye chart that contains rows of letters that become progressively smaller. The smallest row for which the observer can identify the letters is used as the indicator of acuity. Regarded as normal, 20/20 vision means that the person being tested is able to identify at a distance of 20 ft letters of a size that a person with normal vision is expected to identify. A person with 20/100 vision can identify letters at 20 ft only as large as those that a person with normal vision could at 100 ft. *Vernier acuity* is a person's ability to discriminate between broken and unbroken lines, and *resolution acuity* is the ability to distinguish gratings from a stimulus

that covers the same area but is of the same average intensity throughout. All of these measures are variants of *static acuity*, in that they are based on static displays. *Dynamic acuity* refers to the ability to resolve detail when there is relative motion between the stimulus and the observer. Dynamic acuity is usually poorer than static acuity (Morgan et al., 1983), partly due to an inability to keep a moving image within the fovea (Murphy, 1978). A concern in measuring acuity is that the types are not perfectly correlated, and thus an acuity measure of one type may not be a good predictor of ability to perform a task whose acuity requirements are of a different type. For example, the elderly show typically little loss of identification acuity as measured by a standard test, but they seem to have impaired acuity in dynamic situations and at low levels of illumination (Kosnik et al., 1990; Sturr et al., 1990). Thus, performance on a dynamic acuity test may be a better predictor of driving performance for elderly persons (Wood, 2002). In a step toward making assessment of dynamic acuity more routine, Smither and Kennedy (2010) have developed and evaluated a prototype of an automated, portable dynamic visual acuity system using a low-energy laser.

Spatial contrast sensitivity has been shown to provide an alternative, more detailed way for characterizing acuity. A *spatial contrast sensitivity function* can be generated by obtaining threshold contrast values for discriminating sine-wave gratings (for which the bars change gradually rather than sharply from light to dark) of different spatial frequencies from a homogeneous field. The contrast sensitivity function for a typical adult shows maximum sensitivity at a spatial frequency of about three to five cycles per degree of visual angle, with relatively sharp drop-offs at high and low spatial frequencies. Basically what this means is that we are not extremely sensitive to very fine or course gratings. Because high spatial frequencies relate to the ability to perceive detail and low to intermediate frequencies to the more global characteristics of visual stimuli, tests of acuity based on contrast sensitivity may be more analytic concerning aspects of performance that are necessary for performing specific tasks. For example, Evans and Ginsburg (1982) found contrast sensitivity at intermediate and low spatial frequencies to predict the detectability of stop signs in night driving, and contrast sensitivity measures have also been shown to predict performance of flight-related tasks (Gibb et al., 2010). A main difficulty for use of contrast sensitivity functions for practical applications is the long time required to perform such a test. Lesmes et al. (2010) have developed and performed initial validation tests of a quick method that requires only 25 trials to construct a relatively accurate function.

Of concern in human factors and ergonomics is temporal acuity. Because many light sources and displays present flickering stimulation, we need to be aware of the rates of stimulation beyond which flicker will not be perceptible. The *critical flicker frequency* is the highest rate of flicker at which it can be perceived. Numerous factors influence the critical frequency, including stimulus size, retinal location, and level of

surrounding illumination. The critical flicker frequency can be as high as 60 Hz for large stimuli of high intensity, but it typically is less. You may have noticed that the flicker of a computer display screen is perceptible when you are not looking directly at it as a consequence of the greater temporal sensitivity in the peripheral retina.

***Color Vision and Color Specification***    Color is used in many ways to display visual information. Color can be used, as in television and movies, to provide a representation that corresponds to the colors that would be seen if one were physically present at the location that is depicted. Color is also used to highlight and emphasize as well as to code different categories of displayed information. In situations such as these, we want to be sure that the colors are perceived as intended.

In a color-mixing study, the observer is asked to adjust the amounts of component light sources to match the hue of a comparison stimulus. Human color vision is *trichromatic*, which means that any spectral hue can be matched by a combination of three primary colors, one each from the short-, middle-, and long-wavelength positions of the spectrum. This trichromaticity is a direct consequence of having three types of cones that contain photopigments with distinct spectral sensitivity functions. The pattern of activity generated in the three cone systems will determine what hue is perceived. A specific pattern can be determined by a monochromatic light source of a particular wavelength or by a combination of light sources of different wavelengths. As long as the relative amounts of activation in the three cone systems are the same for different physical stimuli, they will be perceived as being of the same hue. This fact is used in the design of color television sets and computer monitors, for which all colors are generated from combinations of pixels of three different colors.

Another phenomenon of additive color mixing is that blue and yellow, when mixed in approximately equal amounts, yield an achromatic (e.g., white) hue, as do red and green. This stands in contrast to the fact that combinations involving one hue from each of the two complementary pairs are seen as combinations of the two hues. For example, when blue and green are combined additively in similar amounts, the resulting stimulus appears blue-green. The pairs that yield an achromatic additive mixture are called *complementary colors*. That these hues have a special relation is evident in other situations as well. When a background is one of the hues from a pair of complementary colors, it will tend to induce the complementary hue in a stimulus that would otherwise be perceived as a neutral gray or white. Similarly, if a background of one hue is viewed for awhile and then the gaze is shifted to a background of a neutral color, an afterimage of the complementary hue will be seen.

The complementary color relations also appear to have a basis in the visual system, but in the neural pathways rather than in the sensory receptors. That is, considerable evidence indicates that output from the cones is rewired into opponent processes at the level of the ganglion cells and beyond. If a neuron's firing rate increases when a blue stimulus is presented, it decreases when a yellow stimulus is presented. Similarly, if a neuron's firing rate increases to a red stimulus, it decreases to a green stimulus. The pairings in the opponent cells always involve blue with yellow and red with green. Thus, a wide range of color appearance phenomena can be explained by the view that the sensory receptors operate trichromatically, but this information is subsequently recoded into an opponent format in the sensory pathways.

The basic color-mixing phenomena are depicted in color appearance systems. A *color circle* can be formed by curving the visual spectrum, as done originally by Isaac Newton. The center of the circle represents white, and its rim represents the spectral colors. A line drawn from a particular location on the rim to the center depicts saturation, the amount of hue that is present. For example, if one picks a monochromatic light source that appears red, points on the line represent progressively decreasing amounts of red as one moves along it to the center. The appearance for a mixture of two spectral colors can be approximated by drawing a chord that connects the two colors. If the two are mixed in equal amounts, the point at the center corresponds to the mixture; if the percentages are unequal, the point is shifted accordingly toward the higher percentage spectral color. The hue for the mixture point can be determined by drawing a diagonal through it; its hue corresponds to that of the spectral hue at the rim, and its saturation corresponds to the proximity to the rim.

The color circle is too imprecise to be used to specify color stimuli, but a system much like it, the CIE (Commission Internationale de l'Éclairage) system, is the most widely used color specification system. The CIE provided a standardized set of color-matching functions, $x(\lambda)$, $y(\lambda)$, and $z(\lambda)$, called the *XYZ tristimulus coordinate system* (see Figure 9). The tristimulus values for a monochromatic stimulus can be used to determine the proportions of three wavelengths ($X$, $Y$, and $Z$, corresponding to red, green, and blue, respectively) needed to match it. For example, the $X$, $Y$, and $Z$ tristimulus values for a 500-nm stimulus are 0.0049, 0.3230, and 0.2720. The proportion of $X$ primary can be determined by dividing the tristimulus value for $X$ by the combined values for $X$ plus $Y$ plus $Z$. The proportion of $Y$ primary can be determined in like manner, and the proportion of $Z$ primary is simply 1 minus the $X$ and $Y$ proportions. The spectral stimulus of 500 nm thus has the following proportions: $x = 0.008$, $y = 0.539$, and $z = 0.453$.

The *CIE color space*, shown in Figure 9, is triangular rather than circular. Location in the space is specified according to the $x$ and $y$ color coordinates. Only $x$ and $y$ are used as the coordinates for the space because $x$, $y$, and $z$ sum to 1.0. The spectral stimulus of 500 nm would be located on the rim of the triangle, in the upper left of the figure. Saturation decreases as proximity to the rim decreases, to an achromatic point labeled $C$ in Figure 9. The color of a stimulus can be specified precisely by using the tristimulus values for each component spectral frequency or approximately by determining the coordinates at the location approximating its appearance.
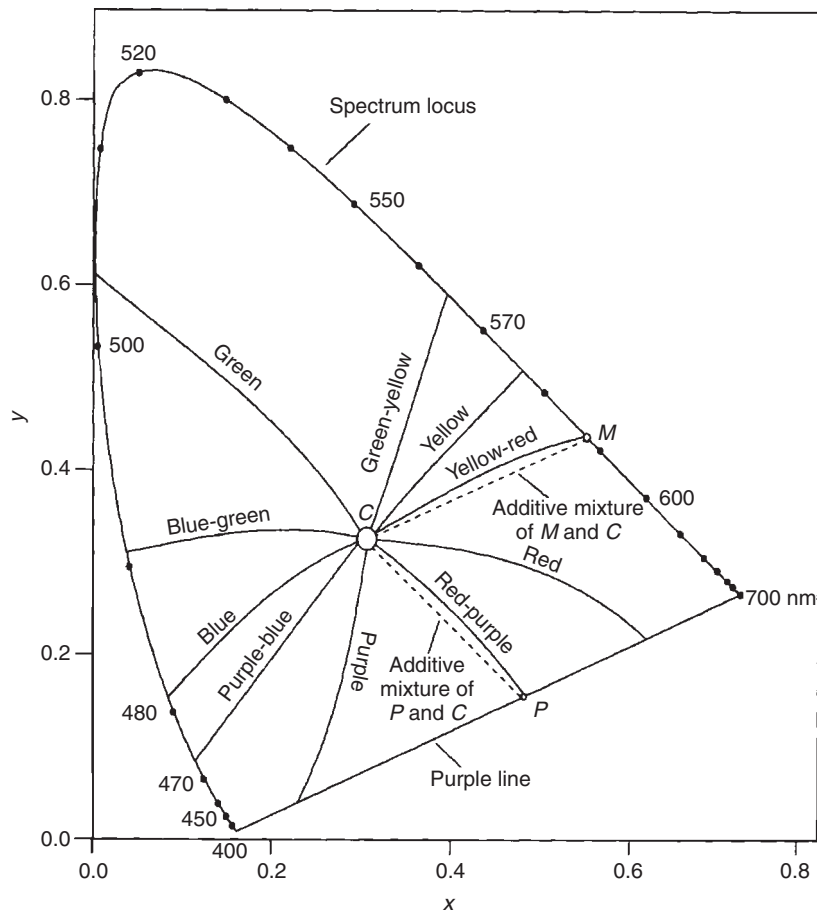
**Figure 9**   CIE color space. The abscissa and ordinate indicate the *x* and *y* values, respectively.

Another widely used color specification system is the *Munsell Book of Colors*. This classification scheme is also a variant of the color circle, but adding in a third dimension that corresponds to lightness. In the Munsell notation, the word *hue* is used as normal, but the words *value* and *chroma* are used to refer to lightness and saturation, respectively. The book contains sheets of color samples organized according to their values on the three dimensions of hue, value, and chroma. Color can be specified by reporting the values of the sample that most closely match those of the stimulus of interest.

When using a colored stimulus, one important consideration is its location in the visual field (Gegenfurtner and Sharpe, 1999). The distribution of cones varies across the retina, resulting in variations in color perception at different retinal locations. For example, because short-wavelength cones are absent in the fovea and only sparsely distributed throughout the periphery, very small blue stimuli imaged in the fovea will be seen as achromatic and the blue component in mixtures will have little impact on the perceived hue. Cones of all three types

decrease in density with increasing eccentricity, with the consequence that color perception becomes less sensitive and stimuli must be larger in order for color to be perceived. Red and green discrimination extends only $20°-30°$ into the periphery, whereas yellow and blue can be seen up to $40°-60°$ peripherally. Color vision is completely absent beyond that point.

Another consideration is that a significant portion of the population has *color blindness*, or, more generally, color vision deficiency. The most common type of color blindness is dichromatic vision. It is a gender-linked trait, with most dichromats being males. The name arises from the fact that such a person can match any spectral hue with a combination of only two primaries; in most cases this disorder can be attributed to a missing cone photopigment. The names *tritanopia, deuteranopia*, and *protanopia* refer to missing the short-, middle-, or long-wavelength pigment, respectively. The latter two types (commonly known as red–green color blindness) are much more prevalent than the former. The point to keep in mind is that color-blind persons

are not able to distinguish all of the colors that a person with trichromatic vision can. Specifically, people with red–green color blindness cannot differentiate middle and long wavelengths (520–700 nm), and the resulting perception is composed of short (blue) versus longer (yellow) wavelength hues. O'Brien et al. (2002) found that the inability to discriminate colors in certain ranges of the spectrum reduces their conspicuity (i.e., the ability to attract attention). Deuteranopes performed significantly worse than trichromats at detecting red, orange, and green color-coded traffic control devices in briefly flashed displays, but not at detecting yellow and blue color-coded signs. Testing for color vision is of importance for certain occupations, such as being an aircraft pilot, where a deficiency in color vision may lead to a crash. Tests for color deficiencies include the Ishihara plates, which require differences in color to be perceived if test patterns are to be identified, and the Farnsworth–Munsell 100-hue test, in which colored caps are to be arranged in a continuous series about four anchor point colors (Wandell, 1995). Because individuals who have some color deficiency but are not severely deficient may be able to pass one test, for example, with the Ishihara plates, multiple color vision tests seem to be necessary to detect lesser color deficiencies (Gibb et al., 2010).

## 3.2 Audition

### 3.2.1 Auditory System

The sensory receptors for hearing are sensitive to sound waves, which are moment-to-moment fluctuations in air pressure about the atmospheric level. These fluctuations are produced by mechanical disturbances, such as a stereo speaker moving in response to signals that it is receiving from a music source and amplifier. As the speaker moves forward and then back, the disturbances in the air go through phases of compression, in which the density of molecules—and hence the air pressure—is increased, and rarefaction, in which the

density and air pressure decrease. With a pure tone, such as that made by a tuning fork, these changes follow a sinusoidal pattern. The frequency of the oscillations (i.e., the number of oscillations per second) is the primary determinant of the sound's pitch, and the amplitude or intensity is the primary determinant of loudness. Intensity is usually specified in *decibels* (dB), which is $20 \log (p/p_0)$, where $p$ is the pressure corresponding to the sound and $p_0$ is the standard value of 20 µPa. When two or more pure tones are combined, the resulting sound wave will be an additive combination of the components. In that case, not only frequency and amplitude become important but also the phase relationships between the components, that is, whether the phases of the cycles for each are matched or mismatched. The wave patterns for most sounds encountered in the world are quite complex, but they can be characterized in terms of component sine waves by means of a Fourier analysis. The auditory system must perform something like a Fourier analysis, since we are capable to a large extent of extracting the component frequencies that make up a complex sound signal, so that the pitches of the component tones are heard.

***Ear*** A sound wave propagates outward from its source at the speed of sound (344 m/s), with the amplitude proportional to $1/(\text{distance})^2$. It is the cyclical air pressure changes at the ear as the sound wave propagates past the observer that starts the sensory process. The outer ear (see Figure 10), consisting of the pinna and the auditory canal, serves to funnel the sound into the middle ear; the pinna will amplify or attenuate some sounds as a function of the direction from which they come and their frequency, and the auditory canal amplifies sounds in the range of approximately 1–2 kHz. A flexible membrane, called the *eardrum* or *tympanic membrane*, separates the outer and middle ears. The pressure in the middle ear is maintained at the atmospheric level by means of the Eustachian tube, which opens into the throat, so any deviations
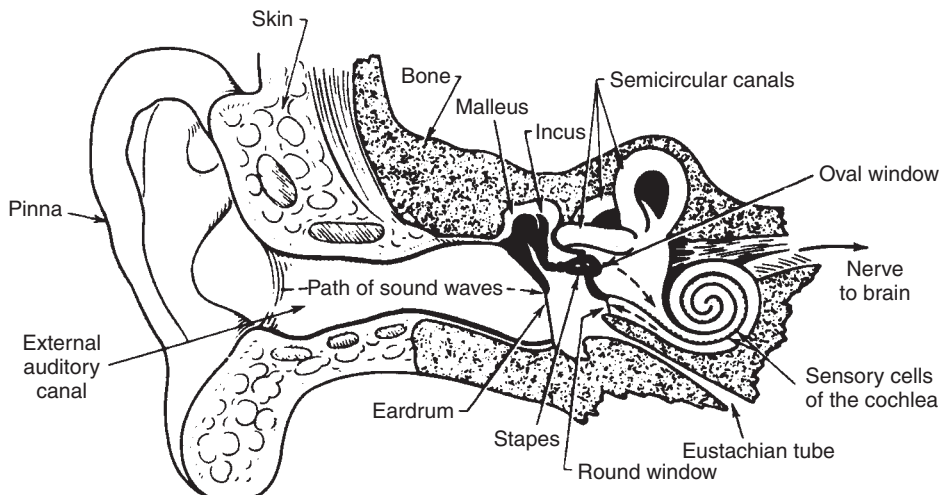


**Figure 10**    Structure of the human ear. (From Schiffman, 1996.)

from this pressure in the outer ear will result in a pressure differential that causes the eardrum to move. Consequently, the eardrum vibrates in a manner that mimics the sound wave that is affecting it. However, changes in altitude, such as those occurring during flight, can produce a pressure differential that impairs hearing until that differential is eliminated, which cannot occur readily if the Eustachian tube is blocked by infection or other causes.

Because the inner ear contains fluid, there is an impedance mismatch between it and the air that would greatly reduce the fluid movement if the eardrum acted on it directly. This impedance mismatch is overcome by a lever system of three bones (the ossicles) in the middle ear: the *malleus, incus,* and *stapes*. The malleus is attached to the eardrum and is connected to the stapes by the incus. The stapes has a footplate that is attached to a much smaller membrane, the oval window, which is at the boundary of the middle ear and the cochlea, the part of the inner ear that is important for hearing. Thus, when the eardrum moves in response to sound, the ossicles move, and the stapes produces movement of the oval window. Muscles attached to the ossicles tighten when sounds exceed 80 dB, thus protecting the inner ear to some extent from loud sounds by lessening their impact. However, because this acoustic reflex takes between 10 and 150 ms to occur, depending on the intensity of the sound, it does not provide protection from percussive sounds such as gunshots.

The *cochlea* is a fluid-filled, spiral structure (see Figure 11). It consists of three chambers, the vestibular and tympanic canals, and the cochlear duct, which separates them except at a small hole at the apex called the *helicotrema*. Part of the wall separating the cochlear duct from the tympanic canal is a flexible membrane called the *basilar membrane*. This membrane is narrower and stiffer nearer the oval window than it is nearer the helicotrema. The organ of Corti, the receptor organ that transduces the pressure changes to neural impulses, sits on the basilar membrane in the cochlear duct. It contains two groups of hair cells whose cilia project into the fluid in the cochlear duct and either touch or approach the tectorial membrane, which is inflexible. When fluid motion occurs in the inner ear, the basilar membrane vibrates, causing the cilia of the hair cells to be bent. It is this bending of the hair cells that initiates a neural signal. One group of hair cells, the inner cells, consists of a single row of approximately 3500 cells; the other group, the outer cells, is composed of approximately 12,000 hair cells arranged in three to five rows. The inner hair cells are mainly responsible for the transmission of sound information from the cochlea to the brain; the outer hair cells act as a cochlear amplifier, increasing the movement of the basilar membrane at frequencies contained in the sounds being received (Hackney, 2010). Permanent hearing loss most often is due to hair cell damage that results from excessive exposure to loud sounds or to certain drugs.

Sound causes a wave to move from the base of the basilar membrane, at the end near the oval window, to its apex. Because the width and thickness of the basilar membrane vary along its length, the magnitude of the displacement produced by this traveling wave at different locations will vary. For low-frequency sounds, the greatest movement is produced near the apex; as the frequency increases, the point of maximal displacement shifts toward the base. Thus, not only does the frequency with which the basilar membrane vibrates vary with the frequency of the auditory stimulus, but so does the location.

**Auditory Pathways** The auditory pathways after sensory transduction show many of the same properties as the visual pathways. The inner hair cells have synapses with the neurons that make up the auditory nerve. The neurons in the auditory nerve show frequency tuning. Each has a preferred or characteristic frequency that corresponds to the location on the basilar membrane of the hair cell from which it receives input but will fire less strongly to a range of frequencies about the preferred one. Neurons can be found with characteristic frequencies for virtually every frequency in the range of hearing. The contour depicting sensitivity of a neuron to different tone frequencies is called a *tuning curve*.
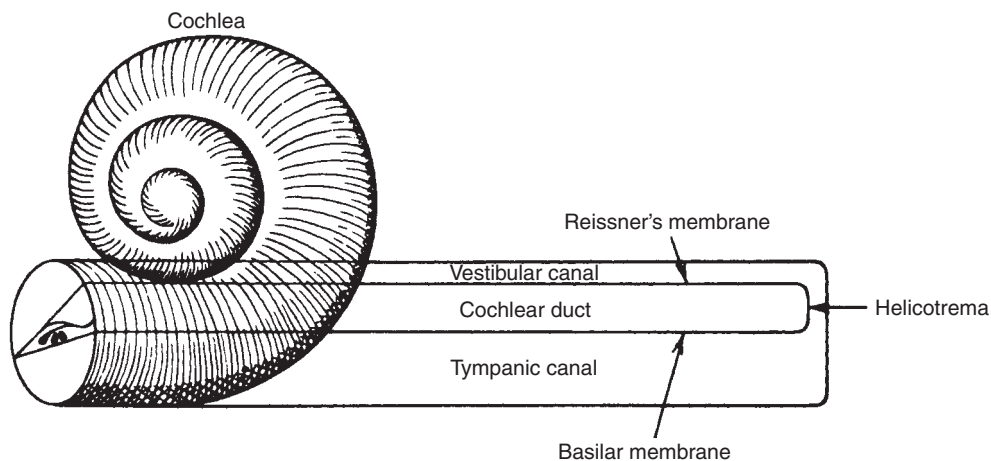


**Figure 11** Schematic of the cochlea uncoiled to show the canals. (From Schiffman, 1996.)

The tuning curves typically are broad, indicating that a neuron is sensitive to a broad range of values, but asymmetric: The sensitivity to frequencies higher than the characteristic frequency is much less than that to frequencies below it. With frequency held constant, there is a dynamic range over which as intensity is increased the neuron's firing rate will increase. This dynamic range is on the order of 25 dB, which is considerably less than the full range of intensities that we can perceive.

The first synapse for the nerve fibers after the ear is the cochlear nucleus, After that point, two separate pathways emerge that seem to have different roles, as in vision. Fibers from the anterior cochlear nucleus go to the superior olive, half to the contralateral side of the brain and half to the ipsilateral side, and then on to the inferior colliculus. This pathway is presumed to be involved in the analysis of spatial information. Fibers from the posterior cochlear nucleus project directly to the contralateral inferior colliculus. This pathway analyzes the frequency of the auditory stimulus. From the inferior colliculus, most of the neurons project to the medial geniculate and then to the primary auditory cortex. Frequency tuning is evident for neurons in all of these regions, with some neurons responding to relatively complex features of stimulation. The auditory cortex has a tonotopic organization, in which cells responsive to similar frequencies are located in close proximity, and contains

neurons tuned to extract complex information. As with vision, the signals from the auditory cortex follow two processing streams (Rauschecker, 2010). The posterior-dorsal stream analyzes *where* a sound is located, whereas the anterior-ventral stream analyzes *what* the sound represents.

### 3.2.2  Basic Auditory Perception

***Loudness and Detection of Sounds***   Loudness for audition is the equivalent of brightness for vision. More intense auditory stimuli produce greater amplitude of movement in the eardrum, which produces higher amplitude movement of the stapes on the oval window, which leads to bigger waves in the fluid of the inner ear and hence higher amplitude movements of the basilar membrane. Thus, loudness is primarily a function of the physical intensity of the stimulus and its effects on the ear, although as with brightness, it is affected by many other factors. The relation between judgments of loudness and intensity follows the power function

$$L = aI^{0.6}$$

where $L$ is loudness, $a$ is a constant, and $I$ is physical intensity.

Just as brightness is affected by the spectral properties of light, loudness is affected by the spectral properties of sound. Figure 12 shows *equal-loudness contours*
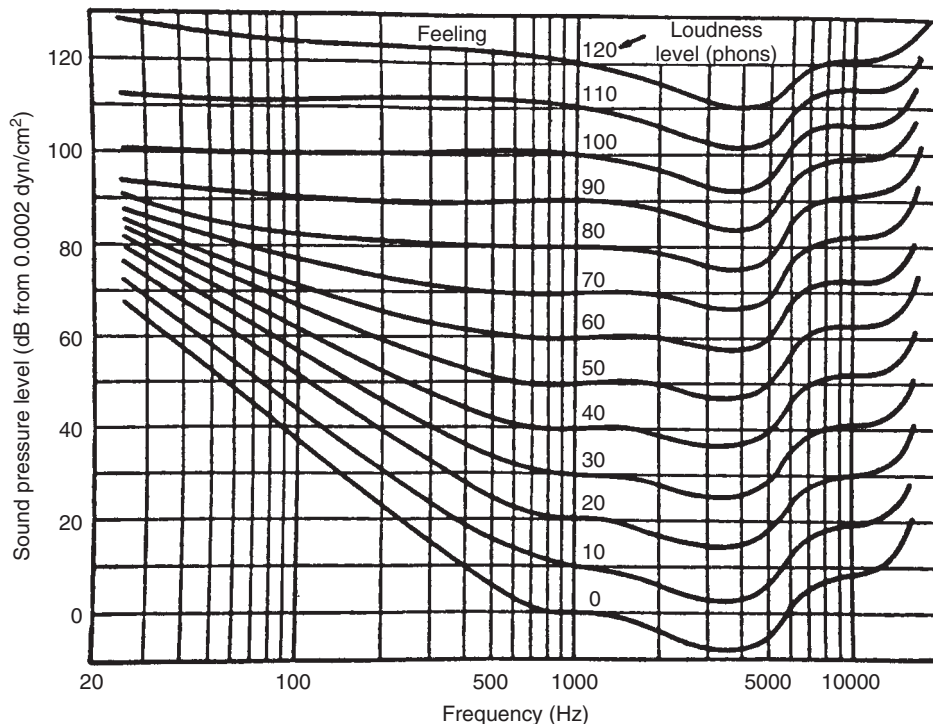


**Figure 12**  Equal-loudness contours. Each contour represents the sound pressure level at which a tone of a given frequency sounds as loud as a 1000-Hz tone of a particular intensity. (From Schiffman, 1996.)

for which a 1000-Hz tone was set at a particular intensity level and tones of other frequencies were adjusted to match its loudness. The contours illustrate that humans are relatively insensitive to low-frequency tones below approximately 200 Hz and, to a lesser extent, to high-frequency tones exceeding approximately 6000 Hz. The curves tend to flatten at high intensity levels, particularly in the low-frequency end, indicating that the insensitivity to low-frequency tones is a factor primarily at low intensity levels. This is why most audio amplifiers include a "loudness" switch for enhancing low-frequency sounds artificially when music is played at low intensities. The curves also show the maximal sensitivity to be in the range 3000–4000 Hz, which is critical for speech perception. The two most widely cited sets of equal-loudness contours are those of Fletcher and Munson (1933), obtained when listening through earphones, and of Robinson and Dadson (1956), obtained for free-field listening.

Temporal summation can occur over a brief period of approximately 200 ms, meaning that loudness is a function of the total energy presented for tones of this duration or less. The bandwidth (i.e., the range of the frequencies in a complex tone) is important for determining its loudness. With the intensity held constant, increases in bandwidth have no effect on loudness until a critical bandwidth is reached. Beyond the critical bandwidth, further increases in bandwidth result in increases in loudness.

Extraneous sounds in the environment can mask targeted sounds. This becomes important for situations such as work environments, in which audibility of specific auditory input must be evaluated with respect to the level of background noise. The degree of masking is dependent on the spectral composition of the target and noise stimuli. Masking occurs only from frequencies within the critical bandwidth. Of concern for human factors is that a masking noise will exert a much greater effect on sounds of higher frequency than on sounds of lower frequency. This asymmetry is presumed to arise primarily from the operation of the basilar membrane.

*Pitch Perception* Pitch is the qualitative aspect of sound that is a function primarily of the frequency of a periodic auditory stimulus. The higher the frequency, the higher the pitch. The pitch of a note played on a musical instrument is determined by what is called its *fundamental frequency*, but the note also contains energy at frequencies that are multiples of the fundamental frequency, called *harmonics* or *overtones*. Observers can resolve perceptually the lower harmonics of a complex tone but have more difficulty resolving the higher harmonics (Plomp, 1964). This is because the perceptual separation of the successive harmonics is progressively less as their frequency increases.

Pitch is also influenced by several factors in addition to frequency. A phenomenon of particular interest in human factors is that of the missing fundamental effect. Here, the fundamental frequency can be removed, yet the pitch of a sound remains unaltered. This suggests that pitch is based on the pattern of harmonics and not just the fundamental frequency. This phenomenon

allows a person's voice to be recognizable over the telephone and music to be played over low-fidelity systems without distorting the melody. The pitch of a tone also varies as a function of its loudness. Equal-pitch contours can be constructed much like equal-loudness contours by holding the stimulus frequency constant and varying its amplitude. Such contours show that as stimulus intensity increases, the pitch of a 3000-Hz tone remains relatively constant. However, tones whose frequencies are lower or higher than 3000 Hz show systematic decreases and increases in pitch, respectively, as intensity increases.

Two different theories were proposed in the nineteenth century to explain pitch perception. According to Ernest Rutherford's (1886) *frequency theory*, the critical factor is that the basilar membrane vibrates at the frequency of an auditory stimulus. This in turn gets transduced into neural signals at the same frequency such that the neurons in the auditory nerve respond at the frequency of the stimulus. Thus, according to this view, it is the frequency of firing that is the neural code for pitch. The primary deficiency of frequency theory is that the maximum firing rate of a neuron is restricted to about 1000 spikes/s. Thus, the firing rate of individual neurons cannot match the frequencies over much of the range of human hearing. Wever and Bray (1937) provided evidence that the range of the auditory spectrum over which frequency coding could occur can be increased by neurons that phase lock and then fire in volleys. The basic idea is that an individual neuron fires at the same phase in the cycle of the stimulus but not on every cycle. Because many neurons are responsive to the stimulus, some neurons will fire on every cycle. Thus, across the group of neurons, distinct volleys of firing will be seen that when taken together match the frequency of the stimulus. Phase locking extends the range for which frequency coding can be effective up to 4000–5000 Hz. However, at frequencies beyond this range, phase locking breaks down.

According to Hermann von Helmholtz's (1877) *place theory*, different places on the basilar membrane are affected by different frequencies of auditory stimulation. He based this proposal on his observation that the basilar membrane was tapered from narrow at the base of the cochlea to broad at its apex. This led him to suggest that it was composed of individual fibers, much like piano strings, that would resonate when the frequency of sound to which it was tuned occurred. The neurons that receive their input from a location on the membrane affected by a particular frequency would fire in its presence, whereas the neurons receiving their input from other locations would not. The neural code for frequency thus would correspond to the particular neurons that were being stimulated. However, subsequent physiological evidence showed that the basilar membrane is not composed of individual fibers.

Von Békésy (1960) provided evidence that the basilar membrane operates in a manner consistent with both frequency and place theory. Basically, he demonstrated that waves travel down the basilar membrane from the base to the apex at a frequency corresponding to that of the tone. However, because the width and thickness

of the basilar membrane vary along its length, the magnitude of the *traveling wave* is not constant over the entire membrane. The waves increase in magnitude up to a peak and then decrease abruptly. Most important, the location of the peak displacement varies as a function of frequency. Low frequencies have their maximal displacement at the apex; as frequency increases, the peak shifts systematically toward the oval window. Although most frequencies can be differentiated in terms of the place at which the peak of the traveling wave occurs, tones of less than 500–1000 Hz cannot be. Frequencies in this range produce a broad pattern of displacement, with the peak of the wave at the apex. Consequently, location coding does not seem to be possible for low-frequency tones. Because of the evidence that frequency and location coding both operate but over somewhat different regions of the auditory spectrum, it is now widely accepted that frequencies less than 4000 Hz are coded in terms of frequency and those above 500 Hz in terms of place, meaning that at frequencies within this range both mechanisms are involved.

### 3.3 Vestibular System and Sense of Balance

The vestibular system provides us with our sense of balance. It contributes to the perception of bodily motion and helps in maintaining an upright posture and the position of the eyes when head movements occur (Lackner, 2010). The sense organs for the vestibular system are contained within a part of the inner ear called the *vestibule*, which is a hollow region of bone near the oval window. The vestibular system includes the *otolith organs*, one called the *utricle* and the other the *saccule*, and three semicircular canals (see Figure 10). The otolith organs provide information about the direction of gravity and linear acceleration. The sensory receptors are hair cells lining the organs whose cilia are embedded in a gelatin-like substance that contains otoliths, which are calcium carbonate crystals. Tilting or linear acceleration of the head in any direction causes a shearing action of the otoliths on the cilia in the utricle, and vertical linear acceleration has the same effect in the saccule. The semicircular canals are placed in three perpendicular planes. They also contain hair cells that are stimulated when relative motion between the fluid inside them and the head is created and thus respond primarily to angular acceleration or deceleration in specific directions.

The vestibular ganglion contains the cell bodies of the afferent fibers of the vestibular system. The fibers project to the vestibular nucleus, where they converge with somatosensory, optokinetic, and motor-related input. These are reciprocally connected with the vestibulocerebellar cortex and nuclei in the cerebellum (Green and Angelaki, 2010). Two functions of the vestibular system, one static and one dynamic, can be distinguished. The *static function*, performed primarily by the utricle and saccule, is to monitor the position of the head in space, which is important in the control of posture. The *dynamic function*, performed primarily by the semicircular canals, is to track the rotation of the head in space. This tracking is necessary for reflexive control of what are called *vestibular eye movements*. If you maintain fixation on an object while rotating your head, the position of the eyes in the sockets will change gradually as the head moves. When your nose is pointing directly toward the object, the eyes will be centered in their sockets, but as you turn your head to the right, the eyes will rotate to the left, and vice versa as the head is turned to the right. These smooth, vestibular eye movements are controlled rapidly and automatically by the brain stem in response to sensing of the head rotation by the vestibular system.

Exposure to motions that have angular and linear accelerations substantially different from those normally encountered, as occurs in aircraft, space vehicles, and ships, can produce erroneous perceptions of attitude and angular motion that result in *spatial disorientation* (Benson, 1990). Spatial disorientation accounts for approximately 35% of all general aviation fatalities, with most occurring at night when visual cues are either absent or degraded and vestibular cues must be relied on heavily. The vestibular sense also is key to producing *motion sickness* (Kennedy et al., 2010). The dizziness and nausea associated with motion sickness are generally assumed to arise from a mismatch between the motion cues provided by the vestibular system, and possibly vision, with the expectancies of the central nervous system. The vestibular sense also contributes to the related problem of simulator sickness that arises when the visual cues in a simulator or virtual reality environment do not correspond well with the motion cues that are affecting the vestibular system (Draper et al., 2001).

### 3.4 Somatic Sensory System

The somatic sensory system is composed of four distinct modalities (Gardner et al., 2000). *Touch* is the sensation elicited by mechanical stimulation of the skin; *proprioception* is the sensation elicited by mechanical displacements of the muscles and joints; *pain* is elicited by stimuli of sufficient intensity to damage tissue; and *thermal sensations* are elicited by cool and warm stimuli. The receptors for these senses are the terminals of the peripheral branch of the axons of ganglion cells located in the dorsal root of the spinal cord. The receptors for pain and temperature, called *nociceptors* and *thermoreceptors*, are bare (or free) nerve endings. Three types of nociceptors exist that respond to different types of stimulation. Mechanical nociceptors respond to strong mechanical stimulation, thermal nociceptors respond to extreme heat or cold, and polymodal nociceptors respond to several types of intense stimuli. Distinct thermoreceptors exist for cold and warm stimuli. Those for cold stimuli respond to temperatures between 1 and $20°C$ below skin temperature, whereas those for warm stimuli respond to temperatures up to $13°C$ warmer than skin temperature,

The mechanoreceptors for touch have specialized endings that affect the dynamics of the receptor to stimulation. Some mechanoreceptor types are rapidly adapting and respond at the onset and offset of stimulation, whereas others are slow adapting and respond throughout the time that a touch stimulus is present. Hairy

skin is innervated primarily by hair follicle receptors. Hairless (glabrous) skin receives innervation from two types: Meissner's corpuscles, which are fast adapting, and Merkel's disks, which are slow adapting. Pacinian corpuscles, which are fast adapting, and Ruffini's corpuscles, which are slow adapting, are located in the dermis, subcutaneous tissue that is below both the hairy and glabrous skin.

The nerve fibers for the skin senses have a center-surround organization of the type found for vision. The receptive fields for the Meissner corpuscles and Ruffini disks are smaller than those for the Pacinian and Ruffini corpuscles, suggesting that the former provide information about fine spatial differences and the latter about coarse spatial differences. The density of mechanoreceptors is greatest for those areas of the skin, such as the fingers and lips, for which two-point thresholds (i.e., the amount of difference needed to tell that two points rather than one are being stimulated) are low. Limb proprioception is mediated by three types of receptors: mechanoreceptors located in the joints, muscle spindle receptors in muscles that respond to stretch, and cutaneous mechanoreceptors. The ability to specify limb positions decreases when the contribution of any of these receptors is removed through experimental manipulation.

The afferent fibers enter the spinal cord at the dorsal roots and follow two major pathways, the dorsal-column medial-lemniscal pathway and the anterolateral (or spinothalamic) pathway. The lemniscal pathway conveys information about touch and proprioception. It receives input primarily from fibers with corpuscles and transmits this information quickly. It ascends along the dorsal part of the spinal column, on the ipsilateral side of the body. At the brain stem, most of its fibers cross over to the contralateral side of the brain and project to the medial lemniscus in the thalamus and from there to the anterior parietal cortex. The fibers in the anterolateral pathway ascend along the contralateral side of the spinal column and project to the reticular formation, midbrain, or thalamus and then to the anterior parietal cortex and other cortical regions. This system is primarily responsible for conveying pain and temperature information.

The somatic sensory cortex is organized in a spatiotopic manner, much as is the visual cortex. That is, it is laid out in the form of a homunculus representing the opposite side of the body, with areas of the body for which sensitivity is greater, such as the fingers and lips, having relatively larger areas devoted to them. There are four different, independent spatial maps of this type in the somatic sensory cortex, with each map receiving its inputs primarily from the receptors for one of the four somatic modalities. The modalities are arranged into columns, with any one column receiving input from the same modality. When a specific point on the skin is stimulated, the population of neurons that receive innervation from that location will be activated. Each neuron has a concentric excitatory–inhibitory center-surround receptive field, the size of which varies as a function of the location on the skin. The receptive fields are smaller for regions of the body in which sensitivity

to touch is highest. Some of the cells in the somatic cortex respond to complex features of stimulation, such as movement of an object across the skin.

*Vibrotaction* has proven to be an effective way of transmitting complex information through the tactile sense (Verrillo and Gescheider, 1992). When mechanical vibrations are applied to a region of skin such as the tips of the fingers, the frequency and location of stimulation can be varied. For frequencies below 40 Hz, the size of the contactor area does not influence the absolute threshold for detecting vibration. For higher frequencies, the threshold decreases with increasing size of the contactor, indicating spatial summation of the energy within the stimulated region. Except for very small contactor areas, sensitivity reaches a maximum for vibrations of 200–300 Hz. A similar pattern of less sensitivity for low-frequency vibrations than for high-frequency vibrations is evident in equal sensation magnitude contours (Verrillo et al., 1969), much like the equal-loudness contours for audition. Because of the sensitivity to vibrotactile stimuli, it has been suggested that vibrotactile stimulation provided through the brake pedal of a vehicle may make an effective frontal collision warning system (de Rosario et al., 2010).

With multicontactor devices, which can present complex spatial patterns of stimulation, masking stimuli presented in close temporal proximity to the target stimulus can degrade identification (e.g., Craig, 1982), as in vision and audition. However, with practice, pattern recognition capabilities with these types of devices can become quite good. As a result, they can be used successfully as reading aids for the blind and to a lesser extent as hearing aids for the hearing impaired (Summers, 1992). Hollins et al. (2002) provide evidence that vibrotaction also plays a necessary and sufficient role in the perception of fine tactile textures.

A distinction is commonly made between active and passive touch (Gibson, 1966; Katz, 1989). *Passive touch* refers to situations in which a person does not move her or his hand, and the touch stimulus is applied passively, as in vibrotaction. *Active touch* refers to situations in which a person moves his or her hand intentionally to manipulate and explore an object. According to Gibson, active touch is the most common mode of acquiring tactile information in the real world and involves a unique perceptual system, which he called *haptics*. Pattern recognition with active touch typically is superior to that with passive touch (Appelle, 1991). However, the success of passive vibrotactile displays for the blind indicates that much information can also be conveyed passively. Passive and active touch can combine in a third type of touch, called *intra-active touch*, in which one body part is used to provide active stimulation to another body part, as when using a finger to roll a ball over the thumb (Bolanowski et al., 2004).

## 3.5 Gustation and Olfaction

Smell and taste are central to human perceptual experience. The taste of a good meal and the smell of perfume can be quite pleasurable. On the other hand, the taste of rancid potato chips or the smell of manure or of a paper mill can be quite noxious. In fact, odor

and taste are quite closely related, in that the taste of a substance is highly dependent on the odor it produces. This is evidenced by the changes in taste that occur when a cold reduces olfactory sensitivity. In human factors, both sensory modalities can be used to convey warnings. For example, ethylmercaptan is added to natural gas to warn of gas leaks because humans are quite sensitive to its odor. Also, as mentioned in Section 2.2.2, there is concern with environmental odors and their influence on people's moods and performance.

The sensory receptors for taste are groups of cells called *taste buds*. They line the walls of bumps on the tongue that are called *papillae*, as well as being located in the throat, the roof of the mouth, and inside the cheeks. Each taste bud is composed of several receptor cells in close arrangement. The receptor mechanism is located in projections from the top end of each cell that lie near an opening called a *taste pore*. Sensory transduction occurs when a taste solution comes in contact with the projections. The fibers from the taste receptors project to several nuclei in the brain and then to the insular cortex, located between the temporal and parietal lobes, and the limbic system.

In 1916, Henning proposed a taste tetrahedron in which all tastes were classified in terms of four primary tastes: sweet, sour, salty, and bitter. This categorization scheme has been accepted since then, although not without opposition. A fifth taste, umami, that of monosodium glutamate (MSG) and described as "savoriness," has also been suggested. People can identify this taste when the MSG is placed in water solutions, and they can also identify it in prepared foods after some training (Sinesio et al., 2009).

For smell, molecules in the air that are inhaled affect receptor cells located in the *olfactory epithelium*, a region of the nasal cavity. An *olfactory rod* extends from each receptor and goes to the surface of the epithelium. Near the end of the olfactory rod is a knob from which olfactory cilia project. These cilia are thought to be the receptor elements. Different receptor types apparently have different receptor proteins that bind the odorant molecules to the receptor. The axons from the smell receptors project to the olfactory bulb, located in the front of the brain, via the olfactory nerve. From there, the fibers project to a cluster of neural structures called the *olfactory brain*.

Olfaction shows several functional attributes (Engen, 1991). For one, a novel odor will almost always cause apprehension and anxiety. As a consequence, odors are useful as warnings. However, odors are not very effective at waking someone from sleep, which is illustrated amply by the need for smoke detectors that emit a loud auditory signal, even though the smoke itself has a distinctive odor. There also seems to be a bias to falsely detect the presence of odors and to overestimate the strength when the odor is present. Such a bias ensures that a miss is unlikely to occur when an odor signal is really present. The sense of smell shows considerable plasticity, with associations of odors to events readily learned and habituation occurring to odors of little consequence. Doty (2003) and Rouby

et al. (2002) provide detailed treatment of the perceptual and cognitive aspects of smell and taste.

## 4 HIGHER LEVEL PROPERTIES OF PERCEPTION

### 4.1 Perceptual Organization

The stimulus at the retina consists of patches of light energy that affect the photoreceptors. Yet we do not perceive patches of light. Rather, we perceive a structured world of meaningful objects. The organizational processes that affect perception go unnoticed in everyday life, until we encounter a situation in which we initially misperceive the situation in some way. When we realize this and our perception now is more veridical, we become aware that the organizational processes can be misled.

Perceptual organization is particularly important for the design of any visual display. If a symbol on a street sign is organized incorrectly, it may well go unrecognized. Similarly, if a warning signal is grouped perceptually with other displays, its message may be lost. The investigation of perceptual organization was initiated by a group of German psychologists called *Gestalt psychologists*, whose mantra was, "The whole is more than the sum of the parts." The demonstrations they provided to illustrate this point were sufficiently compelling that the concept is now accepted by all perceptual psychologists.

According to the Gestalt psychologists, the overriding principle of perceptual organization is that of *prägnanz*. The basic idea of this law is that the organizational processes will produce the simplest possible organization allowed by the conditions (Palmer, 2003). The first step in perceiving a figure requires that it be separated from the background. Any display that is viewed will be seen as a figure or figures against a background. The importance of *figure–ground* organization is illustrated clearly in figures with ambiguous figure–ground organizations, as in the well-known Ruben's vase (see Figure 13). Such figures can be organized with either the light region or the dark region seen as the figure. When a region is seen as the figure, the contour appears to be part of it. Also, the region seems to be in front of the background and takes on a distinct form. When the organization changes so that the region is now seen as the ground, its perceived relation with the other region reverses.

Clearly, when designing displays, one wants to construct them such that the figure–ground organization of the observer will correspond with what is intended. Fortunately, research has indicated factors that influence figure–ground organization. Symmetric patterns tend to be seen as the figure over asymmetric ones; a region that is surrounded completely by another tends to be seen as the figure and the surrounding region as the background; convex contours tend to be seen as the figure in preference to concave contours; the smaller of two regions tends to be seen as the figure and the larger as the ground; and a region oriented vertically or
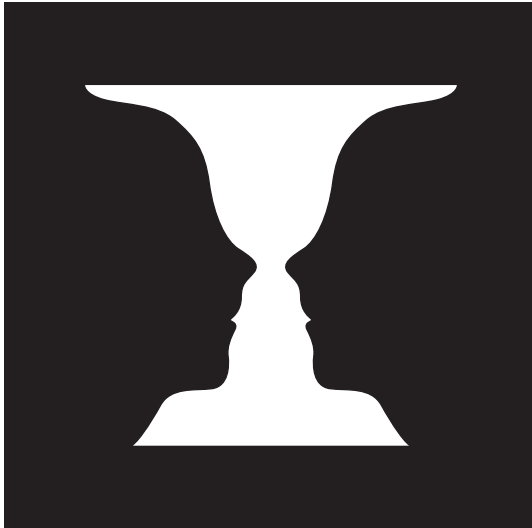
**Figure 13** Ruben's vase, for which two distinct figure–ground organizations are possible. (From Schiffman, 1996.)



**Figure 14** Gestalt organizational principles. (From Proctor and Van Zandt, 2008.)

horizontally will tend to be seen as the figure relative to one that is not so oriented.

In addition to figure–ground segregation being crucial to perception, the way that the figure is organized is important as well (see Figure 14). The most widely recognized *grouping principles* are *proximity*, display elements that are located close together will tend to be grouped together; *similarity*, display elements that are similar in appearance (e.g., orientation or color) will tend to be grouped together; *continuity*, figures will tend to be organized along continuous contours; *closure*, display elements that make up a closed figure will tend to be grouped together; and *common fate*, elements with a common motion will tend to be grouped together. Differences in orientation of stimuli seem to provide a particularly distinctive basis for grouping. As illustrated in Figure 15, when stimuli differ in orientation, those of like orientation are grouped and perceived separately from those of a different orientation. This relation lies behind the customary recommendation that displays for check reading be designed so that the pointers on the dials all have the same orientation when working properly. When something is not right, the pointer on the dial will be at an orientation different from that of the rest of the pointers, and it will "jump out" at the operator.

Two additional grouping principles (see Figure 14) were described by Rock and Palmer (1990). The principle of connectedness is that lines drawn between some elements but not others will cause the connected elements to be grouped perceptually. The principle of common region is that a contour drawn around display elements will cause those elements to be grouped together. Palmer (1992) has demonstrated several important properties of grouping by common region. When multiple, conflicting regions are present, the smaller enclosing
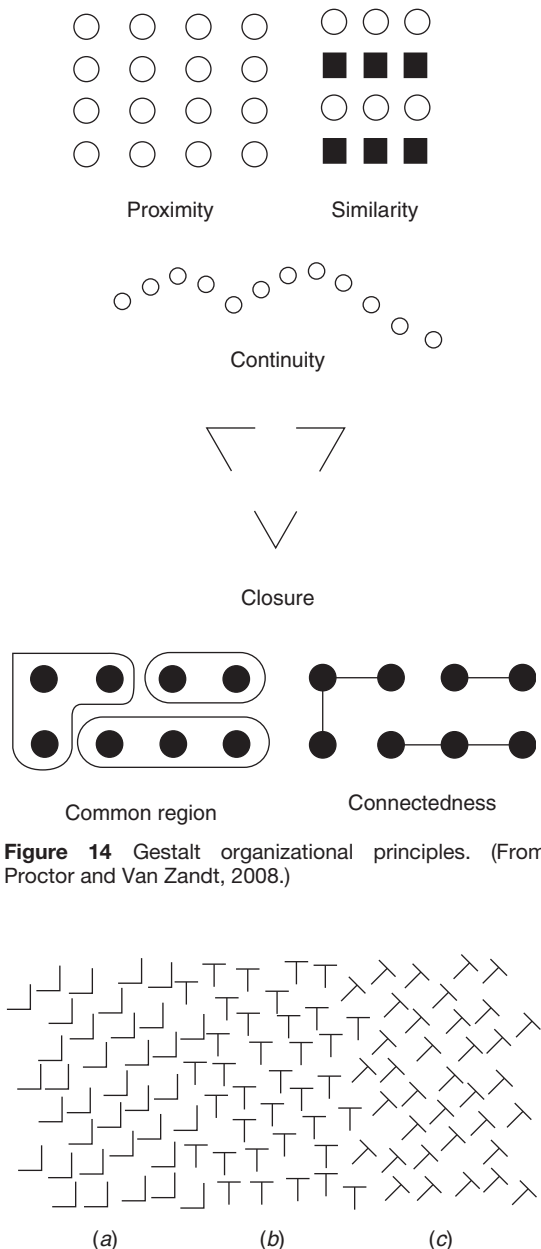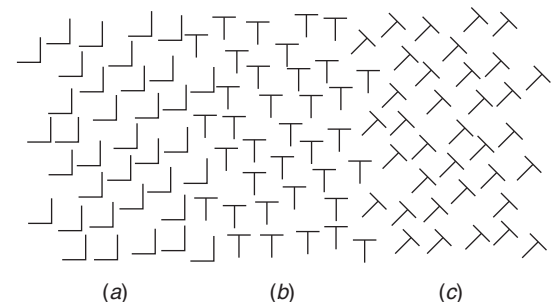


**Figure 15** Tilted-T group appears more distinct from upright T's than do backward-L characters.

region seems to dominate the organization; for nested, consistent regions, the organization appears to be hierarchical. Grouping by common region breaks down when the elements and background region are at different perceived depths, as does grouping by proximity (Rock and Brosgole, 1964), suggesting that such grouping occurs relatively late in processing, after at least some depth perception has occurred.

Although most work on perceptual organization has been conducted with visual stimuli, there are numerous demonstrations where the principles apply as well to auditory stimuli (Julesz and Hirsh, 1972). Grouping by similarity is illustrated in a study by Bregman and Rudnicky (1975) in which listeners had to indicate which of two tones of different frequency occurred first in a sequence. When the two tones were presented in isolation, performance was good. However, when preceded and followed by a single occurrence of a distractor tone of lower frequency, performance was relatively poor. The important finding is that when several occurrences of the distractor tone preceded and followed the critical pair, performance was just as good as when the two tones were presented in isolation. Apparently, the distractor tones were grouped as a distinct auditory stream based on their frequency similarity. Grouping of tones occurs not only with respect to frequency but also on the basis of similarities of their spatial positions, similarities in the fundamental frequencies and harmonics of complex tones, and so on (Bregman, 1990, 1993). Based on findings that the two-tone paradigm often yields instability of streaming under conditions that should be biased toward a particular organization, Denham et al. (2010) concluded that auditory perception is inherently multistable, but rapid switches of attention back to the dominant organization yield the experience of stability.

Another distinction that has received considerable interest over the past 35 years is that between *integral* and *separable* stimulus dimensions (Garner, 1974). The basic idea is that stimuli composed from integral dimensions are perceived as unitary wholes, whereas stimuli composed from separable dimensions are perceived in terms of their distinct dimensions. The operations used to distinguish between integral and separable dimensions are that (1) direct similarity scaling should produce a Euclidean metric for integral dimensions (i.e., the psychological distance between two stimuli should be the square root of the sum of the squares of the differences on each dimension) and a city-block metric for separable dimensions (i.e., the psychological distance should be the sum of the differences on the two dimensions); and (2) in free perceptual classification tasks, stimuli from sets with integral dimensions should be classified together if they are close in terms of the Euclidean metric (i.e., in overall similarity), whereas those from sets with separable dimensions should be classified in the same category if they match on one of the dimensions (i.e., the classifications should be in terms of dimensional structure; Garner, 1974). Perhaps most important for human factors, speed of classification with respect to one dimension is unaffected by its relation to the other dimension if the dimensions are separable but shows strong dependencies if they are integral. For integral dimensions, classifications are slowed when the value of the irrelevant dimension is uncorrelated with the value of the relevant dimension but speeded when the two dimensions are correlated.

Based on these criteria, dimensions such as hue, saturation, and lightness, in any combination, or pitch and loudness have been classified as integral; size and lightness or size and angle are classified as separable
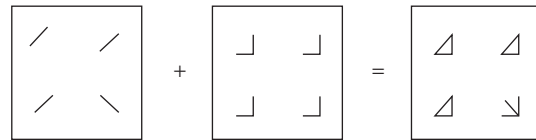


**Figure 16** Configural dimensions. The bracket context helps in discriminating the line whose slope is different from the rest.

(e.g., Shepard, 1991). A third classification, called *configural dimensions* (Pomerantz, 1981), has been proposed for dimensions that maintain their separate codes but have a new relational feature that emerges from their specific configuration. For example, as illustrated in Figure 16, a diagonally oriented line can be combined with the context of two other lines to yield an emergent triangle. Configural dimensions behave much like integral dimensions in speeded classification tasks, although the individual dimensions are still relatively accessible. Potts et al. (1998) presented evidence that the distinction between interacting (integral and configural) and noninteracting (separable) dimensions may be oversimplified. They found that with some instructions and spatial arrangements the dimensions of circle size and tilt of an enclosed line behaved as if they were separable, whereas under others they behaved as if they were integral. Thus, Potts et al. suggest that specific task contexts increase or decrease the salience of dimensional structures and may facilitate or interfere with certain processing strategies.

Wickens and his colleagues have extended the distinction between interactive dimensions (integral and configural) and separable dimensions to display design by advocating what they call the *proximity compatibility principle* (e.g., Wickens and Carswell, 1995). This principle states that if a task requires that information be integrated mentally (i.e., processing proximity is high), that information should be presented in an integral or integrated display (i.e., one with high display proximity). High display proximity can be accomplished by, for example, increasing the spatial proximity of the display elements, integrating the elements so that they appear as a distinct object, or combining them in such a way as to yield a new configural feature. The basic idea is to replace the cognitive computations that the operator must perform to combine the separate pieces of information with a much less mentally demanding pattern recognition process. The proximity compatibility principle also implies that if a task requires that the information be kept distinct mentally (i.e., processing proximity is low), the information should be presented in a display with separable dimensions (i.e., one with low display proximity). However, the cost of high display proximity for tasks that do not require integration of displayed information is typically much less than that associated with low display proximity for tasks that do require information integration.

## 4.2 Spatial Orientation

We live in a three-dimensional world and hence must be able to perceive locations in space relatively accurately

if we are to survive. Many sources of information come into play in the perception of distance and spatial relations (Proffitt and Caudek, 2003), and the consensus view is that the perceptual system constructs the three-dimensional representation using this information as cues.

### 4.2.1  Visual Depth Perception

Vision is a strongly spatial sense and provides us with the most accurate information regarding spatial location. In fact, when visual cues regarding location conflict with those from the other senses, the visual sense typically wins out, a phenomenon called *visual dominance*. There are several areas of human factors in which we need to be concerned about visual depth cues. For example, accurate depth cues are crucial for situations in which navigation in the environment is required; misleading depth cues at a landing strip at an airfield may cause a pilot to land short of the runway. For another, a helmet-mounted display, viewed through a monocle, will eliminate binocular cues and possibly provide information that conflicts with that seen by the other eye. As a final example, it may be desired that a simulator depict three-dimensional relations relatively accurately on a two-dimensional display screen.

One distinction that can be made is between oculomotor cues and visual cues. The *oculomotor cues* are accommodation and vergence angle, both of which we discussed earlier in the chapter. At relatively close distances, vergence and accommodation will vary systematically as a function of the distance of the fixated object from the observer. Therefore, either the signal sent from the brain to control accommodation and vergence angle or feedback from the muscles could provide cues to depth. However, Proffitt and Caudek (2003) conclude that neither oculomotor cue is a particularly effective cue for perceiving absolute depth and both are easily overridden when other depth cues are available.

*Visual cues* can be partitioned into binocular and monocular cues. The binocular cue is *retinal disparity*, which arises from the fact that the two eyes view an object from different locations. An object that is fixated falls on corresponding points of the retinas. This object can be regarded as being located on an imaginary curved plane, called the *horopter*; any other object that is located on this plane will also fall on corresponding points. For objects that are not on the horopter, the images will fall on disparate locations of the retinas. The direction of disparity, uncrossed or crossed (i.e., whether the image from the right eye is located to the right or left of the image from the left eye), is a function of whether the object is in back of or in front of the horopter, respectively, and the magnitude of disparity is a function of how far the object is from the horopter. Thus, retinal disparity provides information with regard to the locations of objects in space with respect to the surface that is being fixated.

The first location in the visual pathway at which neurons are sensitive to disparity differences is the primary visual cortex. However, Parker (2007) emphasizes that "generation of a full, stereoscopic depth percept is a multi-stage process that involves both dorsal and ventral cortical pathways . . . . Both pathways may contribute to perceptual judgements about stereo depth, depending on the task presented to the visual system" (p. 389).

Retinal disparity is a strong cue to depth, as witnessed by the effectiveness of three-dimensional (3D) movies and stereoscopic static pictures, which are created by presenting slightly different images to the two eyes to create disparity cues. Anyone who has seen any of the recent spate of 3D movies realizes how compelling these effects can be. They are sufficiently strong that 3D is now being incorporated into home television and entertainment systems. In addition to enhancing the perception of depth relations in displays of naturalistic scenes, stereoptic displays may be of value in assisting scientists and others in evaluating multidimensional data sets. Wickens et al. (1994) found that a three-dimensional data set could be processed faster and more accurately to answer questions that required integration of the information if the display was stereoptic than if it was not.

The fundamental problem for theories of stereopsis is that of matching. Disparity can be computed only after corresponding features at the two eyes have been identified. When viewing the natural world, each eye receives the information necessary to perceive contours and identify objects, and stereopsis could occur after monocular form recognition. However, one of the more striking findings of the past 40 years is that there do not have to be contours present in the images seen by the individual eyes in order to perceive objects in three dimensions. This phenomenon was discovered by Julesz (1971), who used random-dot stereograms in which a region of dot densities is shifted slightly in one image relative to the other. Although a form cannot be seen if only one of the two images is viewed, when each of the two images is presented to the respective eyes, a three-dimensional form emerges. Random-dot stereograms have been popularized recently through figures that utilize the autostereogram variation of this technique, in which the disparity information is incorporated in a single, two-dimensional display. That stereopsis can occur with random-dot stereograms suggests that matching of the two images can be based on dot densities.

There are many static, or pictorial, monocular cues to depth. These cues are such that people with only one eye and those who lack the ability to detect disparity differences are still able to interact with the world with relatively little loss in accuracy. The monocular cues include retinal size (i.e., larger images appear to be closer) and familiar size (e.g., a small image of a car provides a cue that the car is far away). The cue of interposition refers to an object that appears to block part of the image of another object located in front of it. Although interposition provides information that one object is nearer than another, it does not provide information about how far apart they are. Another cue comes from shading. Because light sources typically project from above, as with the sun, the location of a shadow provides a cue to depth relations. A darker shading at the bottom of a region implies that the region

is elevated, whereas one at the top of a region provides a cue that it is depressed. Aerial perspective refers to blue coloration, which appears for objects that are far away, such as is seen when viewing a mountain at a distance. Finally, the cue of linear perspective occurs when parallel lines receding into the distance, such as train tracks, converge to a point in the image.

Gibson (1950) emphasized the importance of texture gradient, which is a combination of linear perspective and relative size, in depth perception. If one looks at a textured surface such as a brick walkway, the parts of the surface (i.e., the bricks) become smaller and more densely packed in the image as they recede into the distance. The rate of this change is a function of the orientation of the surface in depth with respect to the line of sight. This texture change specifies distance on the surface, and an image of a constant size will be perceived to come from a larger object that is farther away if it occludes a larger part of the texture. Certain color gradients, such as a gradual change from red to gray, provide effective cues to depth as well (Truscianko et al., 1991).

For a stationary observer, there are plenty of cues to depth. However, cues become even richer once the observer is allowed to move. When you maintain fixation on an object and change locations, as when looking out a train window, objects in the background will move in the same direction in the image as you are moving, whereas objects in the foreground will move in the opposite direction. This cue is called *motion parallax*. When you move straight ahead, the optical flow pattern conveys information about how fast your position is changing with respect to objects in the environment. There are also numerous ways in which displays with motion can generate depth perception (Braunstein, 1976).

Of particular concern for human factors is how the various depth cues are integrated. Bruno and Cutting (1988) varied the presence or absence of four cues: relative size, height in the projection plane, interposition, and motion parallax. They found that the four cues combined additively in one direct and two indirect scaling tasks. That is, each cue supported depth perception, and the more cues that were present, the more depth was revealed. Bruno and Cutting interpreted these results as suggesting that a separate module processes each source of depth information. Landy et al. (1995) have developed a detailed model of this general nature, according to which interactions among depth cues occur for the purpose of establishing for each cue a map of absolute depth throughout the scene. The estimate of depth at each location is determined by taking a weighted average of the estimates provided by the individual cues.

Because the size of the retinal image of an object varies as a function of the distance of the object from the observer, perception of size is intimately related to perception of distance. When accurate depth cues are present, good *size constancy* results. That is, the perceived size of the object does not vary as a function of the changes in retinal image size that accompany changes in depth. One implication of this view is
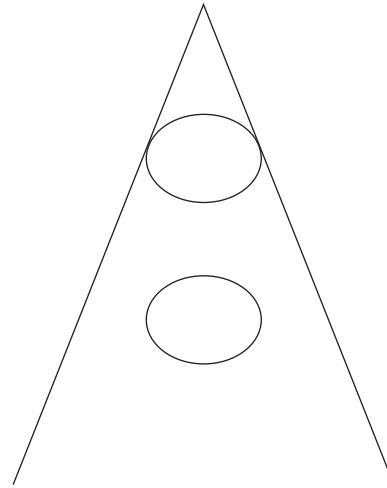


**Figure 17**  Ponzo illusion. The top circle appears larger than the lower circle, due to the linear perspective cue.

that size and shape constancy will break down and illusions appear when depth cues are erroneous. There are numerous illusions of size, such as the Ponzo illusion (see Figure 17), in which one of two stimuli of equal physical size appears larger than another, due at least in part to misleading depth cues. Misperceptions of size and distance also can arise when depth cues are minimal, as when flying at night.

### 4.2.2  Sound Localization

The cues for sound localization on the horizontal dimension involve disparities at the two ears, much as disparities of the images at the two eyes are cues to depth. Two different sources of information, interaural intensity and time differences, have been identified (Yost, 2010). Both of these cues vary systematically with respect to the position of the sound relative to the listener. At the front and back of the listener, the intensity of the sound and the time at which it reaches the ears will be equal. As the position of the sound along the azimuth (i.e., relative to the listener's head) is moved progressively toward one side or the other, the sound will become increasingly louder at the ear closest to it relative to the ear on the opposite side, and it also will reach the ipsilateral ear first. The interaural intensity differences are due primarily to a sound shadow created by the head. Because the head produces no shadow for frequencies less than 1000 Hz, the intensity cue is most effective for relatively high frequency tones. In contrast, interaural time differences are most effective for low-frequency sounds. Localization accuracy is poorest for tones between 1200 and 2000 Hz, because neither the intensity nor time cue is very effective in this intermediate-frequency range (Yost, 2010).

Both the interaural intensity and time difference cues are ambiguous because the same values can be produced by stimuli in more than one location. To locate sounds in the vertical plane and to distinguish whether the sound

is in front of or behind the listener, spectral alterations in the sound wave caused by the outer ears, head, and body (collectively called a head-related transfer function) must be relied on. Because these cues vary mainly for frequencies above 6000 Hz (Yost, 2010), front–back and vertical-location confusions of brief sounds will often occur. Confusions are relatively rare in the natural world because head movements and reflections of sound make the cues less ambiguous than they are in the typical localization experiment (e.g., Guski, 1990; Makous and Middlebrooks. 1990). As with vision, misleading cues can cause erroneous localization of sounds. Caelli and Porter (1980) illustrated this point by having listeners in a car judge the direction from which a siren occurred. Localization accuracy was particularly poor when all but one window were rolled up, which would alter the normal relation between direction and the cues.

## 4.3 Eye Movements and Motion Perception

Because details can be perceived well only at the fovea, the location on which the fovea is fixated must be able to be changed regularly and rapidly if we are to maintain an accurate perceptual representation of the environment and to see the details of new stimuli that appear in the peripheral visual field. Such changes in fixation can be brought about by displacement of the body, movements of the head, eye movements, or a combination of the three. Each eye has attached to it a set of extraocular muscle pairs: medial and lateral rectus, superior and inferior rectus, and superior and inferior obliques. Each pair controls a different axis of rotation, with the two members of the pairs acting antagonistically. Fixation is maintained when all of the muscles are active to similar extents. However, even in this case there is a continuous tremor of the eye as well as slow drifts that are corrected with compensatory micromovements, causing small changes in position of the image on the retina. Because the visual system is insensitive to images that are stabilized on the retina, such as the shadows cast by the blood vessels that support the retinal neurons, this tremor prevents images from fading when fixation is maintained on an object for a period of time.

Two broad categories of eye movements are of deepest concern. *Saccadic eye movements* involve a rapid shift in fixation from one point to another. Typically, up to three saccadic movements will be made each second (Kowler and Coolewijn, 2010). Saccadic movements can be initiated automatically by the abrupt onset of a stimulus in the peripheral visual field or voluntarily. The latency of initiation typically is on the order of 200 ms, and the duration of movement less than 100 ms. One of the more interesting phenomena associated with these eye movements is that of saccadic suppression, which is reduced sensitivity to visual stimulation during the time that the eye is moving. Saccadic suppression does not seem to be due to the movement of the retinal image being too rapid to allow perception or to masking of the image by the stationary images that precede and follow the eye movement. Rather, it seems to have a neurological basis. The loss of sensitivity is much less for high-spatial-frequency

gratings of light and dark lines than for low-spatial-frequency gratings and is absent for colored edges (Burr et al., 1994). Because lesioning studies suggest that the low spatial frequencies are conveyed primarily by the magnocellular pathway, this pathway is probably the locus of saccadic suppression.

*Smooth pursuit movements* are those made when a moving stimulus is tracked by the eyes. Such movements require that the direction of motion of the target be decoded by the system in the brain responsible for eye movements. This information must be integrated with cognitive expectancies and then translated into signals that are sent to the appropriate members of the muscle pairs of both eyes, causing them to relax and contract in unison and the eyes to move to maintain fixation on the target. Pursuit is relatively accurate for relatively slow moving targets, with increasingly greater error occurring as movement speed increases.

Eye movement records provide precise information about where a person is looking at any time. Such records have been used to obtain evidence about strategies for determining where successive saccades are directed when scanning a visual scene and about the extraction of information from the display (see Abernethy, 1988, for a review). Because direction of gaze can be recorded online by appropriate eye-tracking systems, eye gaze computer interface controls have considerable potential applications for persons with physical disabilities and for high-workload tasks (e.g., Goldberg and Schryver, 1995). It is tempting to equate direction of fixation with direction of attention, and in many cases that may be appropriate. However, there is considerable evidence that attention can be directed to different locations in space while fixation is held constant (e.g., Sanders and Houtmans, 1985), indicating that direction of fixation and direction of attention are not always one and the same.

Movements of our eyes, head, and body produce changes in position of images on the retina, as does motion of an object in the environment. How we distinguish between motion of objects in the world and our own motion has been an issue of concern for many years (Crapse and Sommer, 2008). We have already seen that many neurons in the visual cortex are sensitive to motion across the retina. However, detecting changes in position on the retina is not sufficient for motion perception, because those changes could be brought about by our own motion, motion of an object, or a combination of the two. Typically, it has been assumed that the position of the eyes is monitored by the brain, and any changes that can be attributed to eye movements are taken into account. According to inflow theory, first suggested by Sherrington (1906), it is the feedback from the muscles controlling the eyes that is monitored. According to outflow theory, first proposed by Helmholtz (1909), it is the command to the eyes to move (referred to as efference copy or corollary discharge) that is monitored. Evidence, such as that the scene appears to move when an observer who has been paralyzed tries to move her or his eyes (which do not actually move; Stevens et al., 1976; Matin et al., 1982), has tended to support the outflow theory. Crapse and Sommer (2008) present evidence that

corollary discharge is important for many aspects of perception when a person (or other organism) is moving through the world because it allows predictions of consequences of one's own movements. In their words, "CD (corollary discharge) contributes to sensorimotor harmony as primates interact with the world" (p. 552).

Sensitivity to motion is affected by many factors. For one, motion can be detected at a slower speed if a comparison, stationary object is also visible. When a reference object is present, changes of as little as $0.03°$ per second can be perceived (Palmer, 1986). However, this gain in sensitivity for detecting relative motion is at the potential cost of attributing the motion to the wrong object. For example, it is common for movement of a large region that surrounds a smaller object to be attributed to the object, a phenomenon that is called *induced motion* (Mack, 1986). The possibility for misattribution of motion is a concern for any situation in which one object is moving relative to another.

Induced motion is one example of a phenomenon in which motion of an object is perceived in the absence of motion of its image on the retina. The phenomenon of apparent, or stroboscopic, motion is probably the most important of these. This phenomenon of continuous perceived motion occurs when discrete changes in position of stimulation on the retina take place at appropriate temporal and spatial separations. It appears to be attributable to two processes, a short-range process and a long-range process (Petersik, 1989). The short-range process is presumed to reflect relatively low-level directionally sensitive neurons that respond to small spatial changes that occur with short interstimulus intervals. The long-range process is presumed to reflect higher level processes and to respond to stimuli at relatively large retinal separations presented at interstimulus intervals as long as 500 ms. Apparent motion is responsible not only for the motion produced in flashing signs but also for motion pictures and television, in which a series of discrete images is presented.

## 4.4 Pattern Recognition

The organizational principles and depth cues determine form perception, that is, what shapes and objects will be perceived. However, for the information in a display to be conveyed accurately, the objects must be recognized. If there are words, they must be read correctly; if there is a pictograph, the pictograph must be interpreted accurately. In other words, good use of the organizational principles and depth cues by a designer does not ensure that the intended message will be conveyed to the observer.

Concern with the way in which stimuli are recognized and identified is the domain of pattern recognition. Much research on pattern recognition has been conducted with verbal stimuli. The initial step in pattern recognition is typically presumed to be feature analysis. If visual, alphanumeric characters are presented, they are assumed to be analyzed in terms of features such as a vertical line segment, a horizontal line segment, and so on. Such an assumption is generally consistent

with the evidence that neurons in the primary visual cortex respond to specific features of stimulation. Evidence indicates that detection of features provides the basis for letter recognition (Pelli et al., 2006). Confusion matrices obtained when letters are misidentified indicate that an incorrect identification is most likely to involve a letter with considerable feature overlap with the one that was actually displayed (e.g., Townsend, 1971). Detailed evaluations of the features show that line terminations (e.g., the lower termination of *C* vs. *G*) and horizontal lines are most important for letter identification (Fiset et al., 2009).

Letters are composed of features, but they in turn are components of the letter patterns that form syllables and words (see Figure 18). The role played by letter-level information in visual word recognition has been the subject of considerable debate. Numerous findings have suggested that in at least some cases letter-level information is not available prior to word recognition. For example, Healy and colleagues have found that when people perform a letter detection task while reading a prose passage, the target letter is missed more often when it occurs in a very high frequency word such as *the* than when it appears in lower frequency words (e.g., Healy, 1994; Proctor and Healy, 1995). Their results have shown that this "missing-letter" effect is not just due to skipping over the words while reading. To explain these and other results, Greenberg et al. (2004) proposed a guidance-organization model of reading, which has the following properties: Unitization processes facilitate identification of function words
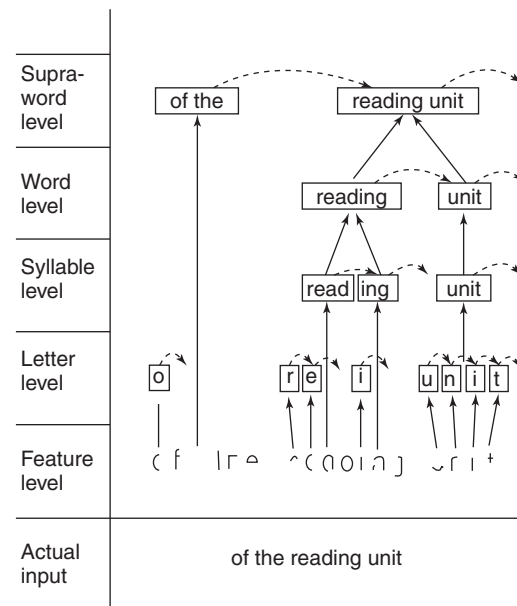


**Figure 18** Levels of representation in reading a short passage of text. Operation of the unitization hypothesis is illustrated by the bypassing of levels that occurs for "of the." (From Healy, 1994.)

that operate as cues for the structural organization of the sentence; this organization then directs attention to the content words, allowing semantic analysis and integration of meaning. In contrast to the unitization hypothesis, Pelli et al. (2003, 2006) have found that a word in isolation cannot be identified unless its letters are separately identifiable and the difficulty in identification of even common words can be predicted from the difficulties in identifying the individual letters. This has led them to conclude that people identify words as letter combinations and, even more broadly, that "everything seen is a pattern of features" (Pelli et al., 2003, p. 752).

The primary emphasis in the accounts just described is on bottom-up processing from the sensory input to recognition of the pattern, but pattern recognition is also influenced by top-down, nonvisual information of several types (Massaro and Cohen, 1994). These include orthographic constraints on the spelling patterns, regularities in the mapping between spelling and spoken sounds, syntactic constraints regarding which parts of speech are permissible, semantic constraints based on coherent meaning, and pragmatic constraints derived from the assumption that the writer is trying to communicate effectively. Interactive activation models, in which lower level sources of information are modified by higher levels, have been popular (e.g., McClelland and Rumelhart, 1981). However, Massaro and colleagues (e.g., Massaro and Cohen, 1994) have been successful in accounting for a range of reading phenomena with a model, which they call the *fuzzy logical model* of perception, in which the multiple sources of information are assumed to be processed independently, rather than interactively, and then integrated.

Reading can be viewed as a prototypical pattern recognition task. The implications of the analysis of reading are that multiple sources of information, both bottom up and top down, are exploited. For accurate pattern recognition, the possible alternatives need to be physically distinct and consistent with expectancies created by the context. More complex than reading, applied tasks such as identifying unwanted activity from computer log files involve pattern recognition, and the difficulty of these tasks can be minimized by taking pattern recognition accounts when displaying the information that goes into the log files.

## 5 SUMMARY

In this chapter we have reviewed much of what is known about sensation and perception. Any such review must necessarily exclude certain topics and be limited in the treatment given to the topics that are covered. Mather (2011) provides an accessible overview of sensation and perception that assumes no prior background, and excellent introductory texts that provide more thorough coverage include Schiffman (2001), Goldstein (2010), Sekuler and Blake (2006), and Wolfe et al. (2009). More advanced treatments of most areas are included in Volume 1 of *Stevens' Handbook of Experimental Psychology* (Pashler and Yantis, 2002) and Volume 1

of *Handbook of Perception and Human Performance* (Boff et al., 1986). *Engineering Data Compendium: Human Perception and Performance* (Boff and Lincoln, 1988) is an excellent, although now somewhat dated, resource for information pertinent to many human engineering concerns. Also, throughout the text we have provided references to texts and review articles devoted to specific topics. These and related sources should be consulted to get an in-depth understanding of the relevant issues pertaining to any particular application involving perception.

Virtually all concerns in human factors and ergonomics involve perceptual issues to at least some extent. Whether dealing with instructions for a consumer product, control rooms for chemical processing or nuclear power plants, interfaces for computer software, guidance of vehicles, office design, and so on, information of some type must be conveyed to the user or operator. To the extent that the characteristics of the sensory systems and the principles of perception are accommodated in the design of displays and the environments in which the human must work, the transmission of information to the human will be fast and accurate and the possibility for injury low. To the extent that they are not accommodated, the opportunity for error and the potential for damage are increased.

## REFERENCES

Abernethy, B. (1988), "Visual Search in Sport and Ergonomics: Its Relationship to Selective Attention and Performer Expertise," *Human Performance*, Vol. 4, pp. 205–235.

Adelson, E. H. (1993), "Perceptual Organization and the Judgment of Brightness," *Science*, Vol. 262, pp. 2042–2044.

Agostini, T., and Proffitt, D. R. (1993), "Perceptual Organization Evokes Simultaneous Lightness Contrast," *Perception*, Vol. 22, pp. 263–272.

Andre, J. (2003), "Controversies Concerning the Resting State of Accommodation," in *Visual Perception: The Influence of H. W. Leibowitz*, J. Andre, D. A. Owens, and L. O. Harvey, Jr., Eds., American Psychological Association, Washington, DC, pp. 69–79.

Appelle, S. (1991), "Haptic Perception of Form: Activity and Stimulus Attributes," in *The Psychology of Touch*, M. A. Heller and W. Schiff, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 169–188.

Araragi, Y., Ito, H., and Sunaga, S. (2009), "Perceptual Filling-in of a Line Segment Presented on Only One Side of the Blind Spot," *Spatial Vision*, Vol. 22, pp. 339–353.

Arend, L. E. (1993), "Mesopic Lightness, Brightness, and Brightness Contrast," *Perception and Psychophysics*, Vol. 54, pp. 469–476.

Baird, J. C., and Berglund, B. (1989), "Thesis for Environmental Psychophysics," *Journal of Environmental Psychology*, Vol. 9, pp. 345–356.

Barber, A. V. (1990), "Visual Mechanisms and Predictors of Far Field Visual Task Performance," *Human Factors*, Vol. 32, pp. 217–233.

Benson, A. J. (1990), "Sensory Functions and Limitations of the Vestibular System," in *Perception and Control of Self-Motion*, R. Warren and A. H. Wertheim, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 145–170.

Berglund, M. B. (1991), "Quality Assurance in Environmental Psychophysics," in *Ratio Scaling of Psychological Magnitude*, S. J. Bolanowski and G. A. Gescheider, Eds., Wiley, New York, pp. 140–162.

Blakeslee, B., Reetz, D., and McCourt, M. E. (2008), "Coming to Terms with Lightness and Brightness: Effects of Stimulus Configuration and Instructions on Brightness and Lightness Judgments," *Journal of Vision*, Vol. 8, pp. 1–14.

Boff, K. R., and Lincoln, J. E. (1988), *Engineering Data Compendium: Human Perception and Performance*, Harvey G. Armstrong Medical Research Laboratory, Wright-Patterson AFB, Dayton, OH.

Boff, K. R., Kaufman, L., and Thomas, J. P., Eds. (1986), *Handbook of Perception and Human Performance*, Vol. 1, *Sensory Processes and Perception*, Wiley, New York.

Bolanowski, S. J., and Gescheider, G. A., Eds. (1991), *Ratio Scaling of Psychological Magnitude*, Lawrence Erlbaum Associates, Mahwah, NJ.

Bolanowski, S. J., Verrillo, R. T., and McGlone, F. (2004), "Passive, Active and Intra-Active (Self) Touch," *Behavioural Brain Research*, Vol. 148, pp. 41–45.

Borg, G. (1998), *Borg's Perceived Exertion and Pain Scales*, Human Kinetics, Champaign, IL.

Braunstein, M. L. (1976), *Depth Perception through Motion*, Academic Press, New York.

Bregman, A. S. (1990), *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, Cambridge, MA.

Bregman, A. S. (1993), "Auditory Scene Analysis: Hearing in Complex Environments," in *Thinking in Sound: The Cognitive Psychology of Human Audition*, S. McAdams and E. Bigand, Eds., Oxford University Press, New York, pp. 10–36.

Bregman, A. S., and Rudnicky, A. I. (1975), "Auditory Segregation: Stream or Streams?" *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 1, pp. 263–267.

Briggs, F., and Usrey, W. M. (2010). "Visual System Structure," in *Encyclopedia of Perception*, Vol. 2, E. B. Goldstein, Ed., Sage, Thousand Oaks, CA, pp. 1130–1134.

Briggs, F., and Usrey, W. M. (2011). "Corticogeniculate Feedback and Visual Processing in the Primate." *Journal of Physiology*, Vol. 589, pp. 33–40.

Bruno, N., and Cutting, J. E. (1988), "Minimodularity and the Perception of Layout," *Journal of Experimental Psychology: General*, Vol. 117, pp. 161–170.

Burr, D. C., Morrone, M. C., and Ross, J. (1994), "Selective Suppression of the Magnocellular Visual Pathway during Saccadic Eye Movement," *Nature*, Vol. 371, pp. 511–513.

Caelli, T., and Porter, D. (1980), "On Difficulties in Localizing Ambulance Sirens," *Human Factors*, Vol. 22, pp. 719–724.

Cornsweet, T. H. (1962), "The Staircase Method in Psychophysics," *American Journal of Psychology*, Vol. 75, pp. 485–491.

Corwin, J. (1994), "On Measuring Discrimination and Response Bias: Unequal Numbers of Targets and Distractors and Two Classes of Distractors," *Neuropsychology*, Vol. 8, pp. 110–117.

Craig, J. C. (1982), "Vibrotactile Masking: A Comparison of Energy and Pattern Maskers," *Perception and Psychophysics*, Vol. 31, pp. 523–529.

Crapse, T. B., and Sommer, M. A. (2008), "Corollary Discharge Circuits in the Primate Brain," *Current Opinion in Neurobiology*, Vol. 18, pp. 552–557.

Denham, S. L., Gylmesi, K., Stefanics, G., and Winkler, I. (2010), "Stability of Perceptual Organization in Auditory Streaming," in *The Neurophysiological Bases of Auditory Perception*, E. A. Lopez-Poveda, A. R. Palmer, and R. Meddis, Eds., Springer, New York, pp. 477–488.

de Rosario, H., Louredo, M., Díaz, I., Soler, A., Gil, J. J., Solaz, J. S., and Jornet, J. (2010), "Efficacy and Feeling of a vibrotactile Frontal collision Warning implemented in a Haptic pedal," *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 13, pp. 80–91.

Donders, F. C. (1868/1969), "On the Speed of Mental Processes," W. G. Koster, Trans., *Acta Psychologica*, Vol. 30, pp. 412–431.

Doty, R. L., Ed. (2003), *Handbook of Olfaction and Gustation*, 2nd ed., Marcel Dekker, New York.

Doucet, M., Leclerc, M., Picard, F., and Niall, K. K. (2010), "Brightness and Contrast of Images with Laser-Based Video Projectors," in *Vision and Displays for Military and Security Applications*, K. K. Niall, Ed., Springer, New York, pp. 27–48.

Dowling, J. E., and Boycott, B. B. (1966), "Organization of the Primate Retina," *Proceedings of the Royal Society, Series B*, Vol. 166, pp. 80–111.

Draper, M. H., Viirre, E. S., Furness, T. A., and Gawron, V. J. (2001), "Effects of Image Scale and System Time Delay on Simulator Sickness within Head-Coupled Virtual Environments," *Human Factors*, Vol. 43, pp. 129–146.

Dzhafarov, E. N., and Colonius, H. (2005), "Psychophysics without Physics: A Purely Psychological Theory of Fechnerian Scaling in Continuous Stimulus Spaces," *Journal of Mathematical Psychology*, Vol. 49, pp. 1–50.

Eadie, T. L., and Doyle, P. C. (2002), "Direct Magnitude Estimation and Interval Scaling of Pleasantness and Severity in Dysphonic and Normal Speakers," *Journal of the Acoustical Society of America*, Vol. 112, pp. 3014–3021.

Ebenholtz, S. M. (1992), "Accommodative Hysteresis as a Function of Target-Dark Focus Separation," *Vision Research*, Vol. 32, pp. 925–929.

Eng, J. (2006), *ROC Analysis: Web-Based Calculator for ROC Curves*, Johns Hopkins University, Baltimore, MD, available: http://www.jrocfit.org, updated May 17, 2006, cited June 20, 2010.

Engen, T. (1991), *Odor Sensation and Memory*, Praeger, New York.

Evans, D. W., and Ginsburg, A. P. (1982), "Predicting Age-Related Differences in Discriminating Road Signs Using Contrast Sensitivity," *Journal of the Optical Society of America*, Vol. 72, pp. 1785–1786.

Farah, M. J. (2000), *The Cognitive Neuroscience of Vision*, Blackwell, Oxford.

Fechner, G. T. (1860/1966), *Elements of Psychophysics*, Vol. 1, H. E. Adler, Trans., Holt, Rinehart and Winston, New York.

Fiset, D., Blais, C., Arguin, M., Tadros, K., Éthier-Majcher, C., Bub, D., and Gosselin, F. (2009), "The Spatio-Temporal Dynamics of Visual Letter Recognition," *Cognitive Neuropsychology*, Vol. 26, pp. 23–35.

Fletcher, H., and Munson, W. A. (1933), "Loudness, Its Definition, Measurement, and Calculation," *Journal of the Acoustical Society of America*, Vol. 5, pp. 82–108.

Gardner, E. P., and Martin, J. H. (2000), "Coding of Sensory Information," in *Principles of Neural Science*, Vol. 4, E. R. Kandel, J. H. Schwartz, and T. M. Jessell, Eds., Elsevier, Amsterdam, pp. 411–429.

Gardner, E. P., Martin, J. H., and Jessell, T. M. (2000), "The Bodily Senses," in *Principles of Neural Science*, Vol. 4, E. R. Kandel, J. H. Schwartz, and T. M. Jessell, Eds., Elsevier, Amsterdam, pp. 430–450.

Garner, W. (1974), *The Processing of Information and Structure*, Lawrence Erlbaum Associates, Mahwah, NJ.

Gegenfurtner, K. R., and Sharpe, L. T. (1999), *Color Vision: From Genes to Perception*, Cambridge University Press, New York.

Gescheider, G. A. (1997), *Psychophysics: The Fundamentals*, 3rd ed., Lawrence Erlbaum Associates, Mahwah, NJ.

Gibb, R., Gray, R., and Scharff, L. (2010), *Aviation Visual Perception: Research, Misperception and Mishaps*, Ashgate, Burlington, VA.

Gibson, J. J. (1950), *The Perception of the Visual World*, Houghton Mifflin, Boston.

Gibson, J. J. (1966), *The Senses Considered as Perceptual Systems*, Houghton Mifflin, Boston.

Gilchrist, A. L. (1977), "Perceived Lightness Depends on Perceived Spatial Arrangement," *Science*, Vol. 195, pp. 185–187.

Gislén, A., Gustafsson, J., and Kröger, R. H. H. (2008), "The Accommodative Pupil Responses of Children and Young Adults at Low and Intermediate Levels of Ambient Illumination," *Vision Research*, Vol. 48, pp. 989–993.

Goldberg, J. H., and Schryver, J. C. (1995), "Eye-Gaze-Contingent Control of the Computer Interface: Methodology and Example for Zoom Detection," *Behavior Research Methods, Instruments, and Computers*, Vol. 27, pp. 338–350.

Goldstein, E. B. (2010), *Sensation and Perception*, 8th ed., Wadsworth, Belmont, CA.

Green, A. M., and Angelaki, D. E. (2010), "Internal Models and Neural Computation in the Vestibular System," *Experimental Brain Research*, Vol. 200, pp. 197–222.

Green, D. M., and Swets, J. A. (1966), *Signal Detection Theory and Psychophysics*, Wiley, New York; reprinted, 1974, by Krieger, Huntington, NY.

Greenberg, S. N., Healy, A. F., Koriat, A., and Kreiner, H. (2004), "The GO Model: A Reconsideration of the Role of Structural Units in Guiding and Organizing Text On Line," *Psychonomic Bulletin & Review*, Vol. 11, pp. 428–433.

Guski, R. (1990), "Auditory Localization: Effects of Reflecting Surfaces," *Perception*, Vol. 19, pp. 819–830.

Hall, D. (2010), "Brain Imaging," in *Encyclopedia of Perception*, Vol. 1, E. B. Goldstein, Ed., Sage, Thousand Oaks, CA, pp. 227–231.

Hackney, C. M. (2010), "Auditory Processing: Peripheral," in *Encyclopedia of Perception*, Vol. 1, E. B. Goldstein, Ed., Sage, Thousand Oaks, CA, pp. 180–183.

Handy, T. C., Ed. (2005), *Event-Related Potentials: A Methods Handbook*, MIT Press, Cambridge, MA.

Harvey, L. O., Jr. (2003), "Living with Uncertainty in an Uncertain World: Signal Detection Theory in the Real World," in *Visual Perception: The Influence of H. W. Leibowitz*, J. Andre, D. A. Owens, and L. O. Harvey, Jr., Eds., American Psychological Association, Washington, DC, pp. 23–41.

Healy, A. F. (1994), "Letter Detection: A Window to Unitization and Other Cognitive Processes in Reading Text," *Psychonomic Bulletin and Review*, Vol. 1, pp. 333–344.

Helmholtz, H. von (1877/1954), *On the Sensation of Tone as a Psychological Basis for the Theory of Music*, 2nd ed., A. J. Ellis, Trans. and Ed., Dover, New York.

Helmholtz, H. von (1909/1962), *Treatise on Physiological Optics*, J. C. P. Southall, Ed. and Trans., Dover, New York.

Henning, H. (1916), "Die Qualitiitsreibe des Geschmacks," *Zeitschrift für Psychologie*, Vol. 74, pp. 203–219.

Heuer, H., Hollendiek, G., Kroger, H., and Romer, T. (1989), "The Resting Position of the Eyes and the Influence of Observation Distance and Visual Fatigue on VDT Work," *Zeitschrift für Experimentelle und Angewandte Psychologie*, Vol. 36, pp. 538–566.

Hollins, M., Bensmaiea, S. J., and Roy, E. A. (2002), "Vibrotaction and Texture Perception," *Behavioural Brain Research*, Vol. 135, pp. 51–56.

Holmes, T. H., and Rahe, R. H. (1967), "The Social Readjustment Rating Scale," *Journal of Psychosomatic Research*, Vol. 11, pp. 213–218.

Hubel, D. H., and Wiesel, T. N. (1977), "Functional Architecture of Macaque Monkey Visual Cortex," *Proceedings of the Royal Society of London, Series B*, Vol. 198, pp. 1–59.

Jackson, G. R., Owsley, C., and McGwin, G., Jr. (1999), "Aging and Dark Adaptation," *Vision Research*, Vol. 39, pp. 3975–3982.

Jacobsen, A., and Gilchrist, A. L. (1988), "The Ratio Principle Holds over a Million-to-One Range of Illumination," *Perception and Psychophysics*, Vol. 43, pp. 1–6.

Jaschinski, W., Jainta, S., Hoormann, J., and Walper, N. (2007), "Objective vs Subjective Measurements of Dark Vergence," *Ophthalmic and Physiological Optics*, Vol. 27, pp. 85–92.

Jaschinski-Kruza, W. (1991), "Eyestrain in VDU Users: Viewing Distance and the Resting Position of Ocular Muscles," *Human Factors*, Vol. 33, pp. 69–83.

Julesz, B. (1971), *Foundations of Cyclopean Perception*, University of Chicago Press, Chicago.

Julesz, B., and Hirsh, I. J. (1972), "Visual and Auditory Perception: An Essay of Comparison," in *Human Communication: A Unified View*, E. E. David and P. B. Denes, Eds., McGraw-Hill, New York, pp. 283–335.

Kanwisher, N., and Duncan, J., Eds. (2004), *Functional Neuroimaging of Visual Cognition: Attention and Performance XX*, Oxford University Press, New York.

Katz, D. (1989), *The World of Touch*, L. E. Krueger, Trans., Lawrence Erlbaum Associates, Mahwah, NJ (original work published 1925).

Kennedy, R. S., Drexler, J., and Kennedy, R. C. (2010), "Research in Visually Induced Motion Sickness," *Applied Ergonomics*, Vol. 41, pp. 494–503.

Kingdom, F. A. A., and Prins, N. (2010), *Psychophysics: A Practical Introduction*, Academic Press, San Diego, CA.

Kolb, H. (1994), "The Architecture of Functional Neural Circuits in the Vertebrate Retina," *Investigative Opthalmology and Visual Science*, Vol. 35, pp. 2385–2403.

Kosnik, W. D., Sekuler, R., and Kline, D. W. (1990), "Self-Reported Visual Problems of Older Drivers," *Human Factors*, Vol. 5, pp. 597–608.

Kowler, E., and Collewijn, H. (2010), "Eye Movements: Behavioral," in *Encyclopedia of Perception*, Vol. 1, E. B. Goldstein, Ed., Sage, Thousand Oaks, CA, pp. 421–425.

Kramer, A. F., and Weber, T. (2000), "Applications of Psychophysiology to Human Factors," in *Handbook of Psychophysiology*, 2nd ed., J. T. Cacioppo, L. G. Tassinary, and G. Berntson, Eds., Cambridge University Press, New York, pp. 794–814.

Lackner, J. R. (2010), "Vestibular System," in *Encyclopedia of Perception*, Vol. 2, E. B. Goldstein, Ed., Sage, Thousand Oaks, CA, pp. 1025–1029.

Land, M., and Tatler, B. (2010), *Looking and Acting: Vision and Eye Movements in Natural Behaviour*, Oxford University Press, New York.

Landy, M. S., Maloney, L. T., Johnston, E. B., and Young, M. (1995), "Measurement and Modeling of Depth Cue Combination: In Defense of Weak Fusion," *Vision Research*, Vol. 35, pp. 389–412.

Lees, M. N., Cosman, J. D., Lee, J. D., Fricke, N., and Rizzo, M. (2010), "Translating Cognitive Neurosciences to the Driver's Operational Environment: A Neuroergonomics Approach," *American Journal of Psychology*, Vol. 123, 391–411.

Leibowitz, H. W., and Owens, D. A. (1975), "Anomalous Myopias and the Intermediate Dark Focus of Accommodation," *Science*, Vol. 189, pp. 646–648.

Lesmes, L. A., Lu, Z.-L., Baek, J., and Albright, T. D. (2010), "Bayesian Adaptive Estimation of the Contrast Sensitivity Function: The Quick CSF Method," *Journal of Vision*, Vol. 10, Number 3, Article 17, pp. 1–21.

Mack, A. (1986), "Perceptual Aspects of Motion in the Frontal Plane," in *Handbook of Perception and Performance*, Vol. 1, *Sensory Processes and Perception*, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., Wiley, New York, pp. 17-1–17-38.

Macmillan, N. A. (2002), "Signal Detection Theory," in *Stevens' Handbook of Experimental Psychology*, Vol. 4, *Methodology in Experimental Psychology*, H. Pashler and J. Wixted, Eds., Wiley, New York, pp. 43–90.

Macmillan, N. A., and Creelman, C. D. (1990), "Response Bias: Characteristics of Detection Theory, Threshold Theory, and 'Nonparametric' Indexes," *Psychological Bulletin*, Vol. 107, pp. 401–413.

Macmillan, N. A., and Creelman, C. D. (2005), *Detection Theory: A User's Guide*, 2nd ed., Lawrence Erlbaum Associates, Mahwah, NJ.

Makous, J. C., and Middlebrooks, J. C. (1990), "Two-Dimensional Sound Localization by Human Listeners," *Journal of the Acoustical Society of America*, Vol. 87, pp. 2188–2200.

Manning, D. J., and Leach, J. (2002), "Perceptual and Signal Detection Factors in Radiography," *Ergonomics*, Vol. 45, pp. 1103–1116.

Marcos, S., Moreno, E., and Navarro, R. (1999), "The Depth-of-Field of the Human Eye from Objective and Subjective Measurements," *Vision Research*, Vol. 39, pp. 2039–2049.

Marks, L. E., and Gescheider, G. A. (2002), "Psychophysical Scaling," in *Stevens' Handbook of Experimental Psychology*, Vol. 4, *Methodology in Experimental Psychology*, H. Pashler and J. Wixted, Eds., Wiley, New York, pp. 91–138.

Massaro, D. W., and Cohen, M. M. (1994), "Visual, Orthographic, Phonological, and Lexical Influences in Reading," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 20, pp. 1107–1128.

Mather, G. (2011), *Essentials of Sensation and Perception*, Routledge, New York.

Matin, L., Picoult, E., Stevens, J., Edwards, M., and McArthur, R. (1982), "Oculoparalytic Illusion: Visual-Field Dependent Spatial Mislocations by Humans Partially Paralyzed with Curare," *Science*, Vol. 216, pp. 198–201.

Matsumoto, M., and Komatsu, H. (2005), "Neural Responses in the Macaque V1 to Bar Stimuli with Various Lengths Presented on the Blind Spot," *Journal of Neurophysiology*, Vol. 93, pp. 2374–2387.

McClelland, J. L., and Rumelhart, D. E. (1981), "An Interactive Activation Model of Context Effects in Letter Perception, I: An Account of Basic Findings," *Psychological Review*, Vol. 88, pp. 375–407.

Merigan, W. H., and Maunsell, J. H. R. (1993), "How Parallel Are the Primate Visual Pathways?" *Annual Review of Neuroscience*, Vol. 16, pp. 369–402.

Milner, A. D., and Goodale, M. A. (1995), *The Visual Brain in Action*, Oxford University Press, New York.

Morgan, M. J., Watt, R. J., and McKee, S. P. (1983), "Exposure Duration Affects the Sensitivity of Vernier Acuity to Target Motion," *Vision Research*, Vol. 23, pp. 541–546.

Mort, D. J., Malhotra, P., Mannan, S. K., Rorden, C., Pambakian, A., Kennard, C., and Husain, M. (2003), "The Anatomy of Visual Neglect," *Brain*, Vol. 126, pp. 1986–1997.

Murphy, B. J. (1978), "Pattern Thresholds for Moving and Stationary Gratings During Smooth Eye Movement," *Vision Research*, Vol. 18, pp. 521–530.

O'Brien, K. A., Cole, B. L., Maddocks, J. D., and Forbes, A. B. (2002), "Color and Defective Color Vision as Factors in the Conspicuity of Signs and Signals," *Human Factors*, Vol. 44, pp. 665–675.

Oyama, T. (1987), "Perception Studies and Their Applications to Environmental Design," *International Journal of Psychology*, Vol. 22, pp. 447–451.

Palmer, J. (1986), "Mechanisms of Displacement Discrimination with and Without Perceived Movement," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 12, pp. 411–421.

Palmer, S. E. (1992), "Common Region: A New Principle of Perceptual Grouping," *Cognitive Psychology*, Vol. 24, pp. 436–447.

Palmer, S. E. (2003), "Visual Perception of Objects," in *Experimental Psychology*, A. F. Healy and R. W. Proctor, Eds., Vol. 4 in *Handbook of Psychology*, I. B. Weiner, Ed.-in-Chief, Wiley, Hoboken, NJ, pp. 179–211.

Parker, A. J. (2007), "Binocular Depth Perception and the Cerebral Cortex," *Nature Reviews Neuroscience*, Vol. 8, pp. 379–391.

Pashler, H., and Yantis, S., Eds. (2002), *Stevens' Handbook of Experimental Psychology*, 3rd ed., Vol. 1, *Sensation and Perception*, Wiley, New York.

Pelli, D. G., Farell, B., and Moore, D. C. (2003), "The Remarkable Inefficiency of Word Recognition," *Nature*, Vol. 423, pp. 752–756.

Pelli, D. G., Burns, C. W., Farell, B., and Moore-Page, D. C. (2006), "Feature Detection and Letter Identification," *Vision Research*, Vol. 46, pp. 4646–4674.

Petersik, J. T. (1989), "The Two-Process Distinction in Apparent Motion," *Psychological Bulletin*, Vol. 106, pp. 107–127.

Plomp, R. (1964), "The Ear as Frequency Analyzer," *Journal of the Acoustical Society of America*, Vol. 36, pp. 1628–1636,

Pomerantz, J. R. (1981), "Perceptual Organization in Information Processing," in *Perceptual Organization*, M. Kubovy and J. R. Pomerantz, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 141–180.

Potts, B. C., Melara, R. D., and Marks, L. E. (1998), "Circle Size and Diameter Tilt: A New Look at Integrality and Separability," *Perception and Psychophysics*, Vol. 60, pp. 101–112.

Proctor, J. D., and Healy, A. F. (1995), "Acquisition and Retention of Skilled Letter Detection," in *Learning and Memory of Knowledge and Skills: Durability and Specificity*, A. F. Healy and L. E. Bourne, Jr., Eds., Sage, Thousand Oaks, CA, pp. 282–289.

Proctor, R. W., and Van Zandt, T. (2008), *Human Factors in Simple and Complex Systems*, 2nd ed., CRC Press, Boca Raton, FL.

Proffitt, D. R., and Caudek, C. (2003), "Depth Perception and the Perception of Events," in *Experimental Psychology*, A. F. Healy and R. W. Proctor, Eds., Vol. 4 in *Handbook of Psychology*, I. B. Weiner, Ed.-in-Chief, Wiley, Hoboken, NJ, pp. 213–236.

Randle, R. (1988), "Visual Accommodation: Mediated Control and Performance," in *International Reviews of Ergonomics*, Vol. 2, D. J. Oborne, Ed., Taylor & Francis, London.

Rauschecker, J. P. (2010), "Auditory Processing: Central," in *Encyclopedia of Perception*, Vol. 1, E. B. Goldstein, Ed., Sage, Thousand Oaks, CA, pp. 176–180.

Robinson, D. W., and Dadson, M. A. (1956), "A Re-Determination of the Equal-Loudness Relations for Pure Tones," *British Journal of Applied Physics*, Vol. 7, pp. 166–181,

Rock, I., and Brosgole, L. (1964), "Grouping Based on Phenomenal Proximity," *Journal of Experimental Psychology*, Vol. 67, pp. 531–538.

Rock, I., and Palmer, S. (1990), "The Legacy of Gestalt Psychology," *Scientific American*, Vol. 263, No. 6, pp. 84–90.

Rose, D. (2006), "Psychophysical Methods," in *Research Methods in Psychology*, 3rd ed., G. M. Breakwell, S. Hammond, C. Fife-Schaw, and J. Smith, Eds., Sage, Thousand Oaks, CA, pp. 160–181.

Rosenblum, L. D. (2010), *See What I'm Saying*, W. W. Norton, NewYork.

Rouby, C., Schaal, B., Dubois, D., Gervais, R., and Holley, A., Eds. (2002), *Olfaction, Taste, and Cognition*, Cambridge University Press, New York.

Rutherford, W. (1886), "A New Theory of Hearing," *Journal of Anatomy and Physiology*, Vol. 21, pp. 166–168.

Sanders, A. F., and Houtmans, M. J. M. (1985), "Perceptual Processing Modes in the Functional Visual Field," *Acta Psychologica*, Vol. 58, pp. 251–261.

Satoh, S., and Usui, S. (2008), "Computational Theory and Applications of a Filling-in Process at the Blind Spot," *Neural Networks*, Vol. 21, pp. 1261–1271.

Scharff, L. F. V. (2003), "Sensation and Perception Research Methods," in *Handbook of Research Methods in Experimental Psychology*, S. F. Davis, Ed., Blackwell, Malden, MA, pp. 263–284.

Schiffman, H. R. (1996), *Sensation and Perception: An Integrated Approach*, 4th ed., Wiley, New York.

Schiffman, H. R. (2001), *Sensation and Perception: An Integrated Approach*, 5th ed., Wiley, New York.

Schiffman, H. R. (2003), "Psychophysics," in *Handbook of Research Methods in Experimental Psychology*, S. F. Davis, Ed., Blackwell, Malden, MA, pp. 441–469.

Sekuler, R., and Blake, R. (2006), *Perception*, 5th ed., McGraw-Hill, New York.

Shang, H., and Bishop, I. D. (2000), "Visual Thresholds for Detection, Recognition, and Visual Impact in Landscape Settings," *Journal of Environmental Psychology*, Vol. 20, pp. 125–140.

Shepard, R. N. (1991), "Integrality Versus Separability of Stimulus Dimensions," in *The Perception of Structure*, G. R. Lockhead and J. R. Pomerantz, Eds., American Psychological Association, Washington, DC, pp. 53–71.

Sherrington, C. S. (1906), *Integrative Action of the Nervous System*, Yale University Press, New Haven, CT.

Sinesio, F., Comendador, F. J., Peparaio, M., and Moneta, E. (2009), "Taste Perception of Umami-Rich Dishes in Italian Culinary Tradition," *Journal of Sensory Studies*, Vol. 24, pp. 554–580.

Smither, J. A., and Kennedy, R. S. (2010), "A Portable Device for the Assessment of Dynamic Visual Acuity," *Applied Ergonomics*, Vol. 41, pp. 266–273.

Snodgrass, J. G., and Corwin, J. (1988), "Pragmatics of Measuring Recognition Memory: Applications to Dementia and Amnesia," *Journal of Experimental Psychology: General*, Vol. 117, pp. 34–50.

Sorkin, R. D. (1999), "Spreadsheet Signal Detection," *Behavior Research Methods, Instruments, and Computers*, Vol. 31, pp. 46–54.

Sternberg, S. (1969), "The Discovery of Processing Stages: Extensions of Donders' Method," in *Attention and Performance II*, W. G. Koster, Ed., North-Holland, Amsterdam.

Stevens, J. K., Emerson, R. C., Gerstein, G. L., Kallos, T., Neufeld, G. R., Nichols, C. W., and Rosenquist, A. C. (1976), "Paralysis of the Awake Human: Visual Perceptions," *Vision Research*, Vol. 16, pp. 93–98.

Stevens, S. S. (1956), "The Direct Estimation of Sensory Magnitudes: Loudness," *American Journal of Psychology*, Vol. 69, pp. 1–25.

Stevens, S. S. (1975), *Psychophysics*, Wiley, New York.

Sturr, F., Kline, G. E., and Taub, H. A. (1990), "Performance of Young and Older Drivers on a Static Acuity Test under Photopic and Mesopic Luminance Conditions," *Human Factors*, Vol. 32, pp. 1–8.

Summers, I. R., Ed. (1992), *Tactile Aids for the Hearing Impaired*, Whurr, London.

Swets, J. A. (1986), "Indices of Discrimination or Diagnostic Accuracy: Their ROCs and Implied Models," *Psychological Bulletin*, Vol. 99, pp. 100–117.

Thurstone, L. L. (1927), "A Law of Comparative Judgment," *Psychological Review*, Vol. 34, pp. 273–286.

Townsend, J. T. (1971), "Theoretical Analysis of an Alphabetic Confusion Matrix," *Perception and Psychophysics*, Vol. 9, pp. 40–50.

Truscianko, T., Montagnon, R., and le Clerc, J. (1991), "The Role of Colour as a Monocular Depth Cue," *Vision Research*, Vol. 31, pp. 1923–1930.

Van Zandt, T. (2002), "Analysis of Response Time Distributions," in *Stevens' Handbook of Experimental Psychology*, Vol. 4, *Methodology in Experimental Psychology*, H. Pashler and J. Wixted, Eds., Wiley, New York, pp. 461–516.

Verrillo, R. T., and Gescheider, G. A. (1992), "Perception via the Sense of Touch," in *Tactile Aids for the Hearing Impaired*, I. R. Summers, Ed., Whurr, London, pp. 1–36.

Verrillo, R. T., Fraioli, A. J., and Smith, R. L. (1969), "Sensation Magnitude of Vibrotactile Stimuli," *Perception and Psychophysics*, Vol. 6, pp. 366–372.

Von Békésy, G. (1960), *Experiments in Hearing*, McGraw-Hill, New York.

Walker, B. N. (2002), "Magnitude Estimation of Conceptual Data Dimensions for Use in Sonification," *Journal of Experimental Psychology: Applied*, Vol. 8, pp. 211–221.

Wallach, H. (1972), "The Perception of Neutral Colors," in *Perception: Mechanisms and Models: Readings from Scientific American*, T. Held and W. Richards, Eds., W.H. Freeman, San Francisco, pp. 278–285.

Wandell, B. A. (1995), *Foundations of Vision*, Sinauer Associates, Sunderland, MA.

Welford, A. T., Ed. (1980), *Reaction Times*, Academic Press, London.

Wever, E. G., and Bray, C. W. (1937), "The Perception of Low Tones and the Resonance-Volley Theory," *Journal of Psychology*, Vol. 3, pp. 101–114.

Wickens, C. D., and Carswell, C. M. (1995), "The Proximity Compatibility Principle: Its Psychological Foundation and Relevance to Display Design," *Human Factors*, Vol. 37, pp. 473–494.

Wickens, C. D., Merwin, D. F., and Lin, E. (1994), "Implications of Graphics Enhancements for the Visualization of Scientific Data: Dimensional Integrality, Stereopsis, Motion, and Mesh," *Human Factors*, Vol. 36, pp. 44–61.

Wickens, T. D. (2001), *Elementary Signal Detection Theory*, Oxford University Press, New York.

Wilson, G. F. (2002), "Psychophysiological Test Methods and Procedures," in *Handbook of Human Factors Testing and Evaluation*, 2nd ed., S. G. Charlton and T. G. O'Brien, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 127–156.

Wolfe, J. M., Kluender, K. R., Levi, D. M., Bartoshuk, L. M., Herz, R. S., Klatzky, R. L., Lederman, S. J., and Merfeld, D. M. (2009), *Sensation and Perception*, 2nd ed., Sinauer Associates, Sunderland, MA.

Wood, J. M. (2002), "Age and Visual Impairment Decrease Driving Performance as Measured on a Closed-Road Circuit," *Human Factors*, Vol. 44, pp. 482–494.

Yost, W.A. (2010), "Auditory Localization: Psychophysics," in *Encyclopedia of Perception*, Vol. 1, E. B. Goldstein, Ed., Sage, Thousand Oaks, CA, pp. 170–174.

# CHAPTER 4

# SELECTION AND CONTROL OF ACTION

**Robert W. Proctor**
**Purdue University**
**West Lafayette, Indiana**


**Kim-Phuong L. Vu**
**California State University–Long Beach**
**Long Beach, California**

*Psychology's search for quantitative laws that describe human behavior is long-standing, dating back to the 1850s. A few notable successes have been achieved, including Fitts's law (1954) and the Hick–Hyman law (Hick, 1952; Hyman, 1953).*

Delaney et al. (1998)

## 1  INTRODUCTION

Research on selection and control of action has a long history, dating to at least the middle of the nineteenth century. Modern-day research in this area has developed contemporaneously with that on human factors and ergonomics (Proctor and Vu, 2010a). Influential works in both areas appeared in the period following World War II, and in many instances, people who played important roles in the development of human factors and ergonomics also made significant contributions to our understanding of selection and control of action. Two such contributions are those alluded to in the opening quote from Delaney et al. (1998), the Hick–Hyman law and Fitts's law, involving selection and control of action, respectively, which are among the few well-established quantitative laws of behavior.

Paul M. Fitts, for whom Fitts's law is named, was perhaps the most widely known of those who made significant contributions to the field of human factors and ergonomics and to basic research on human performance (Pew, 1994). He headed the Psychology Branch of the U.S. Army Airforce Aeromedical Laboratory at its founding in 1945 and is honored by a teaching award in his name given annually by the Human Factors and Ergonomics Society. Although Fitts's primary goal was the design of military aircraft and other machines to accommodate the human operator, he fully appreciated that this goal could only be accomplished against a background of knowledge of basic principles of human performance established under controlled laboratory conditions. Consequently, Fitts made many lasting empirical and theoretical contributions to knowledge concerning selection and control of action, including the quantitative law that bears his name and the principle of stimulus–response (SR) compatibility, both of which are discussed in this chapter.

Since the groundbreaking work of Fitts and others in the 1950s, much research has been conducted on selection and control of action under the headings of human performance, motor learning and control, and motor behavior, among others. Indeed, the relation between perception and action is a very active area of research in psychology and associated fields [see, e.g., the special issue of *Psychological Research* devoted to cognitive control of action (Nattkemper and Ziessler, 2004)]. In the present chapter we review some of the major findings, principles, and theories concerning selection and control of action that are relevant to designing for human use.

## 2 SELECTION OF ACTION

### 2.1 Methods

Selection of action is most often studied in choice–reaction tasks, in which a set of stimulus alternatives is mapped to a set of responses. On each trial, one or more stimuli appear and a response is to be made based on task instructions. Simple responses such as keypresses are typically used because the intent is to study the central decisions involved in selecting actions, not the motoric processes involved in executing the actions. In a choice–reaction study, response time (RT) is characteristically recorded as the primary dependent measure and error rate as a secondary measure. Among the methods used to interpret the RT data are the additive factors method and related ones that allow examination of the selective influence of variables on various processes (Sternberg, 1998). Analyses based on these methods have suggested that the primary variables affecting the duration of action selection, or response selection, processes include SR uncertainty, SR compatibility, response precuing, and sequential dependencies (Sanders, 1998). Analyses of measures in addition to mean RT and percentage of error, including RT distributions, specific types of errors that are made, and psychophysiological/neuroimaging indicators of brain functions, can also be used to obtain information about the nature of action selection.

One well-established principle of performance in choice–reaction tasks, as well as of any tasks for which speeded responses are required, is that speed can be traded for accuracy (Pachella, 1974). The speed–accuracy trade-off function (see Figure 1) can be captured by sequential sampling models of response selection, according to which information accumulates over time after stimulus onset in a later decision stage until a decision is reached (Busemeyer and Diederich, 2010). One class of such models, race models, assume that there is a separate decision unit, or counter, for each response, with the response that is ultimately selected being the one for the counter that "wins the race" and reaches threshold first (e.g., Van Zandt et al., 2000). One point of sequential sampling models is that action selection is a function of both the quality of the stimulus information, which affects the rate at which the information accumulates, and the level of the response thresholds, which is affected by instructions and other factors. Speed–accuracy trade-off methods in which subjects are induced to adopt different speed–accuracy criteria in different trial blocks, or in which biases toward one response category or another are introduced, can be used to examine details of the choice process (e.g., Band et al., 2003).

Many situations outside the laboratory require performance of multiple tasks, either in succession or concurrently. Choice RT methods can be used not only to examine action selection for single-task performance but also for conditions in which two or more task sets must be maintained, and the person is required to switch between the various tasks periodically or to perform the tasks concurrently. Because considerable research on action selection has been conducted using both single
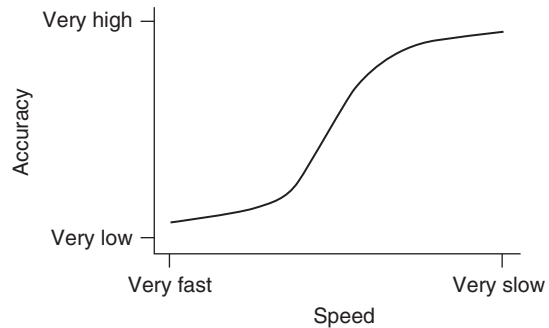


**Figure 1** Speed–accuracy trade-off. Depending on instructions, payoffs, and other factors, when a person must choose a response to a stimulus, he or she can vary the combination of response speed and accuracy between the extremes of very fast with low accuracy or very slow with high accuracy.

and multiple tasks, we cover single- and multiple-task performance separately.

### 2.2 Action Selection in Single-Task Performance

#### 2.2.1 Uncertainty and Number of Alternatives: Hick–Hyman Law

Hick (1952) and Hyman (1953), following up on much earlier work by Merkel (1885; described in Woodworth, 1938), conducted studies showing a systematic increase in choice RT as the number of SR alternatives increased. Both Hick and Hyman were interested in whether effects of SR uncertainty could be explained in terms of information theory, which Shannon (1948) had recently developed in the field of communication engineering. Information theory provides a metric for information transmission in bits (binary digits), with the number of bits conveyed by an event being a function of uncertainty. The average number of bits for a set of $N$ equally likely stimuli is $\log_2 N$. Because uncertainty also varies as a function of the probabilities with which individual stimuli occur, the average amount of information for stimuli that occur with unequal probability will be less than $\log_2 N$. More generally, the average amount of information ($H$) conveyed by a stimulus for a set of size $N$ is

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i$$

where $p_i$ is the probability of alternative $i$. Across trials, all of the information in the stimulus set is transmitted through the responses if no errors are made. However, when errors are made, the amount of transmitted information ($H_T$) will be less than the average information in the stimulus set.

The stimuli in Hick's (1952) study were 10 lamps arranged in an irregular circle, to which subjects responded by pressing one of the 10 keys, on which the fingers from each hand were placed. Hick served as his
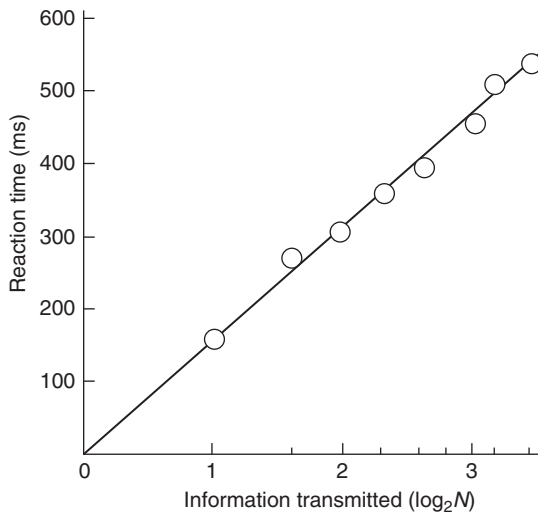
**Figure 2** Hick–Hyman law: Reaction time increases as a function of the amount of information transmitted.

own subject in two experiments (and a third control experiment). In experiment 1, Hick performed blocks of trials with set sizes ranging from 2 to 10 in ascending and descending order, maintaining a high level of accuracy. In experiment 2 he used only the set size of 10 but adopted various speed–accuracy criteria in different trial blocks. For both experiments, RT increased as a logarithmic function of the average amount of information transmitted. Hyman (1953) also manipulated the probabilities of occurrence of the alternative stimuli and sequential dependencies. In both cases, RT increased as a logarithmic function of the average amount of information conveyed by a stimulus, as predicted by information theory.

This relation between RT and the stimulus information that is transmitted in the responses is the Hick–Hyman law (see Figure 2), sometimes called Hick's law, mentioned in the opening quote of the chapter. According to it,

$$RT = a + bH_T$$

where $a$ is basic processing time and $b$ is the amount that RT increases with increases in the amount of information transmitted ($H_T$; $\log_2 N$ for equally likely SR pairs with no errors).

The Hick–Hyman function is obtained in a variety of tasks, although the slope of the function is influenced by several factors (Teichner and Krebs, 1974). The slope is typically shallower for highly compatible SR pairings than for less compatible ones (see later), and it decreases as the amount of practice at a task increases. Thus, the cost associated with high event uncertainty can be reduced by using highly compatible display–control arrangements or giving the operators training on the task. In fact, an essentially zero slope for the Hick–Hyman function, or even a decrease in RTs for larger set sizes, can be obtained with vibrotactile

stimulation of fingers requiring corresponding press responses (ten Hoopen et al., 1982), saccadic eye movements to targets (Marino and Munoz, 2009), and visually guided, aimed hand movements (Wright et al., 2007).

Usher et al. (2002) provided evidence that the Hick–Hyman law may result from subjects trying to maintain a constant accuracy for all set sizes. Usher et al. evaluated race models for which, as mentioned earlier, the response selection process is characterized as involving a separate stochastic accumulator for each SR alternative. Upon stimulus presentation, activation relevant to each alternative builds up dynamically within the respective accumulators, and when the activation in one accumulator reaches a threshold, that response is selected. Response selection is faster with lower than with higher thresholds because a threshold is reached sooner after stimulus presentation. However, this benefit in response speed is obtained at the cost of accuracy because the threshold for an incorrect alternative is more likely to be reached due to the noisy activation process.

With two SR alternatives there are two accumulators, with four alternatives there are four accumulators, and so on. Each additional accumulator provides an extra chance for an incorrect response to be selected. Consequently, if the error rate is to be held approximately constant as the size of the SR set increases, the response thresholds must be adjusted upward. Usher et al. (2002) showed that if the increase in the threshold as $N$ increases is logarithmic, the probability of an incorrect response remains approximately constant. This logarithmic increase in criterion results in a logarithmic increase in RT. Based on their model fits, Usher et al. concluded that the major determinant of the Hick–Hyman law is the increase in likelihood of erroneously reaching a response threshold as the number of SR alternatives increases, coupled with subjects attempting to keep the error rate from increasing under conditions with more alternatives.

### 2.2.2 Stimulus–Response Compatibility

**Spatial Compatibility** SR compatibility refers to the fact that some arrangements of stimuli and responses, or mappings of individual stimuli to responses, are more natural than others, leading to faster and more accurate responding (see Proctor and Vu, 2006, for a review). SR compatibility effects were demonstrated by Fitts and colleagues in two classic studies conducted in the 1950s. Specifically, Fitts and Seeger (1953) had subjects perform eight-choice tasks in which subjects moved a stylus (or a combination of two styluses) to a location in response to a stimulus. Subjects performed with each of nine combinations of three display configurations and three control configurations (see Figure 3), using the most compatible mapping of the stimulus and response elements for each combination. The primary finding was that responses were fastest and most accurate when the display and control configurations corresponded spatially than when they did not. Fitts and Deininger (1954) examined different mappings of the stimulus and response elements. In the case of circular display and control arrangements (see Figure 3a), performance was much worse with a random mapping of the eight
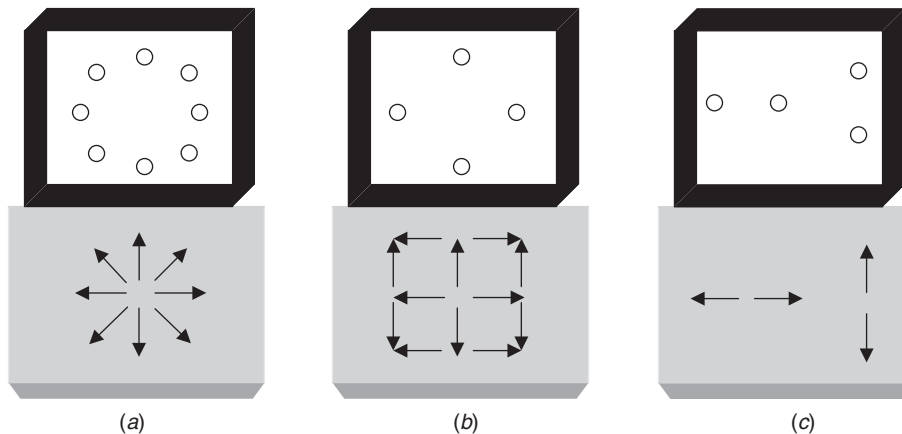
**Figure 3** Three configurations of stimulus sets and response sets used by Fitts and Seeger (1953). Displays are shown in black boxes, with the stimulus lights shown as circles. Response panels are in gray, with the directions in which one or two styluses could be moved shown by arrows. As an example, an upper right stimulus location was indicated by the upper right light for stimulus set (*a*), the upper and right lights for stimulus set (*b*), and the right light of the left pair and upper light of the right pair for stimulus set (*c*). The upper right response was a movement of a stylus to the upper right response location for response sets (*a*) and (*b*), directly for response set (*a*) and indirectly through the right or upper position for response set (*b*), or of two styluses for response set (*c*), one to the right and the other up.

stimulus locations to the eight response locations than with a spatially compatible mapping in which each stimulus was mapped to its spatially corresponding response. This finding demonstrated the basic spatial compatibility effect that has been the subject of many subsequent studies. Almost equally important, performance was much better with a mirror opposite mapping of stimuli to responses than with the random mapping. This finding implies that action selection benefits from being able to apply the same rule regardless of which stimulus occurs.

Spatial compatibility effects also occur when there are only two alternative stimulus positions, left and right, and two responses, left and right keypresses or movements of a joystick or finger, and regardless of whether the stimuli are lights or tones. Moreover, spatial correspondence not only benefits performance when stimulus location is relevant to the task but also when it is irrelevant. If a person is told to press a right key to the onset of a high pitch tone and a left key to onset of a low pitch tone, the responses are faster when the high pitch tone is in a right location (e.g., the right ear of a headphone) than when it is in a left location, and vice versa for the low pitch tone (Simon, 1990). This effect, which is found for visual stimuli as well, is known as the *Simon effect* after its discoverer, J. R. Simon. The Simon effect and its variants have attracted considerable research interest in the past 15 years because they allow examination of many fundamental issues concerning the relation between perception and action (Proctor, 2011).

**Accounts of SR Compatibility** Most accounts of SR compatibility effects attribute them to two factors. One factor is direct, or automatic, activation of the corresponding response. The other is intentional translation of the stimulus into the desired response according to the instructions that have been provided for the task. The Simon effect is attributed entirely to the automatic

activation factor, with intentional translation not considered to be involved because stimulus location is irrelevant to the task. The basic idea is that, because the response set has a spatial property, the corresponding response code is activated automatically by the stimulus at its onset, producing a tendency to select that response regardless of whether it is correct. Evidence suggests that this activation may dissipate across time, through either passive decay or active inhibition, because the Simon effect often decreases as RT becomes longer (Hommel, 1993b; De Jong et al., 1994).

In many situations, stimuli can be coded as left or right with respect to multiple frames of reference, as, for example, when there is a row of eight possible stimulus positions, four in the left hemispace and four in the right, with each of those divided into left and right pairs and left and right elements within the pairs. In such circumstances, stimulus position is coded relative to all frames of reference, with the magnitude of the Simon effect reflecting the sum of the weighted correspondence effects for each position code (e.g., Lamberts et al., 1992). Errors can result if an inappropriate reference frame is weighted more heavily than one that is relevant to the response, as appears to have been the case in the 1989 crash of a British Midland Airways Boeing 737-400 aircraft in which the operating right engine was shut down instead of the nonoperating left engine (Learmount and Norris, 1990). Confusion arose about which engine to shut down because the primary instruments for both engines were grouped in a left panel and the secondary instruments for both engines in a right panel, for which the global left and right panels were not mapped compatibly to the left and right engines (and controls).

SR compatibility proper is also presumed by many researchers to be determined in part by automatic activation of the corresponding response. As for the Simon

effect, the activated response is correct when the mapping is compatible and incorrect when the mapping is incompatible. The most influential dual-route model, that of Kornblum et al. (1990), assumes that this automatic activation occurs regardless of the SR mapping, a strong form of automaticity. However, certain results question this assumption with regard to compatibility effects (e.g., Read and Proctor, 2009), and more recent treatments of automaticity in general suggest that goal independence is not a defining feature (e.g., Moors and De Houwer, 2006). The intentional translation route is also presumed to play an important role in SR compatibility effects, with translation being fastest when a "corresponding" rule can be applied, intermediate when some other rule is applicable (e.g., respond at the opposite position), and slowest when there is no simple rule and the specific response assigned to a stimulus must be retrieved from memory.

**Dimensional Overlap** Although spatial location is an important factor influencing performance, it is by no means the only type of compatibility effect. Kornblum et al. (1990) introduced the term *dimensional overlap* to describe stimulus and response sets that are perceptually or conceptually similar. Left and right stimulus locations overlap with left and right response locations both perceptually and conceptually, and responding is fastest with the SR mapping that maintains spatial correspondence (left stimulus to left response and right stimulus to right response) than with the mapping that does not. The words "left" and "right" mapped to keypress responses also produce a compatibility effect because of the conceptual correspondence between the words and the response dimension, but the effect is typically smaller than that for physical locations due to the absence of perceptual overlap (e.g., Proctor et al., 2002).

SR compatibility and Simon effects have been obtained for a number of different stimulus types with location or direction information, for example, direction of stimulus motion (Galashan et al., 2008) and the direction of gaze of a face stimulus (Ansorge, 2003). They have also been obtained for typing letters on a keyboard, as a function of the positions in which the letters appear on a computer screen relative to the locations of the keys with which they are typed (Logan, 2003), elements in movement sequences (Inhoff et al., 1984), and clockwise versus counterclockwise rotations of a wheel (Wang et al., 2007). Properties such as the durations of stimuli and responses (short and long; Kunde and Stöcker, 2002), positive or negative affective valence of a stimulus in relation to that of a response (Duscherer et al., 2008), and pitch of a tone with that of the vowels in syllable sequences (Rosenbaum et al., 1987) also yield compatibility effects. The point is that compatibility effects are likely to occur for any situation in which the relevant or irrelevant stimulus dimension has perceptual or conceptual overlap with the response dimension.

**Influence of Action and Task Goals** It is important to understand that SR compatibility effects are determined largely by action goals and not the physical responses. This is illustrated by a study conducted by Hommel (1993a) in which subjects made a left or right keypress to a high or low pitch tone, which could occur in the left or right ear. The closure of the response key produced an action effect of turning on a light on the side opposite that on which the response was made. When instructed to turn on the left light to one tone pitch and the right light to the other, a Simon effect was obtained for which responses were faster when the tone location corresponded with the light location than when it did not, even though this condition was noncorresponding with respect to the key that was pressed. Similarly, when holding a wheel at the bottom, for which the direction of hand movements is incongruent with that of wheel movement, some subjects code the responses as left or right with respect to direction of hand movement and others with respect to direction of wheel movement, and these tendencies can be influenced to some extent by instructions that stress one response coding or the other and by controlled visual events (Guiard, 1983; Wang et al., 2007).

Compatibility effects can occur for situations in which there is no spatial correspondence relation between stimuli and responses. One such example is when stimulus and response arrays are orthogonal to each other, one being oriented vertically and the other horizontally (e.g., Proctor and Cho, 2006). Action selection when there is no spatial correspondence has been studied extensively in the literature on display–control population stereotypes, in which the main measure of interest is which action a person will choose when operating a control to achieve a desired outcome. Many studies have examined conditions in which the display is linear and the control is a rotary knob. Their results have yielded several principles relating direction of control motion to display movement (Proctor and Vu, 2010b), including:

- *Clockwise to Right/Up*. Turn the control clockwise to move the controlled element of the display to the right on a horizontal display or up on a vertical display.
- *Clockwise to Increase*. Turn the control clockwise to increase the value of the controlled element of the display.
- *Warrick's*. The controlled element of the display will move in the same direction as the side of the control nearest to the display. This principle is only applicable when the control is to the left or right of a vertical display or below or above a horizontal display.
- *Scale Side*. The controlled element of the display will move in the same direction as that of the side of the control corresponding to the side of the scale markings on the display.

Performance is most consistent when all stereotypes predict the same response to achieve an action goal. When the stereotypes are in conflict (e.g., the clockwise-to-right principle specifies clockwise rotation, whereas Warrick's principle specifies counterclockwise rotation), choices are less consistent across individuals and group differences from experience become more evident. For example, Hoffmann (1997) reported that psychology

students abide more by the clockwise-to-right principle, whereas engineering students tend to adhere to Warrick's principle.

### 2.2.3 Sequential Effects

Because many human–machine interactions involve a succession of responses, it is also important to understand how action selection is influenced by immediately preceding events, which can be evaluated by examining the sequential effects that occur in choice–reaction tasks. The most common sequential effect is that the response to a stimulus is faster when the stimulus and response are the same as those on the preceding trial than when they are not (Bertelson, 1961). This repetition benefit increases in size as the number of SR alternatives becomes larger and is greater for incompatible SR mappings than for compatible ones (Soetens, 1998). Repetition effects have been attributed to two processes, residual activation from the preceding trial when the current trial is identical to it and intentional preparation for what is expected on the next trial (Soetens, 1998). The former contributes to response selection primarily when the interval between a response and onset of the next stimulus is short, whereas the latter contributes primarily when the interval is long.

Although sequential effects with respect to the immediately preceding trial have been most widely studied, higher order repetition effects, which involve the sequence of the preceding two or three stimuli, also occur (Soetens, 1998). For two-choice tasks, at short response–stimulus intervals, where automatic activation predominates, a string of multiple repetitions is beneficial regardless of whether or not the present trial is a repetition of the immediately preceding one. In contrast, at long response–stimulus intervals, where expectancy is important, a prior string of repetition trials is beneficial if the current trial is also a repetition, and a prior string of alternation trials is beneficial if the current trial is an alternation.

When stimuli contain irrelevant stimulus information, as in the Stroop color-naming task in which the task is to name the ink color in which a conflicting color word is printed, RT is typically longer if the relevant stimulus value on a trial (e.g., the color red) is the same as that of the irrelevant information on the previous trial (e.g., the word *red*). This effect is called *negative priming* (Fox, 1995), with reference to the fact that the "priming" from the preceding trial slows RT compared to a neutral trial, for which there is no repetition of the relevant or irrelevant information from that trial. Negative priming was attributed initially to inhibition of the response tendency to the irrelevant information on the previous trial, which then carried over to the current trial. However, the situation is more complex than originally thought, and several other factors may account in whole or in part for negative priming. One such factor is that of episodic retrieval from memory (Neill and Valdes, 1992), according to which stimulus presentation initiates retrieval of the most recent episode involving the stimulus; if the relevant stimulus information was irrelevant on the previous trial, it includes an "ignore" tag, which slows responding. Another factor is that of feature mismatch

(Park and Kanwisher, 1994), according to which symbol identities are bound to objects and locations, and any change in the bindings from the preceding trial will produce negative priming. These accounts are difficult to discriminate because they make similar predictions in many situations (Christie and Klein, 2008).

Similar to negative priming, in which the irrelevant information from the previous trial interferes with processing the relevant feature on the current trial, several studies have also shown that in the Simon task noncorresponding information from the previous trial can alter how the present trial is processed. That is, the Simon effect has been shown to be evident when the preceding trial was one for which the SR locations corresponded and were absent when it was one for which they did not (e.g., Stürmer et al., 2002). A suppression/release hypothesis has been proposed to account for this pattern of results (e.g., Stürmer et al., 2002). According to this hypothesis, the Simon effect is absent following a noncorresponding trial because the direct response selection route is suppressed since automatic activation of the response code corresponding to the stimulus location would lead to the wrong response alternative. This suppression is released, though, following a corresponding trial, which results in the stimulus activating the corresponding response and thus producing a Simon effect.

However, Hommel et al. (2004) noted that the analysis on which the suppression/release hypothesis is based collapses across mapping and location repetitions and nonrepetitions. According to Hommel's (1998b) event file hypothesis, the stimulus features on a trial and the response made to them are integrated into an *event file*. When both stimulus features are repeated on the next trial, the response with which they were integrated on the previous trial is reactivated, and responding is facilitated. When both stimulus features change, neither feature was associated with the previous response, and the change in stimulus features signals a change in the response. Response selection is more difficult on trials for which one stimulus feature repeats and the other changes because one stimulus feature produces reactivation of the previous response and the other signals a change in response.

Hommel et al. (2004) and Notebaert et al. (2001) provided evidence that the pattern of repetition effects in the Simon task can be attributed to feature integration processes of the type specified by the event file hypothesis rather than to suppression/release of the automatic route. That is, responses were faster when the relevant stimulus feature and irrelevant stimulus location both repeated or both changed than when only one stimulus feature repeated. Whether the suppression/release feature integration mechanism accounts for the largest part of the sequential effects is still a matter of debate [cf. Chen and Melara (2009) and Iani et al. (2009)].

### 2.2.4 Preparation and Advance Information

When a stimulus to which a response is required occurs unexpectedly, the response to it will typically be slower than when it is expected. General preparation is studied in choice–reaction tasks by presenting a neutral warning signal at various intervals prior to onset of the imperative

stimulus. A common finding is that RT first decreases as the warning interval increases and then goes up as the warning interval is increased further, but the error rate first increases and then decreases. Bertelson (1967) demonstrated this relation in a study in which he varied the onset between an auditory warning click and a left or right visual stimulus to which a compatible keypress response was to be made. RT decreased by 20 ms for warning intervals of 0–150 ms and increased slightly as the interval increased to 300 ms, but the error rate increased from approximately 7% at the shortest intervals to about 10% at 100- and 150-ms intervals and decreased slightly at the longer intervals. Posner et al. (1973) obtained similar results for a two-choice task in which SR compatibility was manipulated, and compatibility did not interact with the warning interval. These results suggest that the warning tone alters alertness, or readiness to respond, but does not affect the rate at which the information accumulates in the response selection system.

People can also use informative cues to prepare for subsets of stimuli and responses. Leonard (1958) performed a task in which six stimulus lights were assigned compatibly to six response keys operated by the index, middle, and ring fingers of each hand. Of most concern was a condition in which either the three left or three right lights came on, precuing that subset as possible on that trial. RT decreased as precuing interval increased, being similar to that of a three-choice task when the precuing interval was 500 ms. Similar results have been obtained using four-choice tasks in which a benefit for precuing the two left or two right locations occurs within the first 500 ms of precue onset (Miller, 1982; Reeve and Proctor, 1984). However, when pairs of alternate locations are precued, a longer period of time is required to attain the maximal benefit of the precue. Reeve and Proctor (1984) showed that the benefit for precuing the two left or two right responses is also obtained when the hands are overlapped such that the index and middle fingers from the two hands are alternated, indicating that it reflects faster translation of the precued stimulus locations into possible response locations. Proctor and Reeve (1986) attributed this pattern of differential precuing benefits to the left–right distinction being salient for both stimulus and response sets, and Adam et al. (2003) proposed a grouping model that expands on this theme.

### 2.2.5  Acquisition and Transfer of Action–Selection Skill

Response selection efficiency improves with practice or training on a task. This improvement has been attributed to better pattern recognition or chunking of stimuli and responses (Newell and Rosenbloom, 1981), strengthening of associations between stimuli and responses (e.g., Anderson, 1982), and shifting from an algorithmic mode of processing to one based on retrieval of prior instances (Logan, 1988). The general idea behind all of these accounts is that practice results in performance becoming increasingly automatized. For virtually any task, the absolute benefit of a given amount of additional practice is a decreasing function of the amount of prior practice. Newell and Rosenbloom (1981) showed

that the reduction in RT with practice using the mean data from groups of subjects in a variety of tasks is characterized well by a power function:

$$RT = A + BN^{-\beta}$$

where $A$ is the asymptotic RT, $B$ the performance time on the first trial, $N$ the number of practice trials, and $\beta$ the learning rate.

Although the power function for practice has been regarded as a law to which any theory or model of skill acquisition must conform (e.g., Logan, 1988), evidence indicates that it does not provide the best fit for the practice functions of individual subjects. Heathcote et al. (2000) demonstrated that exponential functions of the following form provided better fits than power functions for individual data sets:

$$RT = A + Be^{-\alpha N}$$

where $\alpha$ is the rate parameter. In relatively complex cognitive tasks such as mental arithmetic, individual subject data often show one or more abrupt changes (e.g., Haider and Frensch, 2002; Rickard, 2004), suggesting shifts in strategy. Delaney et al. (1998) showed that in such cases the individual improvement in solution time is fit better by separate power functions for each specific strategy than by a single power function for the entire task.

As noted earlier, practice reduces the slope of the Hick–Hyman function (e.g., Hyman, 1953), indicating that the cost associated with increased SR uncertainty can be offset by allowing more practice. Seibel (1963) showed that after practice with more than 75,000 trials of all combinations of 10 lights mapped directly to 10 keys, RT for a task with 1023 alternatives was only about 25 ms slower than that for a task with 31 alternatives. Practice also benefits performance more for tasks with an incompatible SR mapping than for ones with a compatible mapping (e.g., Fitts and Seeger, 1953). However, as a general rule, performance with an incompatible mapping does not reach the same level as that with a compatible mapping for the same amount of practice (e.g., Fitts and Seeger, 1953; Dutta and Proctor, 1992).

Some evidence suggests that the improvements that occur with practice in spatial choice tasks involve primarily the mappings of the stimuli to spatial response codes and not to the specific motor effectors. Proctor and Dutta (1993) had subjects perform two-choice spatial tasks for 10 blocks of 42 trials each. In alternating trial blocks, subjects performed with their hands uncrossed or crossed such that the right hand operated the left key and the left hand the right key. There was no cost associated with alternating the hand placements for the compatible or incompatible mapping when the mapping of stimulus locations to response locations remained constant across the two hand placements. However, when the mapping of stimulus locations to response locations was switched between blocks so that the hand used to respond to a stimulus remained constant across the two hand placements, there was a substantial cost for

participants who alternated hand placements compared to those who did not, indicating the importance of maintaining a constant location mapping.

Although spatial SR compatibility effects are not eliminated by practice, transfer studies show that changes in processing that occur as one task is practiced can affect performance of a subsequent, different task. Proctor and Lu (1999) had subjects perform with an incompatible spatial mapping of left stimulus to right response and right stimulus to left response for 900 trials. When the subjects then performed a Simon task, for which stimulus location was irrelevant, the Simon effect was reversed: RT was shorter when stimulus location did not correspond with that of the response than when it did. Later studies by Tagliabue et al. (2000) and Vu et al. (2003) showed that as few as 72 practice trials with an incompatible spatial mapping eliminate the Simon effect in the transfer session and that this transfer effect remains present even a week after practice. Thus, a limited amount of practice produces new spatial SR associations that continue to affect performance at least a week later.

The transfer of a spatially incompatible mapping to the Simon task is not an automatic consequence of having executed the spatially incompatible response during practice. Vu (2011) had subjects perform 72 practice trials of a two-choice task for which stimuli occurred in a left or right location and stimulus color was nominally relevant. However, the correct response was always to the side that did not correspond to the stimulus location (i.e., if a left response was to be made to the color red, the red stimulus always occurred in the right location). Thus, the relation between stimulus and response locations was identical to that for a task with an incompatible spatial mapping; if subjects became aware of this spatially noncorresponding relation, they could base response selection on an "opposite" spatial rule instead of on stimulus color. Approximately half of the subjects indicated in a postexperiment interview that they were aware that the noncorresponding response was always correct, whereas half showed no awareness of this relation. Those subjects who were aware of the relation showed a stronger transfer effect to the Simon task than did those who showed no awareness. Related to this finding, Miles and Proctor (2010) showed that imposing an attentional load during practice with an incompatible spatial mapping eliminates transfer of this mapping to the Simon task, indicating that attention is required for the learning to occur.

## 2.3  Action Selection in Multiple-Task Performance

In many activities and jobs, people must engage in multiple tasks concurrently. This is true for an operator of a vehicle, a pilot of an aircraft, a university professor, and a secretary, among others. When more than one task set must be maintained, there is a cost in performance for all tasks even when the person devotes all of his or her attention to only one task at that time. This cost of concurrence that occurs with multiple-task performance has been of considerable interest in human factors and ergonomics. As a result, much research has been devoted to understanding and improving multiple-task performance.

### 2.3.1  Task Switching and Mixing Costs

Since the mid-1990s, there has been considerable interest in task switching (see Kiesel et al., 2010, for a review). In task-switching studies, two distinct tasks are typically performed, one at a time, with the tasks presented either in a fixed sequence (e.g., two trials of one task followed by two trials of the other task) or randomly with the current task indicated by a cue or instruction. The interval between successive trials or between the cue and the imperative stimulus can be varied to allow different amounts of time to prepare for the forthcoming task. Four phenomena are commonly obtained in such situations (Monsell, 2003):

1. *Mixing Cost.* Responses are slower overall compared to when the same task is performed on all trials. This cost represents the "global" demands required to maintain two task sets in working memory and resolve which task is to be performed on the current trial.

2. *Switch Cost.* Responses are slower on trials for which the task switches from the previous trial than for those on which it repeats. This cost reflects the "local" demands associated with switching from the previously performed task to the new task for the current trial.

3. *Preparation Benefit.* The switch cost is reduced if the next task is known, due to either following a predetermined sequence or to being cued, as long as adequate time for preparation is allowed.

4. *Residual Cost.* Although reduced in magnitude, the switch cost is not eliminated, even with adequate time for preparation.

The switch cost is typically attributed to the time needed to change the task set. The fact that the switch cost can be reduced but not eliminated by preparation is often interpreted as evidence for at least two components to the switch cost. One component involves an intentional task set reconfiguration process, and the other reflects exogenous, stimulus-driven processes, of which several have been suggested as possibilities. Rogers and Monsell (1995) proposed that this second component is a part of task set reconfiguration that cannot be accomplished until it is initiated by stimulus components related to the task. Allport et al. (1994) attributed this second component to task set inertia, with the idea that inhibition of the inappropriate task set on the previous trial carries over to the next trial, much as in negative priming. Finally, because the requirement to perform a task a few minutes later can slow performance of the current task, Waszak et al. (2003) proposed that associative retrieval of the task sets associated with the current stimulus is involved in the second component. Monsell (2003) describes the situation as follows: "Most authors now acknowledge a plurality of causes, while continuing to argue over the exact blend" (p. 137).

One important finding in the task-switching literature is that the costs associated with mixing an easy task with

a more difficult one are often larger for the easier task. For example, for Stroop stimuli, in which a color word is printed in an incongruent ink color, the costs are larger for the easy task of naming the word and ignoring the ink color than for the difficult task of naming the ink color and ignoring the word (Allport et al., 1994). Similarly, when compatible and incompatible spatial mappings are mixed within a trial block, responding is not only slowed overall, but the benefit for the compatible mapping is often eliminated (Vu and Proctor, 2008). One way to think of the reduction of the compatibility effect with mixed mappings is that the "automatic" tendency to make the corresponding response must be suppressed because it often leads to the incorrect response. An example of a task environment where operators must maintain both compatible and incompatible spatial mappings is that of driving a coal mine shuttle car, where one mapping is in effect when entering the mine and the other mapping when exiting the mine. Zupanc et al. (2007) found that in a simulated shuttle car driving task, drivers were slower at responding to critical stimuli and made more directional errors when they had to alternate between mappings, as is the case in real mines where forward and reversed maneuvers are necessary, compared to when all trials were performed with a single mapping.

Many of the results for the elimination of the SR compatibility effect with mixing are consistent with a dual-route model of the general type described earlier. According to such an account, response selection can be based on direct activation of the corresponding response when all trials are compatible, but the slower indirect route must be used when compatible trials are mixed with either incompatible trials or trials for which another stimulus dimension is relevant. The most important point for application of the compatibility principle is that the benefit for a task with a compatible mapping may not be realized when that task is mixed with other less compatible tasks.

### 2.3.2 Psychological Refractory Period Effect

Much research on multiple-task performance has focused on what is called the *psychological refractory period* (PRP) *effect* [see Pashler and Johnston (1998) and Lien and Proctor (2002) for reviews]. This phenomenon

refers to slowing of RT for the second of two tasks that are performed in rapid succession. Peripheral sensory and motor processes can contribute to decrements in dual-task performance. For example, if you are looking at the display for a compact disk changer in your car, you cannot respond to visual events that occur outside, and if you are holding a cellular phone in one hand, you cannot use that hand to respond to other events. However, research on the PRP effect has indicated that the central processes involved in action selection seem to be the locus of a major limitation in performance.

In the typical PRP study, the subject is required to perform two speeded tasks. Task 1 may be to respond to a high or low pitch tone by saying "high" or "low" out loud, and task 2 may be to respond to the location of a visual stimulus by making a left or right key press. The stimulus onset asynchrony (SOA, the interval between onsets of the task 1 stimulus, S1, and the task 2 stimulus, S2) is typically varied, either randomly within a block of trials or between blocks. The characteristic PRP effect is that RT is slowed, often considerably, for task 2 when the SOA is short (e.g., 50 ms) compared to when it is long (e.g., 800 ms).

The most widely accepted account of the PRP effect is what has been called the *central bottleneck model* (e.g., Welford, 1952; Pashler and Johnston, 1998). This model assumes that selection of the response for task 2 (R2) cannot begin until response selection for task 1 (R1) is completed (see Figure 4). The central bottleneck model has several testable implications that have tended to be confirmed by the data. First, increasing the duration of response selection for task 2 should not influence the magnitude of the PRP effect because response selection processes occur after the bottleneck. This result has been obtained in several studies in which manipulations such as SR compatibility for task 2 have been found to have additive effects with SOA, that is, to affect task 2 RT similarly at all SOAs (e.g., Pashler and Johnston, 1989; McCann and Johnston, 1992). Second, increasing the duration of stimulus identification processes for task 2 by, for example, degrading S2 should reduce the PRP effect because this increase can be absorbed into the "slack" at short SOAs after which identification of S2 is completed but response selection for the task cannot begin. This predicted underadditive
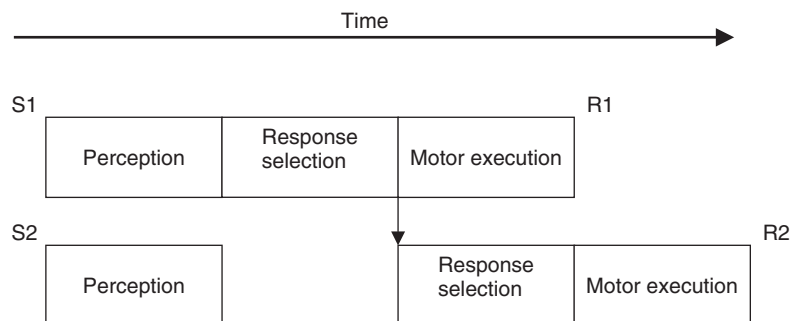


**Figure 4** Central bottleneck model. Response selection for task 2 cannot begin until that for task 1 is completed. S1 and S2 are the stimuli for tasks 1 and 2, respectively, and R1 and R2 are the responses.

interaction has been obtained in several studies (e.g., Pashler and Johnston, 1989).

The central bottleneck influences performance in tasks other than typical choice–reaction tasks. Pashler et al. (2008) found that when the second task required spontaneously choosing an action (whether to accept a card denoting a gamble), response latencies were longer at short SOAs, and this PRP effect was additive with the effects of several decision-related variables. Likewise, Levy et al. (2006) showed that in simulated driving vehicle braking was subject to dual-task slowing. Thus, evidence suggests that the bottleneck imposes a limitation on multitasking performance in real-world environments.

Numerous issues concerning the central response selection bottleneck have been investigated in recent years. One issue is whether all processes associated with action selection are subject to the bottleneck or only a subset. Consistent with the latter view, several studies have shown crosstalk correspondence effects such that the responses for both tasks 1 and 2 are faster when they correspond than when they do not (e.g., Hommel, 1998a; Lien et al., 2005), which suggests that activation of response codes occurs prior to the response selection bottleneck. A second issue is whether the bottleneck is better conceived as being of limited capacity than all-or-none. Navon and Miller (2002) and Tombu and Jolicœur (2003) have argued that the evidence is most consistent with a central-capacity sharing model, in which attentional capacity can be allocated in different amounts to response selection for the two tasks. One finding that the capacity-sharing account can explain that is difficult for the all-or-none bottleneck model is that RT for task 1, as well as that for task 2, sometimes increases at short SOAs.

Another issue is whether there is a structural bottleneck at all, or whether the bottleneck reflects a strategy adopted to perform the dual tasks as instructed. Meyer and Kieras (1997) developed a computational model, implemented within their EPIC (executive-process interactive control) architecture, which consists of perceptual, cognitive, and motor components, that does not include a limit on central-processing capacity. The specific model developed for the PRP effect, called the strategic response deferment model, includes an analysis of the processes involved in the performance of each individual task and of the executive control processes that coordinate the joint performance of the two tasks. Attention begins at the perceptual level, orienting focus (i.e., moving the eyes) on sensory input. Limits in the systems are attributed to the sensory and motor effectors, but not to the central processes. Central limitations arise from individuals' strategies for satisfying task demands (e.g., making sure that the responses for the two tasks are made in the instructed order). Specifically, according to the model, the PRP effect occurs when people adopt a conservative strategy of responding with high accuracy at the expense of speed. EPIC computational models can be developed for multitasking in real-world circumstances such as human–computer interaction and military aircraft operation as well as for the PRP effect.

The view that the bottleneck is strategic implies that it should be possible to bypass its limitations. Greenwald and Shulman (1973) provided evidence suggesting that this is the case when two tasks are "ideomotor" compatible: that is, the feedback from the response is similar to the stimulus. Their ideomotor-compatible tasks were moving a joystick to a positioned left- or right-pointing arrow (task 1) and saying the name of an auditorily presented letter (task 2). Greenwald and Shulman's experiment 2 showed no PRP effect when both tasks were ideomotor compatible, although an effect was apparent when only one task was. However, other experiments in which the two tasks were ideomotor compatible, including Greenwald and Shulman's experiment 1, have consistently shown small PRP effects (e.g., Lien et al., 2002). Regardless of whether such tasks bypass the bottleneck, dual-task interference is much smaller with two ideomotor-compatible tasks than for most other pairs of tasks.

Under certain conditions, the PRP effect can be virtually eliminated with considerable practice, a finding that some authors have interpreted as evidence against a central bottleneck (e.g., Schumacher et al., 2001). However, this elimination is accomplished primarily through the reduction of RT for task 1, which leaves open the possibility that the bottleneck is "latent" and not affecting performance (Ruthruff et al., 2003). That is, because the speed of performing task 1 improves with practice, even at short SOAs, task 1 response selection can be completed prior to the time at which response selection for task 2 is ready to begin. For practical purposes, though, the messages to take from the PRP research is that it is difficult to select different actions concurrently, but many factors can reduce the magnitude of the cost associated with trying to do so.

## 3 MOTOR CONTROL AND LEARNING

### 3.1 Methods

Whereas action selection focuses primarily on choice between action goals, motor control is concerned mainly with the execution of movements to carry out the desired actions. Tasks used to study motor control typically require movement of one or more limbs, execution of sequences of events, or control of a cursor following a target that is to be tracked. For example, a person may be asked to make an aimed movement from a start key to a target location under various conditions, and measures such as movement time and accuracy can be recorded. Some issues relevant to human factors include the nature of movement representation, the role of sensory feedback in movement execution, the way in which motor actions are sequenced, and the acquisition of perceptual–motor skills.

### 3.2 Control of Action

Motor control is achieved in two different ways, open loop and closed loop. *Open-loop control* is based on an internal model, called a *motor plan* or *motor program*, which provides a set of movement commands. Two pieces of evidence for motor plans include the

fact that deafferented monkeys, which cannot receive sensation from the deafferented limb, can still make movements including walking and climbing (e.g., Taub and Berman, 1968) and the time to initiate a movement increases as the number of elements to be performed increases (e.g., Henry and Rogers, 1960; Monsell, 1986). *Closed-loop control*, in contrast, relies on sensory feedback, comparing the feedback to a desired state and making the necessary corrections when a difference is detected. The advantages and disadvantages of open- and closed-loop control are the opposite of each other. A movement under open-loop control can be executed quickly, without a delay to process feedback, but at a cost of limited accuracy. In contrast, closed-loop control is slower but more accurate. Not surprisingly, both types of control are often combined, with open-loop control used to approximate a desired action and closed-loop control serving to reduce the deviation of the actual state from the intended state as the action is executed.

### 3.2.1 Fitts's Law

As indicated in the quote with which the chapter began, Fitts's law, which specifies the time to make aimed movements to a target location (Fitts, 1954), is one of the most widely established quantitative relations in behavioral research. As originally formulated by Fitts, the law is

$$\text{Movement time} = a + b \, \log_2(2D/W)$$

where $a$ and $b$ are constants, $D$ is the distance to the target, and $W$ is the target width (see Figure 5). Two important points of Fitts's law are that (1) movement time increases as movement distance increases and (2) movement time decreases as target width increases. It is a speed–accuracy relation in the sense that movement time must be longer when more precise movements are required. Fitts's law provides an accurate description of movement time in many situations, although alternative formulations can provide better fits for certain specific situations. The speed–accuracy relation captured by Fitts's law is a consequence of both open- and closed-loop components. Meyer et al. (1988) provided the most complete account of the relation, a stochastic optimized-submovement model. This model assumes that aimed movements consist of a primary submovement and an optional secondary submovement. Fitts's law arises as a consequence of (1) programming each movement to minimize average movement time while maintaining a high frequency of "hitting" the target and (2) making the secondary, corrective submovement when the index of difficulty is high.

Fitts's law is of considerable value in human factors because it is quite robust and is applicable to many tasks of concern to human factors professionals. The relation holds not only for tasks that require movement of a finger to a target location (e.g., when using an ATM machine) but also for tasks such as placing washers on pegs and inserting pins into holes (Fitts, 1954), using tweezers under a microscope (Langolf and Hancock, 1975), and making aimed movements underwater (Kerr, 1973). Variants of Fitts's law can also be used to model
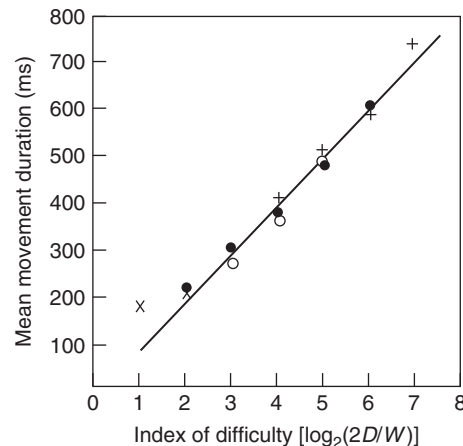


**Figure 5** Fitts's law: Movement time increases as a function of the index of difficulty [$\log_2(2D/W)$].

limb and head movements with extended probes such as screwdrivers and helmet-mounted interfaces (Baird et al., 2002). The slope of the Fitts's law function has been used to evaluate the efficiency of various ways for moving a computer cursor to a target position. For example, Card et al. (1978) showed that a computer mouse produced smaller slopes than text keys, step keys (arrows), and a joystick for the task of positioning a cursor on a desired area of text and pressing a button or key.

Size and distance are only two of many constraints that influence movement time (Heuer and Massen, in press). Movement time will be longer, for example, if the target must be grasped instead of just touched. Moreover, for objects that must be grasped, movement time depends on properties of the object, being longer to one that has to be grasped cautiously (e.g., a knife) than one that does not.

### 3.2.2 Motor Preparation and Advance Specification of Movement Properties

Movement of a limb is preceded by preparatory processes at various levels of the motor system. For a simple voluntary movement such as a keypress, a negative potential in the electroencephalogram (EEG) begins as much as 1 s before the movement itself, with this potential being stronger over the contralatateral cerebral hemisphere (which controls the finger) 100–200 ms before responding. This asymmetry, called the *lateralized readiness potential*, provides an index of being prepared to respond with a limb on one or the other side of the body (Masaki et al., 2004). In reaction tasks, this preparation may involve what is sometimes called a *response set*, or a readiness to respond, that is, response activation just below the threshold for initiating the response. However, motor preparation depends on the response that is to be performed. As noted, simple RT increases as the number of components of which the to-be-executed movement is composed increases.

Also, motor preparation is sensitive to the end state of an action. For example, when executing an action that requires grasping a bar with a pointer and placing it in a specified target position, the bar will be grasped in a manner that minimizes the awkwardness, or maximizes the comfort, of the final position in which the arm will end up (Rosenbaum et al., 1990).

Advance specification of movement parameters has been studied using a choice–RT procedure in which subjects must choose between aimed-movement responses that differ in, for example, arm (left or right), direction (toward or away), and extent (near or far). One or more parameters are precued prior to presentation of the stimulus to which the person is to respond, the idea being that RT will decrease if those parameters can be specified in advance (Rosenbaum, 1983). The results of such studies have generally supported the view that movement features can be specified in variable order; that is, there is a benefit of precuing any parameter in isolation or in combination with another. Thus, the results support those described in the section on action selection, which indicated that people can take advantage of virtually any advance information that reduces the possible stimulus and response events. A disadvantage of using the movement precuing technique to infer characteristics of parameter specification is that the particular patterns of results may be determined more by SR compatibility than by the motoric preparation process itself (e.g., Goodman and Kelso, 1980; Dornier and Reeve, 1992).

### 3.2.3   Visual Feedback

Another issue in the control of movements is the role of visual feedback. In a classic study, Woodworth (1899) had people repeatedly draw lines of a specified length on a roll of paper moving through a vertical slot in a tabletop. The rate of movement in drawing the lines was set by a metronome that beat from 20 to 200 times each minute, with one complete movement cycle to be made for each beat. Subjects performed the task with their eyes open or closed. At rates of 180 per minute or greater, movement accuracy was equivalent for the two conditions, indicating that visual feedback had no effect on performance. However, at rates of 140 per minute or less, performance was better with the eyes open. Consequently, Woodworth concluded that the minimum time required to process visual feedback was 450 ms.

Subsequent studies have reduced this estimate substantially. Keele and Posner (1968) had people perform a discrete movement of a stylus to a target that, in separate pacing conditions, was to be approximately 150, 250, 350, or 450 ms in duration. The lights turned off at the initiation of the movement on half of the trials, without foreknowledge of the performer. Movement accuracy was better with the lights turned on than off in all but the fastest pacing condition, leading Keele and Posner to conclude that the minimum duration for processing visual feedback is between 190 and 260 ms. Moreover, when people know in advance whether visual feedback will be present, results indicate that feedback can be used for movements with durations of only slightly longer than 100 ms (Zelaznik et al., 1983).

It might be thought that the role of visual feedback would decrease as a movement task is practiced, but evidence indicates that vision remains important. For example, Proteau and Cournoyer (1992) had people perform 150 trials of a task of moving a stylus to a target with either full vision, vision of both the stylus and the target, or vision of the target only. Performance during these practice trials was best with full vision and worst with vision of the target only. However, when the visual information was eliminated in a subsequent transfer block, performance was worst for those people who had practiced with full vision and worst for those who had practiced with vision of only the target. What appears to happen is that participants rely on the visual feedback for accurate performance without developing an adequate internal model of the task. Heuer and Hegele (2008) found similar results to those of Proteau and Cournoyer for a task requiring performance with a novel visuomotor gain when continuous visual feedback was provided. But when only terminal visual feedback about the final positions of the movements was provided, visuomotor adaptation to the practice conditions occurred. Thus, the kind of feedback used during practice will influence what the performer learns.

### 3.3   Coordination of Effectors

To perform many tasks well, it is necessary to coordinate the effectors. For example, when operating a manual transmission, the movements of the foot on the gas pedal must be coordinated with the shifting of gears controlled by the arm and hand. This example illustrates that one factor determining the coordination pattern is the constraints imposed by the task that is to be performed. These coordination patterns are flexible within the structural constraints imposed by the action system.

For tasks involving bimanual movements, there is a strong tendency toward mirror symmetry; that is, it is generally easy to perform symmetric movements of the arms, as in drawing two circles simultaneously with each hand. Moreover, intended asymmetric movement patterns will tend more toward symmetry in duration and timing than they should. This symmetry tendency has been studied extensively for tasks involving bimanual oscillations of the index fingers: It is easier to maintain the instructed oscillatory pattern if told to make symmetrical movements of the fingers inward and outward together than if told to make parallel movements leftward and rightward together (see Figures 6a,b). The symmetry tendency in bimanual oscillatory movements and for other bimanual tasks has traditionally been attributed to coactivation of homologous muscles (e.g., Kelso, 1984).

However, Mechsner et al. (2001) presented evidence that the bias is toward spatial symmetry and not motor symmetry. To dissociate motor symmetry from spatial symmetry, Mechsner et al. had subjects perform with the palm up for one hand and the palm down for the other (see Figures 6c,d). A tendency toward coactivation of homologous muscles would predict that, in this case, the bias should be toward parallel oscillation, whereas a tendency toward spatial symmetry should still show the bias toward symmetrical oscillation. The latter result
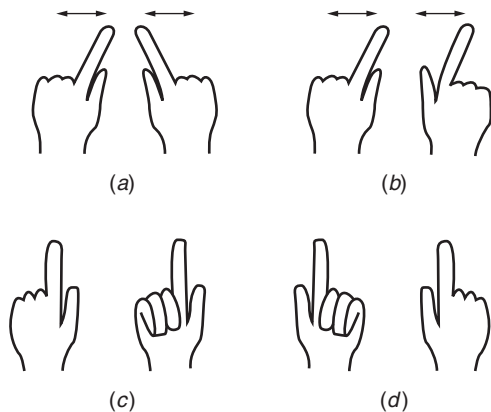
**Figure 6** (*a*) Symmetric and (*b*) asymmetric patterns of finger movements for which symmetric movements are typically easier. Parts (*c*) and (*d*) show conditions examined by Mechsner et al. (2001), in which spatial symmetry and muscle homology are dissociated.

was in fact obtained, with the bias toward symmetrical oscillation being just as strong when one palm faced up and the other down as when both hands were palm down or both palm up. Mechsner et al. and Mechsner and Knoblich (2004) obtained similar results for tasks in which two fingers of each hand are periodically tapped together by comparing congruous conditions for which the fingers from the two hands were the same (e.g., index and middle fingers of each hand or middle and ring fingers of each hand) and incongruous conditions for which they were different (index and middle finger for one hand and middle and ring finger for the other). Mechsner and Knoblich concluded that "homology of active fingers, muscular portions, and thus motor commands plays virtually no role in defining preferred coordination patterns, in particular the symmetry tendency" (p. 502) and that the symmetry advantage "originates at a more abstract level, in connection with planning processes involving perceptual anticipation" (p. 502). Evidence from behavioral and neuroscientific investigations has supported this conclusion that a major source of constraint in bimanual control is a consequence of the manner in which the action goals are represented (Oliveira and Ivry, 2008).

### 3.4 Sequencing and Timing of Action

How sequences of actions are planned and executed is one of the central problems of concern in the area of motor control (Rosenbaum, 2010). Most discussions of this problem originate with Lashley's (1951) well-known book chapter in which he presented evidence against an associative chaining account of movement sequences, according to which the feedback from each movement in the sequence provides the stimulus for the next movement. Instead, Lashley argued that the sequences are controlled centrally by motor plans.

Considerable evidence is consistent with the idea that these motor plans are structured hierarchically.

For example, Povel and Collard (1982) had subjects perform sequences of six taps with the four fingers on a hand (excluding the thumb). A sequence was practiced until it could be performed from memory, and then trials were conducted for which the sequence was to be executed as rapidly as possible. The sequences differed in terms of the nature and extent of their structure. For example, the patterns 1−2−3−2−3−4 and 2−3−4−1−2−3, where the numbers 1, 2, 3, and 4 designate the index, middle, ring, and little fingers, respectively, can each be coded as two separate ordered subsets. Povel and Collard found that the pattern of latencies between each successive tap was predicted well by a model that assumed the memory representation for the sequence was coded in a hierarchical decision tree (see Figure 7), with the movement elements represented at the lowest level, which was then interpreted by a decoding process that traversed the decision tree from left to right. Interresponse latencies were predicted well by the number of links that had to be traversed in the tree between successive responses. For example, for the sequences shown above, the longest latencies were between the start signal and the first tap and between the third and fourth taps, both of which required two levels of the tree to be traversed.

Although many results in tasks requiring execution of sequential actions are in agreement with predictions of hierarchical models, it should be noted that it is not so simple to rule out serial association models. Context-sensitive association models, which allow elements farther back than just the immediately preceding one to affect performance, can generate many of the same result patterns as hierarchical models (e.g., Wickelgren, 1969).

Beginning with a study by Nissen and Bullemer (1987), numerous experiments have been conducted on incidental learning of trial sequences in choice–reaction tasks. Nissen and Bullemer had subjects perform a four-choice RT task in which the stimulus on a trial appeared in one of four horizontal locations, and the response was the corresponding location of one of
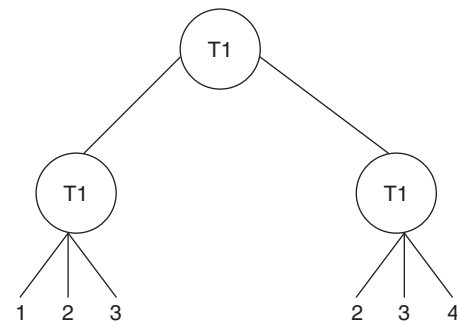


**Figure 7** Hierarchical representation of the movement sequence 1−2−3−2−3−4. T1 represents the operation *transpose to an adjacent finger*. The tree traversal model predicts longer latencies for the first and fourth elements in the movement sequence, as Povel and Collard (1982) found.

four buttons also arranged in a row, made with the middle and index fingers of the left and right hands. Subjects received eight blocks of 100 trials for which the stimuli were presented in random order or in a sequence that repeated every 10 trials. There was a slight decrease in RT of about 20 ms across blocks with the random order but a much larger one of about 150 ms for the repeating sequence. Nissen and Bullemer presented evidence that they interpreted as indicating that such sequence learning can occur without awareness, but this remains a contentious issue (see, e.g., Fu et al, 2010; Rünger and Frensch, 2010).

Of most interest to present concerns is the nature of the representation that is being learned in sequential tasks. Most studies have found little evidence for perceptual learning of the stimulus sequence (e.g., Willingham et al., 2000), although Gheysen et al. (2009) did find perceptual learning for a task that required attending to the stimulus sequence and maintaining information in working memory. Their study also showed evidence of there being a nonperceptual component to the sequence learning, as have most other studies. In general, results have indicated that this learning is not effector specific, because it can transfer to a different set of effectors (e.g., Cohen et al., 1990). Willingham et al. (2000) concluded that the sequence learning occurs in a part of the motor system involving response locations but not specific effectors or muscle groups. They showed that subjects who practiced the task using a keyboard with one arrangement of response keys during a training phase showed no benefit from the repeating stimulus sequence when subsequently transferred to a keyboard with a different arrangement of response keys. In another experiment, Willingham et al. also showed that subjects who switched from performing the task with the hands crossed in practice to performing with them uncrossed in a transfer session, such that the hand operating each key was switched, showed no cost relative to subjects who used the uncrossed hand placement throughout. Willingham et al. rejected an explanation in terms of SR associations because Willingham (1999) found excellent transfer as long as the response sequence remained the same in both the practice and transfer sessions even when the stimulus set was changed from digits to spatial locations or the mapping of spatial stimuli to responses was changed. Note that Willingham et al.'s conclusions are similar to those reached by Mechsner and Knoblich (2004) for bimanual coordination in that much of the motor control and learning occur at a level of spatial response relations rather than the muscles used to execute the actions.

Whereas in some situations the speed with which a sequence of actions is executed is important, in others the timing of the actions is crucial. One influential model of response timing is that of Wing and Kristofferson (1973), who developed it to explain the timing of successive, discrete tapping responses. According to this model, two processes control the timing of the responses. One is an internal clock that generates trigger pulses that can be used to time the delay (by the number of pulses) and initiate motor responses. The other is a delay process between when a trigger pulse initiates a response and when the movement is actually executed. The interval between successive pulses is assumed to be an independent random variable, as is the interval between a trigger pulse and the response that it initiates. One key prediction of the model is that the variance of the interval between responses should increase as the delay between the responses increases, due to the variability of the internal clock. Another prediction is that adjacent interresponse intervals should be negatively correlated, due to the variability of the delay process. These predictions have tended to be confirmed, and Wing and Kristofferson's model has generally been successful for timing of discrete responses such as tapping. However, evidence has suggested that, for more continuous motor acts such as drawing circles at a certain rate with the dominant hand, the timing is an emergent property of the movement rather than the consequence of a central timer (e.g., Ivry et al., 2002; Studenka and Zelaznik, 2008).

## 3.5 Motor Learning and Acquisition of Skill

Performance of virtually any perceptual–motor skill improves substantially with practice, as with the sequence learning described in the previous section, becoming faster, more accurate, and more fluid. Considerable research has been devoted to understanding the ways in which training and augmented feedback (knowledge of results or performance) should be scheduled to optimize the acquisition of motor skill. Some of this research is described in the following sections.

### 3.5.1 Practice Schedules

A long-standing issue in the study of motor skill has been whether better learning results from a distributed practice schedule, in which there is a break or rest period between performance attempts, or a massed practice schedule, for which there is not. Although distributed practice often leads to better performance during acquisition of a motor skill, it does not necessarily result in better learning and retention. Lee and Genovese (1988) conducted a meta-analysis of the literature on the distribution of practice and concluded that distributed practice does result in more learning for motor tasks that require continuous movements, such as cycling. Donovan and Radosevich (1999) reached a similar conclusion from a larger scale meta-analysis, with their findings indicating that distributed practice is more beneficial for simple tasks than for complex ones.

For discrete tasks, however, massed practice may even be beneficial to learning. Carron (1969) found better retention for massed than distributed practice when the task required the three discrete steps of picking up a dowel out of a hole, flipping the ends of the dowel, and putting the dowel back into the hole. Lee and Genovese (1989) directly compared discrete and continuous versions of a task in which the interval between when a stylus was lifted from one plate and moved to another was to be 500 ms. A single movement was made for the discrete version of the task, whereas 20 movements back and forth were performed in succession for the continuous version. Massed practice produced better retention than distributed practice did for the

discrete task, but distributed practice produced better retention for the continuous task.

The conclusions of Lee and Genovese (1988, 1989) and Donovan and Radosevich (1999) hold for distribution of practice within a session. For performance across practice sessions, evidence suggests that shorter practice sessions spread over more days are more effective than longer sessions spread over fewer days. For example, Baddeley and Longman (1978) gave postal trainees 60 h of practice learning to operate mail-sorting machines. The trainees who received this practice in 1-h/day sessions over 12 weeks learned the task much better than those who received the practice in 2-h sessions twice daily over three weeks. One factor contributing to the smaller benefit in learning with the longer, more massed sessions is that the sessions may get tiresome, causing people's attention to wander.

Another issue is whether different tasks or task variations should be practiced individually, in distinct practice blocks, or mixed together within a practice block. Retention and transfer of motor tasks have been shown typically to be better when the tasks are practiced in random order than in distinct blocks, even though performance during the practice session is typically better under blocked conditions. This finding, called the *contextual interference effect*, was first demonstrated by Shea and Morgan (1979). They had subjects knock down three of six barriers in a specified order as quickly as possible when a stimulus light occurred. During the acquisition phase, each subject performed three different versions of the task, which differed with respect to the barriers that were to be knocked down and their order. For half of the subjects, the three barrier conditions were practiced in distinct trial blocks, whereas for the other half, the barrier conditions were practiced in a random order. Although performance during acquisition was consistently faster for the blocked group than for the random group, performance on retention tests conducted 10 min or 10 days later was faster for the random group.

The contextual interference effect has been replicated in numerous studies and tasks (see, e.g., Magill and Hall, 1990; Wright et al., 2005). Shea and Morgan (1979) originally explained the contextual interference effect as follows: Because performance during practice is more difficult for the random group than for the blocked group, the random group is forced to use multiple processing strategies, leading to more elaborate long-term memory representations and better retention. Lee and Magill (1985) proposed instead that the benefit of random practice arises from subjects often forgetting how the task to be performed on the current trial was done previously, requiring that an action plan be reconstructed. This reconstruction process results in a more highly developed memory trace. Although these accounts differ slightly in their details, they make the similar point that random practice schedules lead to better long-term retention because they require deeper or more elaborate processing of the movements. Evidence suggests that this processing is reflected in greater activation of the motor cortex (Lin et al., 2009).

Because real-world perceptual–motor skills may be quite complex, another issue that arises is whether it is beneficial for learning to practice parts of a task in isolation before performing the whole, integrated task. Three types of part-task practice can be distinguished (Wightman and Lintern, 1985): *Segmentation* involves decomposing a task into successive subtasks, which are performed in isolation or in groups and then recombined; *fractionation* involves separate performance of subtasks that typically are performed simultaneously and then recombining them; *simplification* involves practicing a reduced version of a task that is easier to perform (such as using training wheels on a bicycle) before performing the complete version. Part-task training is often beneficial, and the results can be striking, as illustrated by Frederiksen and White's (1989) study, involving performance of a video game called Space Fortress that entailed learning and coordinating many perceptual–motor and cognitive components. In that study, subjects who received part-task training on the key task components performed about 25% better over the last five of eight whole-game transfer blocks than did subjects who received whole-game practice (see Figure 8), and this difference showed no sign of diminishing.

Although part-task training is beneficial for complex tasks that require learning complex rules and relations and coordinating the components, it is less beneficial for motor skills composed of several elements, such as a tennis serve, where practicing one element in isolation shows at most small transfer to the complete task (e.g., Lersten, 1968). Evidence suggests, though, that part-task practice with the first half of a movement sequence prior to practice of the whole sequence causes participants to code the two halves as separate parts, resulting in better performance than that of a whole-task group when required to perform just the second half in a transfer test (Park et al., 2004).

### 3.5.2 Provision of Feedback

Intrinsic feedback arises from movement, and this sensory information is a natural consequence of action. For example, as described previously, several types of visual and proprioceptive feedback are typically associated with moving a limb from a beginning location to a target location. Of more concern for motor learning, though, is extrinsic, or augmented, feedback, which is information that is not inherent to performing a task itself. Two types of extrinsic feedback are typically distinguished, *knowledge of results*, which is information about the outcome of the action, and *knowledge of performance*, which is feedback concerning how the action was executed.

Knowledge of results (KR) is particularly important for motor learning when the intrinsic feedback for the task itself does not provide an indication of whether the goal was achieved. For example, in learning to throw darts at a target, the extrinsic KR is not of extreme importance because intrinsic visual feedback provides information about the amount of error in the throws. However, even in this case, KR may provide motivation to the performer and reinforcement of their actions, and knowledge of performance (e.g., whether the throwing motion was appropriate) may also be beneficial. If the task is one of learning to throw darts in the dark, KR
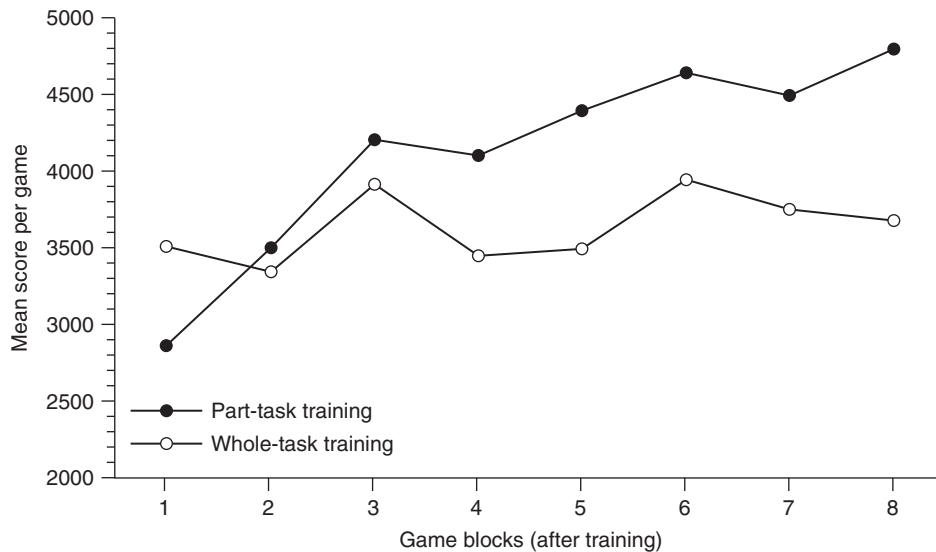
**Figure 8** Mean score of Frederiksen and White's (1989) subjects on eight game blocks of Space Fortress in the transfer session after receiving part- or whole-task training in a prior session.

increases in importance because there is no longer visual feedback to provide information about the accuracy of the throws. Many issues concerning KR have been investigated, including the precision of the information conveyed and the schedule by which it is conveyed.

Feedback can be given with varying precision. For example, when performing a task that requires contact with a target at a specified time window, say, 490–500 ms after movement initiation, the person may be told whether or not the movement was completed within the time window (qualitative KR) or how many milliseconds shorter or longer the movement was than allowed by the window (quantitative KR). Qualitative feedback can be effective, particularly at early stages of practice when the errors are often large, but people tend to learn better when KR is quantitative than when it is just qualitative (Magill and Wood, 1986; Reeve et al., 1990).

Although it may seem that it is best to provide feedback on every trial, research has indicated to the contrary. For example, Winstein and Schmidt (1990) had people learn to produce a lever movement pattern consisting of four segments in 800 ms. Some subjects received KR after every trial during acquisition, whereas others received KR on only half of the trials. The two groups performed similarly during acquisition, but those subjects who received feedback after every trial did substantially worse than the other group on a delayed retention test. Similar results have been obtained for a more naturalistic golf-putting task (Ishikura, 2008). Summary KR, for which feedback about a subset of trials is not presented until the subset is completed, has also been found to be successful (Lavery, 1962; Schmidt et al., 1989). Schmidt et al. had people learn a timed lever movement task similar to that used by Winstein and Schmidt (1990), providing summary KR after 1, 5, 10,

or 15 trials. A delayed retention test showed that learning was best when summary KR was provided every 15 trials and worst when KR was provided every trial. The apparent reason why it is best not to provide feedback on every performance attempt is that the person comes to depend on it. Thus, much like blocked practice of the same task, providing feedback on every trial does not force the person to engage in the more effortful information processing that is necessary to produce enduring memory traces needed for long-term performance.

## 4 SUMMARY AND CONCLUSIONS

Human–machine interactions involve a succession of reciprocal actions taken by the human and the machine. For performance of the human component to be optimal, it is necessary not only to consider how the machine should display information regarding its states and activities to the human, but also to take into account the processes by which the human selects and executes actions in the sequence of the interaction. Selection and control of action have been studied since the earliest days of research on human performance, and research in these areas continues to produce significant empirical and theoretical advances, several of which have been summarized in this chapter. Because the purpose of the chapter is to provide readers with an overview of the topic of selection and control of action, readers are encouraged to refer to more detailed information on topics of interest in chapters by Rosenbaum (2002), Heuer and Massen (in press), and Proctor and Vu (in press); and books by Rosenbaum (2010), Proctor and Dutta (1995), Sanders (1998), and Schmidt and Lee (1999); and other sources.

This chapter showed that the relations between choice uncertainty and response time, captured by the

Hick–Hyman law, movement difficulty and movement time, conveyed by Fitts's law, and amount of practice and performance time, depicted by the power law of practice, follow quantitative laws that can be applied to specific research and design issues in human factors and ergonomics. In addition, many qualitative principles are apparent from research that is directly applicable to human factors:

- The relative speed and accuracy of responding in a situation depends in part on the setting of response thresholds, or how much noisy evidence needs to be sampled before deciding which alternative action to select.
- Sequential sampling models can capture the relations between speed and accuracy of performance in various task conditions.
- Response time increases as the number of alternatives increases, but the cost of additional alternatives is reduced when compatibility is high or the performer is highly practiced.
- Spatially compatible relations and mappings typically yield better performance than spatially incompatible ones.
- Compatibility effects are not restricted to spatial relations but occur for stimulus and response sets that have perceptual or conceptual similarity of any type.
- Compatibility effects occur when an irrelevant dimension of the stimulus set shares similarity with the relevant response dimension.
- For many situations in which compatible mappings are mixed with less compatible ones, the benefit of compatibility is eliminated.
- When actions are not performed in isolation, the context of preceding events can affect performance significantly.
- Advance information can be used to prepare subsets of responses.
- Improvements in response selection efficiency with practice that occur in a variety of tasks involve primarily spatial locations of the actions and their relation to the stimuli, not the effectors used to accomplish the actions.
- Small amounts of experience with novel relations may influence performance after a long delay, even when those relations are no longer relevant to the task.
- Costs that are associated with mixing and switching tasks can be only partly overcome by advance preparation.
- It is difficult to select an action for more than one task at a time, although the costs in doing so can be reduced by using highly compatible tasks and with practice.
- Many constraints influence movement time, and the particular way in which an action will be carried out needs to be accommodated when designing for humans.

- Feedback of various types is important for motor control and acquisition of perceptual–motor skills.
- The tendency toward symmetry in preferred bimanual coordination patterns is primarily one of spatial symmetry, not of homologous muscles.
- Practice and feedback schedules that produce the best performance of perceptual–motor skills during the acquisition phase often do not promote learning and retention of the skills.
- Part-task training can be an effective means of teaching someone how to perform complex tasks.

Beyond these general laws and principles, research has yielded many details concerning the factors that are critical to performance in specific situations. Moreover, models of various types, some qualitative and some quantitative, have been developed for various domains of phenomena that provide relatively accurate descriptions and predictions of how performance will be affected by numerous variables. The laws, principles, and model characteristics can be incorporated into cognitive architectures such as EPIC (Meyer and Kieras, 1997) and ACT-R (Anderson et al., 2004; Byrne, 2001), along with other facts, to develop computational models that enable quantitative predictions to be derived for complex tasks of the type encountered in much of human factors and ergonomics.

## REFERENCES

Adam, J. J., Hommel, B., and Umiltà, C. (2003), "Preparing for Perception and Action, I: The Role of Grouping in the Response-Cuing Paradigm," *Cognitive Psychology*, Vol. 46, pp. 302–358.

Allport, A., Styles, E. A., and Hsieh, H. (1994), "Shifting Intentional Set: Exploring the Dynamic Control of Tasks," in *Attention and Performance XV*, C. Umiltà and M. Moscovitch, Eds., MIT Press, Cambridge, MA, pp. 421–452.

Anderson, J. R. (1982), "Acquisition of Cognitive Skill," *Psychological Review*, Vol. 89, pp. 369–406.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004), "An Integrated Theory of the Mind," *Psychological Review*, Vol. 111, pp. 1036–1060.

Ansorge, U. (2003), "Spatial Simon Effects and Compatibility Effects Induced by Observed Gaze Direction," *Visual Cognition*, Vol. 10, pp. 363–383.

Baddeley, A. D., and Longman, D. J. A. (1978), "The Influence of Length and Frequency of Training Session on the Rate of Learning to Type," *Ergonomics*, Vol. 21, pp. 627–635.

Baird, K. M., Hoffmann, E. R., and Drury, C. G. (2002), "The Effects of Probe Length on Fitts' Law," *Applied Ergonomics*, Vol. 33, pp. 9–14.

Band, G. P. H., Ridderinkhof, K. R., and van der Molen, M. W. (2003), "Speed–Accuracy Modulation in Case of Conflict: The Roles of Activation and Inhibition," *Psychological Research*, Vol. 67, pp. 266–279.

Bertelson, P. (1961), "Sequential Redundancy and Speed in a Serial Two-Choice Responding Task," *Quarterly Journal of Experimental Psychology*, Vol. 13, pp. 90–102.

Bertelson, P. (1967), "The Time Course of Preparation," *Quarterly Journal of Experimental Psychology*, Vol. 19, pp. 272–279.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Thousand Oaks, CA: Sage.

Byrne, M. D. (2001), "ACT-R/PM and Menu Selection: Applying a Cognitive Architecture to HCI," *International Journal of Human–Computer Studies*, Vol. 55, pp. 41–84.

Card, S. K., English, W. K., and Burr, B. J. (1978), "Evaluation of the Mouse, Rate-Controlled Isometric Joystick, Step Keys, and Text Keys for Text Selection on a CRT," *Ergonomics*, Vol. 21, pp. 601–613.

Carron, A. (1969), "Performance and Learning in a Discrete Motor Task under Massed and Distributed Practice," *Research Quarterly*, Vol. 40, pp. 481–489.

Chen, S., and Melara, R. D. (2009), "Sequential Effects in the Simon Task: Conflict Adaptation or Feature Iintegration?" *Brain Research*, Vol. 1297, pp. 89–100.

Christie, J. J., and Klein, R. M. (2008), "On Finding Negative Priming from Distractors," *Psychonomic Bulletin and Review*, Vol. 15, pp. 866–873.

Cohen, A., Ivry, R. I., and Keele, S. W. (1990), "Attention and Structure in Sequence Learning," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 16, pp. 17–30.

De Jong, R., Liang, C.-C., and Lauber, E. (1994), "Conditional and Unconditional Automaticity: A Dual-Process Model of Effects of Spatial Stimulus–Response Correspondence," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 20, pp. 731–750.

Delaney, P. F., Reder, L. M., Staszewski, J. J., and Ritter, F. E. (1998), "The Strategy-Specific Nature of Improvement: The Power Law Applies by Strategy within Task," *Psychological Science*, Vol. 9, pp. 1–7.

Donovan, J. J., and Radosevich, D. J. (1999), "A Meta-Analytic Review of the Distribution of Practice Effect:Now You See It, Now You Don't," *Journal ofApplied Psychology*, Vol. 84, 795–805.

Dornier, L. A., and Reeve, T. G. (1992), "Estimation of Compatibility Effects in Precuing of Arm and Direction Parameters," *Research Quarterly for Exercise and Sport*, Vol. 61, pp. 37–49.

Duscherer, K., Holender, D., and Molenaar, E. (2008), "Revisiting the Affective Simon Effect," *Cognition and Emotion*, Vol. 22, pp. 193–217.

Dutta, A., and Proctor, R. W. (1992), "Persistence of Stimulus–Response Compatibility Effects with Extended Practice," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 18, pp. 801–809.

Fitts, P. M. (1954), "The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement," *Journal of Experimental Psychology*, Vol. 47, pp. 381–391.

Fitts, P. M., and Deininger, R. L. (1954), "S–R Compatibility: Correspondence among Paired Elements within Stimulus and Response Codes," *Journal of Experimental Psychology*, Vol. 48, pp. 483–492.

Fitts, P. M., and Seeger, C. M. (1953), "S–R Compatibility: Spatial Characteristics of Stimulus and Response Codes," *Journal of Experimental Psychology*, Vol. 46, pp. 199–210.

Fox, E. (1995), "Negative Priming from Ignored Distractors in Visual Selection: A Review," *Psychonomic Bulletin and Review*, Vol. 2, pp. 145–173.

Frederiksen, J. R., and White, B. Y. (1989), "An Approach to Training Based upon Principled Task Decomposition," *Acta Psychologica*, Vol. 71, pp. 89–146.

Fu, Q., Dienes, Z., and Fu, X. (2010), "Can Unconscious Knowledge Allow Control in Sequence Learning?" *Consciousness and Cognition: An International Journal*, Vol. 19, pp. 462–474.

Galashan, D., Wittfoth, M., Fehr, T., and Herrmann, M. (2008), "Two Simon Tasks with Different Sources of Conflict: An ERP Study of Motion and Location-Based Compatibility Effects," *Biological Psychology*, Vol. 78, pp. 246–252.

Gheysen, F., Gevers, W., De Schutter, E., Van Waelvelde, H., and Fias, W. (2009), "Disentangling Perceptual from Motor Implicit Sequence Learning with a Serial Color-Matching Task," *Experimental Brain Research*, Vol. 197, pp. 163–174.

Goodman, D., and Kelso, J. (1980), "Are Movements Prepared in Parts Not Under Compatible (Naturalized) Conditions?" *Journal of Experimental Psychology: General*, Vol. 109, pp. 475–495.

Greenwald, A. G., and Shulman, H. G. (1973), "On Doing Two Things at Once, II: Elimination of the Psychological Refractory Period Effect," *Journal of Experimental Psychology*, Vol. 101, pp. 70–76.

Guiard, Y. (1983), "The Lateral Coding of Rotation: A Study of the Simon Effect with Wheel Rotation Responses," *Journal of Motor Behavior*, Vol. 15, pp. 331–342.

Haider, H., and Frensch, P. A. (2002), "Why Aggregated Learning Follows the Power Law of Practice When Individual Learning Does Not: Comment on Rickard (1997, 1999), Delaney et al. (1998), and Palmeri (1999)," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 28, pp. 392–406.

Heathcote, A., Brown, S., and Mewhort, D. J. K. (2000), "The Power Law Repealed: The Case for an Exponential Law of Practice," *Psychonomic Bulletin and Review*, Vol. 7, pp. 185–207.

Henry, F. M., and Rogers, D. E. (1960), "Increased Response Latency for Complicated Movements and a 'Memory Drum' Theory for Neuromotor Reaction," *Research Quarterly*, Vol. 31, pp. 448–458.

Heuer, H., and Hegele, M., (2008). "Constraints on Visuo-Motor Adaptation Depend on the Typeof Visual Feedback During Practice," *Experimental Brain Research*, Vol. 185, pp. 101–110.

Heuer, H., and Massen, C. (in press), "Motor Control," in *Experimental Psychology*, A. F. Healy and R. W. Proctor, Eds., Vol. 4 in *Handbook of Psychology*, 2nd ed., I. B. Weiner, Ed.-in-Chief, Wiley, Hoboken, NJ.

Hick, W. E. (1952), "On the Rate of Gain of Information," *Quarterly Journal of Experimental Psychology*, Vol. 4, pp. 11–26.

Hoffmann, E. R. (1997), "Strength of Component Principles Determining Direction of Turn Stereotypes-Linear Displays and Rotary Controls," *Ergonomics*, Vol. 40, pp. 199–222.

Hommel, B. (1993a), "Inverting the Simon Effect by Intention," *Psychological Research*, Vol. 55, pp. 270–279.

Hommel, B. (1993b), "The Relationship between Stimulus Processing and Response Selection in the Simon Task: Evidence for a Temporal Overlap," *Psychological Research*, Vol. 55, pp. 280–290.

Hommel, B. (1998a), "Automatic Stimulus–Response Translation in Dual-Task Performance," *Journal of Experimental*

*Psychology: Human Perception and Performance*, Vol. 24, pp. 1368–1384.

Hommel, B. (1998b), "Event Files: Evidence for Automatic Integration of Stimulus–Response Episodes," *Visual Cognition*, Vol. 5, pp. 183–216.

Hommel, B., Proctor, R. W., and Vu, K.-P. L. (2004), "A Feature-Integration Account of Sequential Effects in the Simon Task," *Psychological Research*, Vol. 68, pp. 1–17.

Hyman, R. (1953), "Stimulus Information as a Determinant of Reaction Time," *Journal of Experimental Psychology*, Vol. 45, pp. 188–196.

Iani, C., Rubichi, S., Gherri, E., and Nicoletti, R. (2009), "Co-occurrence of Sequential and Practice Effects in the Simon Task: Evidence for Two Independent Mechanisms Affecting Response Selection. *Memory and Cognition*, Vol. 37, pp. 358–367.

Inhoff, A. W., Rosenbaum, D. A., Gordon, A. M., and Campbell, J. A. (1984), "Stimulus–Response Compatibility and Motor Programming of Manual Response Sequences," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 10, pp. 724–733.

Ishikura, T. (2008), "Reduced Relative Frequency of Knowledge of Results without Visual Feedback in Learning a Golf-Putting Task," *Perceptual and Motor Skills*, Vol. 106, pp. 225–233.

Ivry, R., Spencer, R. M., Zelaznik, H. N., and Diedrichsen, J. (2002), "The Cerebellum and Event Timing," in *The Cerebellum: Recent Developments in Cerebellar Research*, Vol. 978, S. M. Highstein and W. T. Thach, Eds., New York Academy of Sciences, New York, pp. 302–317.

Keele, S. W., and Posner, M. I. (1968), "Processing of Visual Feedback in Rapid Movements," *Journal of Experimental Psychology*, Vol. 77, pp. 155–158.

Kelso, J. A. S. (1984), "Phase Transitions and Critical Behavior in Human Bimanual Coordination," *American Journal of Physiology: Regulatory, Integrative and Comparative Physiology*, Vol. 15, pp. R1000–R1004.

Kerr, R. (1973), "Movement Time in an Underwater Environment," *Journal of Motor Behavior*, Vol. 5, pp. 175–178.

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., and Koch, I. (2010), "Control and Interference in Task Switching—A Review. *Psychological Bulletin*," Vol. 136, pp. 849–874.

Kornblum, S., Hasbroucq, T., and Osman, A. (1990), "Dimensional Overlap: Cognitive Basis for Stimulus–Response Compatibility—A Model and Taxonomy," *Psychological Review*, Vol. 97, pp. 253–270.

Kunde, W., and Stöcker, C. (2002), "A Simon Effect for Stimulus–Response Duration," *Quarterly Journal of Experimental Psychology*, Vol. 55A, pp. 581–592.

Lamberts, K., Tavernier, G., and d'Ydewalle, G. (1992), "Effects of Multiple Reference Points in Spatial Stimulus–Response Compatibility," *Acta Psychologica*, Vol. 79, pp. 115–130.

Langolf, G., and Hancock, W. M. (1975), "Human Performance Times in Microscope Work," *AIEE Transactions*, Vol. 7, pp. 110–117.

Lashley, K. S. (1951), "The Problem of Serial Order Behavior," in *Cerebral Mechanisms in Behavior*, L. A. Jeffress, Ed., Wiley, New York, pp. 112–131.

Lavery, J. J. (1962), "Retention of Simple Motor Skills as a Function of Type of Knowledge of Results," *Canadian Journal of Psychology*, Vol. 16, pp. 300–311.

Learmount, D., and Norris, G. (1990), "Lessons to be Learned," *Flight International*, October 31–November 6, pp. 24–26.

Lee, T. D., and Genovese, E. D. (1988), "Distribution of Practice in Motor Skill Acquisition: Learning and Performance Effects Reconsidered," *Research Quarterly for Exercise and Sport*, Vol. 59, pp. 277–287.

Lee, T. D., and Genovese, E. D. (1989), "Distribution of Practice in Motor Skill Acquisition: Different Effects for Discrete and Continuous Tasks," *Research Quarterly for Exercise and Sport*, Vol. 70, pp. 59–65.

Lee, T. D., and Magill, R. A. (1985), "Can Forgetting Facilitate Skill Acquisition?" in *Differing Perspectives on Motor Memory, Learning, and Control*, D. Goodman, R. B. Wilberg, and I. M. Franks, Eds., North-Holland, Amsterdam, pp. 3–22.

Leonard, J. A. (1958), "Partial Advance Information in a Choice Reaction Task," *British Journal of Psychology*, Vol. 49, pp. 89–96.

Lersten, K. C. (1968), "Transfer of Movement Components in a Motor Learning Task," *Research Quarterly*, Vol. 39, pp. 575–581.

Levy, J., Pashler, H., and Boer, E. (2006), "Central Interference in Driving: Is There Any Stopping the Psychological Refractory Period?" *Psychological Science*, Vol. 17, pp. 228–235.

Lien, M.-C., and Proctor, R. W. (2002), "Stimulus–Response Compatibility and Psychological Refractory Period Effects: Implications for Response Selection," *Psychonomic Bulletin and Review*, Vol. 9, pp. 212–238.

Lien, M.-C., Proctor, R. W., and Allen, P. A. (2002), "Ideomotor Compatibility in the Psychological Refractory Period Effect: 29 Years of Oversimplification," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 28, pp. 396–409.

Lien, M.-C., McCann, R. S., Ruthruff, E., and Proctor, R. W. (2005), "Dual-task Performance with Ideomotor Compatible Tasks: Is the Central Processing Bottleneck Intact, Bypassed, or Shifted in Locus?" *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 31, pp. 122–144.

Lin, C.-H., Fisher, B. E., Wu, A. D., Ko, Y.-A., Lee, L.-Y., and Winstein, C. J. (2009), "Neural Correlate of the Contextual Interference Effect in MotorLearning: A Kinematic Analysis," *Journal of Motor Behavior*, Vol. 41, pp. 232–242.

Logan, G. D. (1988), "Toward an Instance Theory of Automatization," *Psychological Review*, Vol. 95, pp. 492–527.

Logan, G. D. (2003), "Simon-Type Effects: Chronometric Evidence for Keypress Schemata in Typewriting," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 29, pp. 741–757.

Magill, R. A., and Hall, K. G. (1990), "A Review of Contextual Interference in Motor Skill Acquisition," *Human Movement Science*, Vol. 9, pp. 241–289.

Magill, R. A., and Wood, C. A. (1986), "Knowledge of Results Precision as a Learning Variable in Motor Skill Acquisition," *Research Quarterly for Exercise and Sport*, Vol. 57, pp. 170–173.

Marino, R. A., and Munoz, D. P. (2009), "The Effects of Bottom-Up Target Luminance and Top-Down Spatial Target Predictability on Saccadic Reaction Times," *Experimental Brain Research*, Vol. 197, pp. 321–335.

Masaki, H., Wild-Wall, N., Sangals, J., and Sommer, W. (2004), "The Functional Locus of the Lateralized Readiness Potential," *Psychophysiology*, Vol. 41, pp. 220–230.

McCann, R. S., and Johnston, J. C. (1992), "Locus of the Single-Channel Bottleneck in Dual-Task Interference," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 18, pp. 471–484.

Mechsner, F., and Knoblich, G. (2004), "Do Muscles Matter for Coordinated Action?" *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 30, pp. 490–503.

Mechsner, F., Kerzel, D., Knoblich, G., and Prinz, W. (2001, November 1), "Perceptual Basis of Bimanual Coordination," *Nature*, Vol. 414, pp. 69–73.

Merkel, J. (1885), "Die zeitliche Verhältnisse de Willenstatigkeit" [The Temporal Relations of Activities of the Will], *Philosophische Studien*, Vol. 2, pp. 73–127.

Meyer, D. E., and Kieras, D. E. (1997), "A Computational Theory of Executive Cognitive Processes and Multiple-Task Performance, 1: Basic Mechanisms," *Psychological Review*, Vol. 104, pp. 3–66.

Meyer, D. E., Abrams, R. A., Kornblum, S., Wright, C. E., and Smith, J. E. K. (1988), "Optimality in Human Motor Performance: Ideal Control of Rapid Aimed Movements," *Psychological Review*, Vol. 95, pp. 340–370.

Miles, J. D., and Proctor, R. W. (2010), "Attention Is Required for the Acquisition But Not Expression of New Spatial Associations," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 36, 1554–1560.

Miller, J. (1982), "Discrete Versus Continuous Stage Models of Human Information Processing: In Search of Partial Output," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 8, pp. 273–296.

Monsell, S. (1986), "Programming of Complex Sequences: Evidence from the Timing of Rapid Speech and Other Productions," in *Generation and Modulation of Action Patterns*, H. Heuer and C. Fromm, Eds., Springer-Verlag, Berlin, pp. 72–86.

Monsell, S. (2003), "Task Switching," *Trends in Cognitive Sciences*, Vol. 7, pp. 134–140.

Moors, A., and De Houwer, J., (2006). "Automaticity: A Theoretical and Conceptual Analysis," *Psychological Bulletin*, Vol. 132, pp. 297–326.

Nattkemper, D., and Ziessler, M. (2004), "Cognitive Control of Action: The Role of Action Effects," *Psychological Research*, Vol. 68, pp. 71–73.

Navon, D., and Miller, J. (2002), "Queuing or Sharing? A Critical Evaluation of the Single-Bottleneck Notion," *Cognitive Psychology*, Vol. 44, pp. 193–251.

Neill, W. T., and Valdes, L. A. (1992), "Persistence of Negative Priming: Steady State or Decay?" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 18, pp. 565–576.

Newell, A., and Rosenbloom, P. S. (1981), "Mechanisms of Skill Acquisition and the Law of Practice," in *Cognitive Skills and Their Acquisition*, J. R. Anderson, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1–55.

Nissen, M. J., and Bullemer, P. (1987), "Attentional Requirements of Learning: Evidence from Performance Measures," *Cognitive Psychology*, Vol. 19, pp. 1–32.

Notebaert, W., Soetens, E., and Melis, A. (2001), "Sequential Analysis of a Simon Task: Evidence for an Attention-Shift Account," *Psychological Research*, Vol. 65, pp. 170–184.

Oliveira, F. T. P., and Ivry, R. B. (2008), "The Representation of Action Insights from Bimanual Coordination," *Current Directions in Psychological Science*, Vol. 17, pp. 130–135.

Pachella, R. G. (1974), "The Interpretation of Reaction Time in Information-Processing Research," in *Human Information Processing: Tutorials in Performance and Cognition*, B. H. Kantowitz, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 41–82.

Park, J., and Kanwisher, N. (1994), "Negative Priming for Spatial Locations: Identity Mismatching, Not Distractor Inhibition," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 20, pp. 613–623.

Park, J.-H., Wilde, H., and Shea, C. H. (2004), "Part-Whole Practice of Movement Sequences," *Journal of Motor Behavior*, Vol. 36, pp. 51–61.

Pashler, H., and Johnston, J. C. (1989), "Chronometric Evidence for Central Postponement in Temporally Overlapping Tasks," *Quarterly Journal of Experimental Psychology*, Vol. 41A, pp. 19–45.

Pashler, H., and Johnston, J. C. (1998), "Attentional Limitations in Dual-Task Performance," in *Attention*, H. Pashler, Ed., Psychology Press, Hove, East Sussex, England, pp. 155–189.

Pashler, H., Harris, C. R., and Nuechterlein, K. H. (2008), "Does the Central Bottleneck Encompass Voluntary Selection of Hedonically Based Choices?" *Experimental Psychology*, Vol. 55, pp. 313–321.

Pew, R. W. (1994), "Paul Morris Fitts," in *Division 21 Members Who Made Distinguished Contributions to Engineering Psychology*, H. L. Taylor, Ed., Division 21 of the American Psychological Association, Washington, DC, pp. 23–44.

Posner, M. I., Klein, R., Summers, J., and Buggie, S. (1973), "On the Selection of Signals," *Memory and Cognition*, Vol. 1, pp. 2–12.

Povel, D.-J., and Collard, R. (1982), "Structural Factors in Patterned Finger Tapping," *Acta Psychologica*, Vol. 52, pp. 107–124.

Proctor, R. W. (2011), "Playing the Simon Game: Use of the Simon Task for Investigating Human Information Processing," *Acta Psychologica*, *136*, 182–188.

Proctor, R. W., & Cho, Y. S. (2006), "Polarity Correspondence: A General Principle for Performance of Speeded Binary Classification Tasks," *Psychological Bulletin*, Vol. 132, pp. 416–442.

Proctor, R. W., and Dutta, A. (1993), "Do the Same Stimulus–Response Relations Influence Choice Reactions Initially and After Practice?" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 19, pp. 922–930.

Proctor, R. W., and Dutta, A. (1995), *Skill Acquisition and Human Performance*, Sage, Thousand Oaks, CA.

Proctor, R. W., and Lu, C.-H. (1999), "Processing Irrelevant Location Information: Practice and Transfer Effects in Choice-Reaction Tasks," *Memory and Cognition*, Vol. 27, pp. 63–77.

Proctor, R. W., and Reeve, T. G. (1986), "Salient-Feature Coding Operations in Spatial Precuing Tasks," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 12, pp. 277–285.

Proctor, R. W., and Vu, K.-P. L. (2006), *Stimulus-Response Compatibility Principles: Data, Theory, and Application*, CRC Press, Boca Raton, FL.

Proctor, R. W., and Vu, K.-P. L. (2010a), "Cumulative Knowledge and Progress in Human Factors," *Annual Review of Psychology*, Vol. 61, pp. 623–651.

Proctor, R. W., and Vu, K.-P. L. (2010b), "Universal and Culture-Specific Effects of Display-Control Compatibility," *American Journal of Psychology*, Vol. 123, 425–435.

Proctor, R. W., and Vu, K.-P. L. (in press), "Action Selection," in *Experimental Psychology*, A. F. Healy and R. W. Proctor, Eds., Vol. 4 in *Handbook of Psychology*, 2nd ed., I. B. Weiner, Ed.-in-Chief, Wiley, Hoboken, NJ.

Proctor, R. W., Wang, H., and Vu, K.-P. L. (2002), "Influences of Conceptual, Physical, and Structural Similarity on Stimulus-Response Compatibility," *Quarterly Journal of Experimental Psychology*, Vol. 55A, pp. 59–74.

Proteau, L., and Cournoyer, J. (1992), "Vision of the Stylus in a Manual Aiming Task," *Quarterly Journal of Experimental Psychology*, Vol. 44A, pp. 811–828.

Read, L. E., and Proctor, R. W. (2009), "Influence of Mapping Complexity on Negative Priming for Incompatible Spatial Mappings," *Psychonomic Bulletin and Review*, Vol. 16, pp. 1118–1123.

Reeve, T. G., and Proctor, R. W. (1984), "On the Advance Preparation of Discrete Finger Responses," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 10, pp. 541–553.

Reeve, T. G., Dornier, L., and Weeks, D. J. (1990), "Precision of Knowledge of Results: Consideration of the Accuracy Requirements Imposed by the Task," *Research Quarterly for Exercise and Sport*, Vol. 61, pp. 284–291.

Rickard, T. C. (2004), "Strategy Execution and Cognitive Skill Learning: An Item-Level Test of Candidate Models," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 30, pp. 65–82.

Rogers, R. D., and Monsell, S. (1995), "Cost of a Predictable Switch between Simple Cognitive Tasks," *Journal of Experimental Psychology: General*, Vol. 124, pp. 207–231.

Rosenbaum, D. A. (1983), "The Movement Precuing Technique: Assumptions, Applications, and Extensions," in *Memory and Control of Action*, R. A. Magill, Ed., North-Holland, Amsterdam, pp. 231–274.

Rosenbaum, D. A. (2002), "Motor Control," in *Stevens' Handbook of Experimental Psychology*, 3rd ed., H. Pashler and S. Yantis, Eds., Vol. 1, *Sensation and Perception*, Wiley, New York, pp. 315–339.

Rosenbaum, D. A. (2010), *Human Motor Control*, 2nd ed., Academic, San Diego, CA.

Rosenbaum, D. A., Gordon, A. M., Stillings, N. A., and Feinstein, M. H. (1987), "Stimulus–Response Compatibility in the Programming of Speech," *Memory and Cognition*, Vol. 15, pp. 217–224.

Rosenbaum, D. A., Marchak, F., Barnes, H. J., Vaughan, J., Slotta, J. D., and Jorgensen, M. J. (1990), "Constraints for Action Selection: Overhand Versus Underhand Grips," in *Attention and Performance XIII: Motor Representation and Control*, M. Jeannerod, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 321–342.

Rünger, D., and Frensch, P. A. (2010), "Defining Consciousness in the Context of Incidental Sequence Learning: Theoretical Considerations and Empirical Implications," *Psychological Research*, Vol. 74, pp. 121–137.

Ruthruff, E., Johnston, J. C., Van Selst, M., Whitsell, S., and Remington, R. (2003), "Vanishing Dual-Task Interference after Practice: Has the Bottleneck Been Eliminated or Is It Merely Latent?" *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 29, pp. 280–289.

Sanders, A. F. (1998), *Elements of Human Performance*, Lawrence Erlbaum Associates, Mahwah, NJ.

Schmidt, R. A., and Lee, T. D. (1999), *Motor Control and Learning: A Behavioral Emphasis*, 3rd ed., Human Kinetics, Champaign, IL.

Schmidt, R. A., Young, D. E., Swinnen, S., and Shapiro, D. C. (1989), "Summary Knowledge of Results for Skill Acquisition: Support for the Guidance Hypothesis," *Journal of Experimental Psychology: Learning, Memory and Cognition*, Vol. 15, pp. 352–359.

Schumacher, E. H., Seymour, T. L., Glass, J. M., Fencsik, D. E., Lauber, E. J., Kieras, D. E., and Meyer, D. E. (2001), "Virtually Perfect Time Sharing in Dual-Task Performance: Uncorking the Central Cognitive Bottleneck," *Psychological Science*, Vol. 12, pp. 101–108.

Seibel, R. (1963), "Discrimination Reaction Time for a 1,023-Alternative Task," *Journal of Experimental Psychology*, Vol. 66, pp. 215–226.

Shannon, C. E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656, available: http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html.

Shea, J. B., and Morgan, R. L. (1979), "Contextual Interference Effects on Acquisition, Retention, and Transfer of a Motor Skill," *Journal of Experimental Psychology: Human Learning and Memory*, Vol. 5, pp. 179–187.

Simon, J. R. (1990), "The Effects of an Irrelevant Directional Cue on Human Information Processing," in *Stimulus–Response Compatibility: An Integrated Perspective*, R. W. Proctor and T. G. Reeve, Eds., North-Holland, Amsterdam, pp. 31–86.

Soetens, E. (1998), "Localizing Sequential Effects in Serial Choice Reaction Time with the Information Reduction Procedure," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 24, pp. 547–568.

Sternberg, S. (1998), "Discovering Mental Processing Stages: The Method of Additive Factors," in *An Invitation to Cognitive Science: Methods, Models, and Conceptual Issues*, 2nd ed., Vol. 4, D. Scarborough and S. Sternberg, Eds., MIT Press, Cambridge, MA, pp. 703–863.

Studenka, B. E., and Zelaznik, H. N. (2008), "The Influence of Dominant versus Non-Dominant Hand on Event and Emergent Motor Timing," *Human Movement Science*, Vol. 27, pp. 29–52.

Stürmer, B., Leuthold, H., Soetens, E., Schroeter, H., and Sommer, W. (2002), "Control over Location-Based Response Activation in the Simon Task: Behavioral and Electrophysiological Evidence," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 28, pp. 1345–1363.

Tagliabue, M., Zorzi, M., Umiltà, C., and Bassignani, F. (2000), "The Role of LTM Links and STM Links in the Simon Effect," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 26, pp. 648–670.

Taub, E., and Berman, A. J. (1968), "Movement and Learning in the Absence of Sensory Feedback," in *The Neuropsychology of Spatially Oriented Behavior*, S. J. Freeman, Ed., Dorsey, Homewood, IL, pp. 173–192.

Teichner, W. H., and Krebs, M. J. (1974), "Laws of Visual-Choice Reaction Time," *Psychological Review*, Vol. 81, pp. 75–98.

ten Hoopen, G., Akerboom, S., and Raymakers, E. (1982), "Vibrotactual Choice Reaction Time, Tactile Receptor Systems, and Ideomotor Compatibility," *Acta Psychologica*, Vol. 50, pp. 143–157.

Tombu, M., and Jolicœur, P. (2003), "A Central Capacity Sharing Model of Dual-Task Performance," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 29, pp. 3–18.

Usher, M., Olami, Z., and McClelland, J. L. (2002), "Hick's Law in a Stochastic Race Model with Speed–Accuracy Tradeoff," *Journal of Mathematical Psychology*, Vol. 46, pp. 704–715.

Van Zandt, T., Colonius, H., and Proctor, R. W. (2000), "A Comparison of Two Response-Time Models Applied to Perceptual Matching," *Psychonomic Bulletin and Review*, Vol. 7, pp. 208–256.

Vu, K.-P. L. (2011), "Unintentional and Intentional Learning of Noncorresponding Stimulus-Response Associations in the Simon Task," *Acta Psychologica*, Vol. 136, pp. 217–224.

Vu, K.-P. L., and Proctor, R. W. (2008), "Age Differences in Response Selection for Pure and Mixed Stimulus-Response Mappings and Tasks," *Acta Psychologica*, Vol. 129, pp. 49–60.

Vu, K.-P. L., Proctor, R. W., and Urcuioli, P. (2003), "Transfer Effects of Incompatible Location-Relevant Mappings on Subsequent Visual or Auditory Simon Tasks," *Memory and Cognition*, Vol. 31, pp. 1146–1152.

Wang, D.-Y. D., Proctor, R. W., and Pick, D. F. (2007), "Coding Controlled and Triggered Cursor Movements as Action Effects: Influences on the Auditory Simon Effect for Wheel-Rotation Responses," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 33, pp. 657–669.

Waszak, F., Hommel, B., and Allport, A. (2003), "Task-Switching and Long-Term Priming: Role of Episodic Stimulus-Task Bindings in Task-Shift Costs," *Cognitive Psychology*, Vol. 46, pp. 361–413.

Welford, A. T. (1952), "The 'Psychological Refractory Period' and the Timing of High-Speed Performance—A Review and a Theory," *British Journal of Psychology*, Vol. 43, pp. 2–19.

Wickelgren, W. A. (1969), "Context-Sensitive Coding, Associative Memory, and Serial Order in (Speech) Behavior," *Psychological Review*, Vol. 76, pp. 1–15.

Wightman, D. C., and Lintern, G. (1985), "Part-Task Training for Tracking and Manual Control," *Human Factors*, Vol. 27, pp. 267–283.

Willingham, D. B. (1999), "Implicit Motor Sequence Learning Is Not Purely Perceptual," *Memory and Cognition*, Vol. 27, pp. 561–572.

Willingham, D. B., Wells, L. A., Farrell, J. M., and Stemwedel, M. E. (2000), "Implicit Motor Sequence Learning Is Represented in Response Locations," *Memory and Cognition*, Vol. 28, pp. 366–375.

Wing, A. M., and Kristofferson, A. B. (1973), "Response Delays and the Timing of Discrete Motor Responses," *Perception and Psychophysics*, Vol. 14, pp. 5–12.

Winstein, C. J., and Schmidt, R. A. (1990), "Reduced Frequency of Knowledge of Results Enhances Motor Skill Learning," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 16, pp. 677–691.

Woodworth, R. S. (1899), "The Accuracy of Voluntary Movement," *Psychological Review*, Vol. 3 (Monograph Suppl.), pp. 1–119.

Woodworth, R. S. (1938), *Experimental Psychology*, Holt, New York.

Wright, C. E., Marino, V. F., Belovsky, S. A., and Chubb, C. (2007), "Visually Guided, Aimed Movements Can Be Unaffected by Stimulus-Response Uncertainty," *Experimental Brain Research*, Vol. 179, pp. 475–496.

Wright, D. L., Magnuson, C. E., and Black, C. B. (2005), "Programmingand Reprogramming Sequence Timing Following High and Low Contextual Interference Practice," *Research Quarterly for Exercise and Sport*, Vol. 76, pp. 258–266.

Zelaznik, H. N., Hawkins, B., and Kisselburgh, L. (1983), "Rapid Visual Feedback Processing in Single-Aiming Movements," *Journal of Motor Behavior*, Vol. 15, pp. 217–236.

Zupanc, C. M., Burgess-Limerick, R. J., and Wallis, G. (2007), "Performance Consequences of Alternating Directional Control-Response Compatibility: Evidence from a Coal Mine Shuttle Car Simulator," *Human Factors*, Vol. 49, pp. 628–636.

# CHAPTER 5

# INFORMATION PROCESSING

**Christopher D. Wickens**
**University of Illinois**
**Savoy, Illinois**


**C. Melody Carswell**
**University of Kentucky**
**Lexington, Kentucky**

## 1 INTRODUCTION

Information processing lies at the heart of human performance. In a plethora of situations in which humans interact with systems, the operator must perceive information, transform that information into different forms, take actions on the basis of the perceived and transformed information, and process the feedback from that action, assessing its effect on the environment. These characteristics apply whether information processing is defined in terms of the classic open-loop information-processing model that derives from much of psychological research (Figure 1*a*) or the closed-loop model of Figure 1*b*, which has its roots within both control engineering (e.g., Pew and Baron, 1978; McRuer, 1980; Jagacinski and Flach, 2003) and more recent conceptualizing in ecological psychology (Flach et al., 1995; Hancock et al., 1995). In either case, *transformations* must be made on the information as it flows through the human operator. These transformations take time
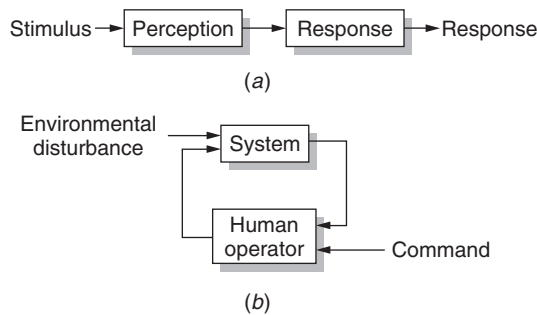
**Figure 1** Two representations of information processing: (*a*) traditional open-loop representation from cognitive psychology; (*b*) closed-loop system, following the tradition of engineering feedback models.

and may be the source of error. Understanding their nature, their time demands, and the kinds of errors that result from their operation is critical to predicting and modeling human–system interaction.

In this chapter we describe characteristics of the different important *stages* of information processing, from perception of the environment to acting on that environment. We try to do so in a way that is neither too specific to any particular system nor so generic that the relevance of the information-processing model to system design is not evident. We begin by contrasting three ways in which information processing has been treated in applied psychology, and then we describe processes and transformations related to *attention, perception, memory* and *cognition, action selection*, and *multiple-task performance*.

## 2 THREE APPROACHES TO INFORMATION PROCESSING

The classic information-processing approach to describing human performance owes much to the seminal work of Broadbent (1958, 1972), Neisser (1967), Sternberg (1969), Posner (1978), and others in the decades of the 1950s, 1960s, and 1970s, who applied the metaphor of the digital computer to human behavior. In particular, as characterized by the representation in Figure 1*a*, information was conceived as passing through a finite number of discrete stages. These stages were identifiable, not only by experimental manipulations, but also by converging evidence from brain physiology. For example, it makes sense to distinguish a perceptual stage from one involving the selection and execution of action, because of the morphological distinctions between perceptual and motor cortex.

There is also a human factors rationale for the stage distinction made by information-processing psychology. This is because different task or environmental factors appear to influence processing differentially at the different stages, a distinction that has certain design implications. For example, the *aging process* appears to affect the selection and execution of actions more than

the speed of perceptual encoding (Strayer et al., 1987). Immersed displays may improve perceptual–motor interaction, even as they inhibit the allocation of attention (Olmos et al., 2000), and different sources of workload may have different influences on the different stages (Wickens and Hollands 2000). Decision-making biases can be characterized by whether they influence perception, diagnosis, or action selection (Wallsten, 1980; Wickens and Hollands, 2000), and the different stages may also be responsible for the commission of qualitatively different kinds of errors (Reason, 1990; see Chapter 27). The support that automation provides to the human operator can also be well represented within the information-processing stage taxonomy (Parasuraman et al., 2000, 2008). Within the stage approach, there is no need to assume that processing starts at stage 1. For example, if one has an intention to act, processing can start with the response.

In contrast to the stage approach, the *ecological approach* to describing human performance provides much greater emphasis on the integrated flow of information through the human rather than on the distinct, analyzable stage sequence (Gibson, 1979; Flach et al., 1995; Hancock et al., 1995). The ecological approach also emphasizes the human's integrated interaction with the *environment* to a greater extent than does the stage approach, which can sometimes characterize information processing in a more context-free manner. Accordingly, the ecological approach focuses very heavily on modeling the perceptual characteristics of the environment to which the user is "tuned" and responds in order to meet the goals of a particular task. Action and perception are closely linked, since to act is to change what is perceived, and to perceive is to change the basis of action in a manner consistent with the closed-loop representation shown in Figure 1*b*.

As a consequence of these properties, the ecological approach is most directly relevant to describing human behavior in interaction with the natural environment (e.g., walking or driving through natural spaces or manipulating objects directly). However, as a direct outgrowth, this approach is also quite relevant to the design of controls and displays that mimic characteristics of the natural environment—the concept of *direct-manipulation interfaces* (Hutchins et al., 1985). As a further outgrowth, the ecological approach is relevant to the design of interfaces that mimic characteristics of how users *think* about a physical process, even if the process itself is not visible in a way that can be represented directly. In this regard, the ecological approach has been used as a basis for designing effective displays of energy conversion processes such as those found in a nuclear reactor (Vicente and Rasmussen, 1992; Moray et al., 1994; Vicente et al., 1995; Burns, 2000; Vicente, 2002; Burns et al., 2004).

Because of its emphasis on interaction with the natural (and thereby familiar) environment, the ecological approach is closely related to other approaches to performance modeling that emphasize people working with domains and systems about which they are experts. This feature characterizes, for example, the study of *naturalistic decision making* (Zsambock and Klein, 1997;

Kahneman and Klein, 2009; see Chapter 3), which is often set up in contrast to the representation of decision making within an information-processing framework (Wickens and Hollands, 2000).

Both the stage-based approach and the ecological approach have a great deal to offer to human factors, and the position we take in this chapter is that aspects of each can and should be selected, as they are more appropriate for analysis of the operator in a particular system. For example, the ecological approach is highly appropriate for modeling vehicle control, but less so for describing processes in reading, understanding complex instructions under stress, or dealing with highly symbolic logical systems (e.g., the logic of computers, information retrieval systems, or decision tree analysis; see Chapter 8). Finally, both approaches can be fused harmoniously, as when, for example, the important constraints of the natural environment are analyzed carefully to understand the information available for perception and the control actions allowable for action execution in driving, but the more context-free limits of information processing can be used to understand how performance might break down from a high load that is imposed on memory or dual-task performance requirements in a car.

A final approach, that of *cognitive engineering*, or *cognitive ergonomics* (Rasmussen et al., 1995; Vicente, 1999, 2002; Bizants and Roth, 2007; Jenkins et al., 2009), is somewhat of a hybrid of the two described above. The emphasis of cognitive engineering is, on the one hand, based on a very careful understanding of the environment and task constraints within which an operator works, a characteristic of the ecological approach. On the other hand, as suggested by the prominence of the word *cognitive*, the approach places great emphasis on modeling and understanding the knowledge structures that expert operators have of the domains in which they must work and, indeed,

the knowledge structures of computer agents in the system. Thus, whereas the ecological approach tends to be more specifically applied to human interaction with physical systems and environments (particularly those that obey the constraints of Newtonian physics), cognitive engineering is relevant to the design of almost any system about which the human operator can acquire knowledge, including the very symbolic computer systems, which have no physical analogy.

Whether human performance is approached from an information-processing, ecological, or cognitive engineering point of view, we assert here that, in almost any task, a certain number of mental processes involved in selecting, interpreting, retaining, or responding to information may be implemented, and it is understanding the vulnerabilities of these processes and capitalizing, where possible, on their strengths which can provide an important key to effective human factors of system design.

In this chapter we adopt as a framework the information-processing model depicted in Figure 2 (Wickens and Hollands, 2000). Here stimuli or events are *sensed and attended* (Section 3) and that information received by our sensory system is *perceived*, that is, provided with some meaningful interpretation based on memory of past experience (Section 4). That which is perceived may be responded to directly, through a process of action selection (decision of what act to take) and execution (Section 6). Alternatively, it may be stored temporarily in working memory, a system that may also be involved in thinking about or transforming information that was not sensed and perceived but was generated internally (e.g., mental images, rules, Section 5). Working memory is of limited capacity and heavily demanding of attention in its operation but is closely related to our large-capacity long-term memory, a system that stores vast amounts of information about the world, including both facts and procedures, and
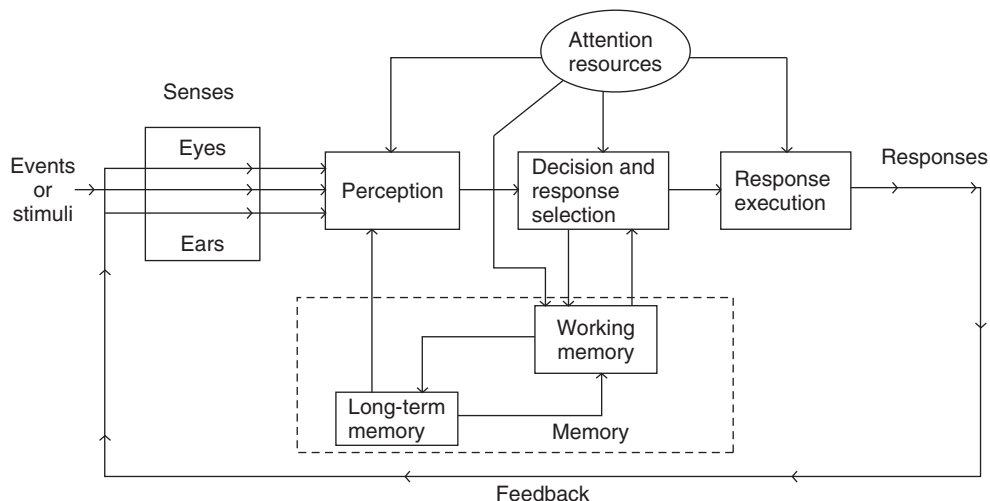


**Figure 2**  Model of human information processing. (Adapted from Wickens, 1992.)

retains that information without attention but not always fully available for retrieval.

As noted in the figure and highlighted in the ecological approach, actions generally produce feedback which is then sensed to complete the closed-loop cycle. In addition, human attention, a limited resource, plays two critical roles in the information-processing sequence (Wickens, 2007; Wickens and McCarley, 2008). As a selective agent, it chooses and constrains this information that will be perceived (Section 3). As a task management agent, it constrains what tasks (or mental operations) can be performed concurrently (Section 7).

## 3   SELECTING INFORMATION

Since Broadbent's (1958) classic book, it has been both conventional and important to model human information processing as, in part, a filtering process. This filtering is assumed to be carried out by the mechanisms of *human attention* (Kahneman, 1973; Parasuraman et al., 1984; Damos, 1991; Pashler, 1998; Johnson and Proctor, 2004). Attention, in turn, may be conceptualized as having three *modes*: Selective attention chooses what to process in the environment, focused attention characterizes the efforts to sustain processing of those elements while avoiding distraction from others, and *divided attention* characterizes the ability to process more than one attribute or element of the environment at a given time.

We discuss below the human factors implications of the selective and focused attention modes and their relevance to visual search and discuss those of divided attention in more detail in Sections 4.6 and 7.

### 3.1   Selective Attention

In complex environments, selective attention may be described in terms of how it is influenced by the combined force of four factors: salience, effort, expectancy, and value (Wickens et al., 2003). These influences can often be revealed by eye movements when visual selective attention is assessed by visual scanning; obviously, however, eye movements cannot reflect the selectivity between vision and the other sensory modalities, such as auditory or tactile (Sarter, 2007).

1. *Salient features* of the environment will attract, or "capture," attention. Thus, auditory sounds tend to be more attention grabbing than visual events, leading to the choice of sounds to be used in alarms (Stanton, 1994; see also Chapterx 24). Within a visual display, the onset of a stimulus (e.g., increase in brightness from zero, or the appearance of an object where one was not present previously) tends to be the most salient or attention-attracting property (Yantis, 1993; Egeth and Yantis, 1997); other features, such as uniqueness, can also attract attention, but these are typically less powerful than onsets. The prominent role of onsets as attention-capturing devices can explain the value of repeated onsets ("flashing") as a visual alert. Salient events

are sometimes described as those that govern *bottom-up* or stimulus-driven allocation of attention, to be contrasted with knowledge-driven or *top-down* features of selective attention, which we describe next in the context of expectations.

2. *Expectancy* refers to knowledge regarding the probable time and location of information availability. For example, frequently changing areas are scanned more often than slowly changing areas (Senders, 1980). Thus, drivers will keep their eyes on the road more continuously when the car is traveling fast on a curvy road than when traveling slowly on a straight one. The former has a higher "bandwidth." Also, expectancy defines the role of cueing in guiding attention. For example, an auditory warning may direct attention toward the display indicator that the warning is monitoring, because the operator will now expect to see an abnormal reading on that display.

3. *Value* is a third factor. High-frequency changes are not, however, sufficient to direct attention. The driver will not look out the side window despite the fact that there is a lot of perceptual "action" there, because information in the side window is generally not relevant to highway safety. It has already passed. Thus, selective attention is also driven by the *value* of information received at different locations. This describes the importance of knowing that information in carrying out useful tasks, or the costs of failing to note important information. It is valuable for the driver to look forward, because of the cost of failing to see a roadway hazard or of changing direction toward the side of the road. Thus, the effect of expectancy (bandwidth or frequency) on the allocation of selective attention is modulated by the value, as if the probability of attending somewhere, $p(A)$, is equal to the *expected value* of information sources to be seen at that location (Moray, 1986; Wickens et al., 2003; Wickens et al., 2008). Indeed, Moray (1986) and Wickens et al. (2003) find that well-trained, highly skilled operators scan the environment very much as if their attention is driven primarily and nearly exclusively by the multiplicative function of expectancy and value. Thus, we may think of the well-trained operator as developing scanning *habits* that internalize the expectancy and value of sources in the environment, defining an appropriately calibrated *mental model* (Bellenkes et al., 1997).

4. The final factor that may sometimes influence attention allocation is a negative one, and this factor, unique to eye movements, is the *effort* required to move attention around the environment. Small attention movements, such as scanning from one word to the next in a line of text or a quick glance at the speedometer in a car, require little effort. However, larger movements, such as shifting the eyes *and* head to check the side-view mirrors in a car, or coupling these

with rotation of the trunk to check the "blind spot" before changing lanes, requires considerably more information access effort (Ballard et al, 1995; Wickens, 1993). Indeed, the role of the information access effort also generalizes to the effort costs of using the fingers to manipulate a keyboard and access printed material that might otherwise be accessed by a simple scan to a dedicated display (Gray and Fu, 2001). The effort to shift attention to more remote locations may have a minimal effect on the well-rested operator with a well-trained mental model who knows precisely the expectancy and value of information sources (Wickens et al., 2003). However, the combination of fatigue (depleting effort) and a less well calibrated mental model can seriously inhibit accessing information at effortful locations, even when such information may be particularly valuable.

Collectively, the forces of salience, effort, expectancy, and value on attention can be represented by a visual attention model called SEEV, in which $P(A) = S - EF + EX \times V$ (Wickens et al., 2003, 2008, 2009a; Horrey et al., 2006). Good design should try to reduce these four components to two by making valuable information sources salient (correlating salience and value) and by minimizing the effort required to access valuable *and* frequently used (expected) sources. For example, head-up displays (HUDs) in aircraft and automobiles are designed to minimize the information access effort of selecting the view outside and the information contained in important instruments (Fadden et al., 2001; Wickens et al., 2004). Reduced information access effort can also be achieved through effective layout of display instruments (Wickens et al., 1997).

While SEEV may predict what **is** attended, and salience highlights the roll of bottom-up attention capture, a large body of research has also recently focused on *attentional blindness,* particularly *change blindness,* highlighting the events in the world that are not noticed, even when they may be valuable (Simons and Levin, 1997 Rensink, 2002; Wickens et al., 2009a).

It seems that the human's attention system is not well designed to notice unexpected events, particularly when they appear in peripheral vision, under conditions of high workload. Naturally if these events are not salient as well (e.g., the offset of a stimulus, or a change in a word, say from "off" to "on"), then noticing will degrade still further. While in many circumstances such events may be noticed a majority of the time, when the events are safety critical (as in the above, "on" designating the activation of power) and relatively rare (referred to as "black swans"; Taleb, 2007; Wickens et al., 2009a), human miss rates as low as 10–20% can illuminate the safety concerns of change blindness (Wickens, 2009).

## 3.2   Focused Attention

While selective attention dictates where attention should travel, the goal of focused attention is to maintain processing of the desired source and avoid the distracting influence of potentially competing sources. The primary sources of breakdowns in focused attention are certain physical properties of the visual environment (clutter) or the auditory environment (noise), which will nearly guarantee some processing of those environments, whether or not such processing is desired. Thus, any visual information source within about $1°$ of visual angle of a desired attentional focus will disrupt processing of the latter to some extent (Broadbent, 1982). Any sound within a certain range of frequency and intensity of an attended sound will have a similar disruptive effect on auditory focused attention (Banbury et al., 2001; see Chapter 3). However, even beyond these minimum limits of visual space and auditory frequency, information sources can be disruptive of focused attention if they are salient.

## 3.3   Discrimination and Confusability

A key to design that can address issues of both selective and focused attention is concern for *discrimination* between information sources. Making sources discriminable by space, color, intensity, frequency, or other physical differences has two benefits. First, it will allow the display viewer to *parse* the world into its meaningful components on the basis of these physical features, thereby allowing selective attention to operate more efficiently (Treisman, 1986; Yeh and Wickens, 2001; Wickens et al., 2004). For example, an air traffic controller who views on her display all of the aircraft within a given altitude range depicted in the same color can easily select all of those aircraft for attention to ascertain which ones might be on conflicting flight paths. Parsing via a discrimination will be effective as long as all elements that are rendered physically similar (and therefore are parsed together) share in common some characteristic that is relevant for the user's *task* (as in the example above, all aircraft at the same altitude represent potential conflicts).

Second, when elements are made more discriminable by some physical feature, it is considerably easier for the operator to *focus* attention on one and ignore distraction from another, even if the two are close together in space (or are similar in other characteristics). Here, again, in our air traffic control example, it will be easier for the controller to focus attention on the converging tracks of two commonly colored aircraft if other aircraft are colored differently than if all are depicted in the same hue.

Naturally, the converse of difference-based discriminability is similarity- (or identity-) based confusion between information sources, a property that has many negative implications for design. For example, industrial designers may strive for consistency or uniformity in the style of a particular product interface by making all touchpad controls the same shape or size. Such stylistic uniformity, however, may result in higher rates of errors from users activating the wrong control because it looks so similar to the control intended.

## 3.4   Visual Search

Discrimination joins with selective and focused attention when the operator is engaged in *visual search*, looking for something in a cluttered environment (Wickens and

McCarley, 2008). The task may characterize looking for a sign by the roadway (Holohan et al., 1978), conflicting aircraft in an air traffic control display (Remington et al., 2000), a weapon in an X-rayed luggage image (McCarley et al., 2004), a feature on a map (Yeh and Wickens, 2001), or an item on a computer menu (Fisher et al., 1989). *Visual search models* are designed to predict the time required to find a target. Such time predictions can be very important for both safety (e.g., if the eyes need to be diverted from vehicle control while searching) and productivity (e.g., if jobs require repeated searches, as in quality control inspection or menu use).

The simplest model of visual search, based on a *serial self-terminating search* (Neisser et al., 1964), assumes that a *search space* is filled with items most of which are *nontargets* or *distracters*. The mean time to find a target is modeled to be RT = $NT/2$, where $N$ is the number of items in the space, $T$ is the time to examine each item and determine that it is *not* a target before moving on to the next, and division by 2 reflects the fact that on average the target will be reached after half of the space is searched, but sometimes earlier and sometimes later. Hence, the variance in search time will also grow with $N$. Importantly, in many displays, we can think of $N$ as a very functional measure of *clutter* (Yeh and Wickens, 2001).

The elegant and simple prediction of the serial self-terminating search model often provides a reasonable accounting for data (Yeh and Wickens 2001; Remington et al., 2000) but is also thwarted (but search performance is improved) by three factors that characterize search in many real-world search tasks: bottom-up parallel processing, top-down processing, and target familiarity. The first two can be accommodated by the concept of a *guided search model* (Wolfe, 2007; Wolfe and Horowitz, 2004). Regarding parallel processing, as noted in Section 3.1, certain features (e.g., uniqueness, flashing) will capture attention because they can be preattentively processed or processed in parallel (rather than in series) with all other elements in the search field. Hence, if the target is known to contain such features, it will be found rapidly, and search time will be unaffected by the number of nontarget items in the search field. This is because all nontarget items can be discriminated automatically (as discussed in Section 1) and thereby eliminated from imposing any search costs (Yeh and Wickens, 2001; L. D. Wickens, Alexander et al., 2004). For example, in a police car dispatcher display, all cars currently available for dispatching can be highlighted, and the dispatcher's search for the vehicle closest to a trouble spot can proceed more rapidly. Stated in other terms, search is "guided" to the subset of items containing the single feature which indicates that they are relevant. If there is more than a single such item, the search may be serial between those items that remain. *Highlighting* (Fisher et al., 1989; L. D. Wickens, Alexander et al., 2004; Remington et al., 2000) is a technique that capitalizes on this guided search.

Regarding top-down processing, search may also be guided by the operator's knowledge of *where* the target is most likely to be found. Location expectancy, acquired with practice and expertise, will create search strategies that scan the most likely locations first, to the extent that such predictability exists in the searched environments. For example, tumors may be more likely to appear in some parts of an organ than others, and skilled radiologists capitalize on this in examining an X ray in a way that novices do not (Kundel and Nodine, 1978). However, such a strategy may not be available to help the scanner of luggage X rays for weapons, because such weapons may be hidden anywhere in the luggage rather than in a predicable location (McCarley et al., 2004).

A second influence of top-down processing on search is the expectancy of whether a target will be present or not, the "target prevalence rate." Wolfe et al. (2005) observe that a low expectancy for targets will lead searchers to terminate their search prematurely, even though the target may still be present in the cluttered search field.

A third factor that can speed visual search, target familiarity is, like guided search, related to experience and learning and, like parallel search, related to salient features. Here we find that repeated exposures to the same consistent target can speed the search for that target and, in particular, reduce the likelihood that the target may be looked at (fixated) but not actually detected (McCarley et al., 2004). With sufficient repetition looking for the same target (or target possessing the same set of features), the expert tunes his or her sensitivity to discriminate target from nontarget features, and with extensive practice, the target may actually "pop out" of the nontargets, as if its discriminating features are processed preattentively (Schneider and Shiffrin, 1977). Further, even if a target does not become sufficiently salient to pop out when viewed in the visual periphery, repeated exposure can help ensure that it will be detected and recognized once the operator has fixated on it (McCarley et al., 2004).

The information processing involved in visual search culminates in a target detection decision, which sometimes may be every bit as important as the search operations that preceded it. In the following section we examine this detection process in its own right.

## 4 PERCEPTION AND DATA INTERPRETATION

### 4.1 Detection as Decision Making

At the top of many display design checklists is a reminder that critical targets must be detectable in the environment for which they are intended (e.g., Travis, 1991; Sanders and McCormick, 1993). Assuring such detectability might seem to be a simple matter of knowing enough about the limits of the operator's sensory systems to choose appropriate levels of physical stimulation, for example, appropriate wavelengths of light, frequencies of sound, or concentrations of odorants. Human sensitivity to the presence and variation of different physical dimensions is reviewed in Chapter 3, and these data must be considered limiting factors in the design of displays. Yet the detectability of any critical signal is also a function of the operator's goals, knowledge, and expectations. As noted in our discussion of

**Table 1  Joint Contingent Events Used in Signal Detection Theory Analysis**

| Operator's Decision (Response Criterion) | State of the World | |
|---|---|---|
| | Signal | No Signal (Noise) |
| Signal | Hit | False alarm |
| No signal | Miss | Correct rejection |

visual search, there are plenty of opportunities for targets that are clearly above threshold to be missed when the operator is hurried and the display is cluttered. As we noted in our discussion of change blindness above, the magnitude of the superthreshold changes in a scene that can be missed is often surprising (Rensink, 2002).

The interpretive and vulnerable nature of even the simplest signal detection task becomes most apparent when we consider that missing a target is not the only possible detection error; we may also make *false alarms*, responding as if a signal is present when it is not (see Table 1). *Signal detection theory* (SDT) provides a valuable conceptual and computational framework for describing the processes that can lead to both types of errors (Green and Swets, 1989; Wickens, 2002; MacMillan and Creelman, 2005). In SDT, signals are never assumed to occur against a "clean" background of zero stimulation. Instead, all signals occur against a background of fluctuating *noise*. The noise arises from both internal (e.g., baseline neuronal activity) and external sources. The observer's detection task is thus, in reality, a *decision* task: Is the level of stimulation experienced at any moment the results of high levels of noise or does it represent the presence of a signal? Because noise is always present and is always fluctuating in intensity, detection errors are inevitable.

To deal with the uncertainty inherent in detection, SDT asserts that operators choose a level of sensory excitation to serve as a *response criterion*. If excitation exceeds this criterion, they will respond as if a signal is present. Operators who raise their response criteria, making them more conservative, will also increase the likelihood of missing targets. Lowering their criteria, however, will decrease the number of misses at the expense of increased false alarms. Signal detection theory provides a way to describe the criterion set by a particular operator performing a particular detection task and of determining the optimality of the selected criterion in the face of task characteristics such as signal probabilities and the relative repercussions (i.e., practical outcomes) of making the two types of errors.

Signal detection theory formally demonstrates that as signal probability increases, response criteria should be lowered in order to minimize overall error rates. People performing laboratory detection tasks tend to adjust their response criteria in the direction prescribed by SDT; however, they do *not* tend to adjust them far enough (Green and Swets, 1989). Probability-related shifts in response criteria also seem to occur in a wide variety of operational settings. For example, Lusted (1976) found that physicians' criteria for detecting particular medical conditions were influenced by the base rate of the abnormality (probability of signal). Similarly, industrial inspectors adjusted their criteria for fault detection based on estimated defect rates (Drury and Addison, 1973), although they fail to adjust their criteria enough when defect rates fall below 5% (Harris and Chaney, 1969). Many errors in the judicial process may also be linked to the biasing effects of implicit and potentially unreliable clues about signal probability. Saks et al. (2003) argue that such probability estimates influence the performance of forensic scientists asked to detect critical matches in evidence such as fingerprints, bite marks, and bomb residues. Research has also demonstrated an effective intervention for operators with overly low response criteria: Inserting "false signals" into some inspection tasks can increase perceived signal probability and, as a result, shift response criteria downward (Baker, 1961; Wilkinson, 1964).

A second factor that should influence the setting of the response criterion, according to SDT, is the relative costs associated with misses and false alarms and the relative benefits of correct responses. As an extreme example, if there were dire consequences associated with a miss and absolutely no costs for false alarms, the operator should adopt the lowest criterion possible and simply respond as if the signal is there at *every* opportunity. Usually, however, circumstances are not so simple. For example, a missed (or delayed) air space conflict by the air traffic controller or a missed tumor by the radiologist may have enormous costs, possibly in terms of human lives. However, actions taken because of false alarms, such as evasive flight maneuvers or unnecessary surgery, also have costs. The operator should adjust his or her response criterion downward to the degree that misses are more costly than false alarms and upward to the extent that avoiding false alarms is more important.

An important use of SDT in human factors research is often to diagnose the source of unsatisfactory detection performance. Has the operator failed to appropriately calibrate his or her response criterion to actual signal probabilities and response outcomes? Or, are limitations in the sensitivity of the operator's own (internal) signal-processing systems at fault? Depending on the answers to these questions, interventions can be devised to enhance detection performance. In the case of sensory limitations, engineering innovations may be required to enhance the fundamental signal-to-noise ratio, or greater target exposure may be necessary to enhance the operator's sensory tuning to critical target features (Gold et al., 1999). For example, attempts to increase the performance of airport luggage screeners have led to the development of a threat image projection (TIP) system for on-the-job training (Schwaninger and Wales, 2009). The system intermittently projects "false threat" images onto actual X-ray images, giving screeners greater exposure to potential targets (increasing overall sensitivity) and increasing their estimates of signal probability as well (thus, keeping their response criteria relatively low).

The job of the baggage screener exemplifies a combination of demands that can prove particularly challenging to operators—detection of low-probability signals

over lengthy monitoring periods. Such *vigilance tasks*, and the predictable *vigilance decrements* in detection performance that occur as the watch progresses, have been the target of extensive research over much of the last century (Warm, 1984). In reviewing this literature, Warm et al. (2008) have argued that vigilance tasks are a part of many modern jobs and are both more mentally demanding and stressful than is often realized. As a way of dealing with such demands and the relatively poor detection performance that often results, designers often develop alarms and alerts to assist or sometimes replace the operator (Stanton, 1994). Yet considerable evidence suggests that such automation does not eliminate the detection problem. Automated alerts must also confront challenging issues of distinguishing signals from highly similar noise (e.g., friend and foe on a military image display), and such alert systems can be counted on to create errors. Thus, the alarm designer, rather than the human monitor, is now the agent responsible for adjusting the response criterion of the alarm, to trade off misses versus false alarms, and designers are often tempted to make this adjustment in such a way that signals are never missed by their systems. (Consider the possible litigation if a fire alarm fails to go off.) However, when the human and automation are considered as a total system, the resulting increase in automation false alarms can have serious consequences (Sorkin and Woods, 1985; Dixon et al., 2007). These consequences arise because a high false-alarm rate can lead to serious issues of *automation mistrust*, in which people may ignore the alarms altogether (Sorkin, 1989) and experience the "cry wolf" phenomenon (Breznitz, 1983).

The analysis of diagnostic systems also reveals that the problems of high false-alarm rates will be further amplified to the extent that the signals to be detected themselves occur rarely—the "low-base-rate problem," as is often the case with alarm systems (Parasuraman et al., 1997), so that a large majority of the alarms that do sound will be false alarms. Answers to these problems lie in part in making available to human perception the raw data of the signal domain that is the basis of the alarm (Wickens et al, 2009b); and this appears to mitigate or even eliminate the cry-wolf effect. There is some evidence that *likelihood alarms* that can signal their own degrees of certainty in graded form (rather than a two-state on–off logic) will assist (Sorkin and Woods, 1985; St. Johns and Manes, 2002). Finally, it is reasonable to assume that training operators as to the nature of the mandatory miss/false-alarm trade-off, and the inevitable high false-alarm rate when there are low-base-rate events, should mitigate problems of distrust to some degree.

## 4.2   Expectancy, Context, and Identification

We have seen that our knowledge and expectations about the world help determine the efficiency of both our visual search and signal detection performance. Based on our past experience, attention can be directed to locations where targets are more likely to occur and response criteria can be adjusted according to the perceived likelihood or value of signals. We now consider how our ability to identify enormous numbers of objects from a variety of vantage points, often under less than optimal conditions, is also dependent on prior knowledge. Much research in object perception has been devoted to determining how the co-occurrence of particular stimulus attributes helps define individual objects (e.g., Biederman, 1987). However, research in cognitive psychology and computer vision has also provided compelling demonstrations of how our knowledge of statistical dependencies across entire scenes and events helps us make "educated guesses" about the likely identity of constituent objects (Oliva and Torralba, 2010).

Many studies, ranging from those using simple, artificial stimuli to those using complex, naturalistic scenes, have demonstrated that objects and attributes are recognized more quickly when they are embedded in consistent contexts (i.e., those in which the targets are naturally and frequently found) rather than when presented alone or in inconsistent contexts. Words are more efficiently identified when embedded in meaningful sentences rather than alone (e.g., Tulving et al., 1964). Individual letters are recognized more efficiently when embedded in words rather than when presented alone or as part of nonwords (Reicher, 1969). Caricature facial features require less physical detail for recognition when they are embedded in a face (Palmer, 1975). Photographs of naturalistic scenes also seem to enhance the identification of objects typically found there (Biederman et al., 1981). Even relatively subtle (but meaningful) object relations can influence identification; for example, an image of a glass is identified more rapidly when presented with an image of a pitcher that has its spout oriented toward the glass rather than away from it (Green and Hummel, 2006).

Explanations for context effects generally assume that added items in the stimulus array will increase the odds that the operator will recognize at least some portion of it. Even if the portion immediately recognized is not the target object, the recognition is still useful because it reduces the likelihood that some stimuli will be encountered while increasing the likelihood of others. In this way, the total set of possible objects, words, or letters from which the observer must choose becomes smaller and more manageable [see Massaro (1979) and McClelland and Rumelhart (1981) for formal models].

It is important to note that not all research on context effects focuses on our use of knowledge about contingencies among and within perceptual objects. Some researchers are more interested in identifying global *scene statistics* that might constrain the processing (and enhance the efficiency) of our interpretation of local scene elements (e.g., Ariely, 2001). These global properties may not be as subjectively accessible as the objects that form the scene, but they may provide an efficient means to reduce perceptual load. These statistics may include the mean size and variance of a set of objects, the center of mass, and textural properties (Oliva and Torralba, 2010). Evidence is accumulating that we do, in fact, extract global scene statistics such as mean object size, even of nonattended display elements (e.g., Chong and Treisman, 2005).

Taken together, these findings suggest that the old design maxim that "less is more" may well be wrong when the goal is to support the operator's identification of task-critical information. Unlike performance in visual search tasks, where additional nontarget stimuli or events usually cause declines in performance, identification tasks often benefit from the presence of additional stimuli, as long as those stimuli are normally encountered in spatial or temporal proximity to the target. In fact, operator expectancies can be used to offset degraded stimulus conditions such as poor print reproductions, faulty lighting, brief stimulus exposures, presentation to peripheral vision, or even the momentary diversion of attention. Therefore, the redundant use of red, an octagonal shape, and the letters S–T–O–P enhances the identification of a stop sign, as does its expected location to the driver's right immediately before an intersection. A "less is more" stop sign that only used the letter S as a distinguishing feature would not be advised!

In addition to the design implications of providing a *consistent* context for critical information, context effects also warn of the dangers of contextual *inconsistency*. The detection of such inconsistencies by safety professionals may be particularly useful for identifying environmental hazards. For example, a sidewalk interrupted by isolated steps may be a dangerous spot for a fall if not surrounded by changes in scene texture that typically indicate abrupt changes in landscape elevation [see Cohen (2003) for further examples of expectancy effects in trips and falls]. The availability of normative scene statistics may one day contribute to the identification of such hazards and is consistent with the ecological approach to human information processing.

## 4.3 Judgments of Two-Dimensional Position and Extent

Both detection and identification are categorical judgments. Sometimes, however, we may be more interested in determining specific qualitative properties of a stimulus such as its location and the magnitude of its various properties (e.g., length, volume, orientation). These judgments are critical for manual control and locomotion (see Section 6) as well as for the interpretation of maps, graphs, and dynamic analog indicators. In this section we focus mainly on spatial judgments of static formats before turning to their dynamic counterparts.

It is well known that the spatial judgments required to read even the most everyday graphs are prone to systematic distortions, knowledge of which can sometimes be used to manipulate a graph's message [e.g., Penrose (2008) evaluates the prevalence of such distortions in corporate annual reports]. Some examples include:

1. Our overestimation of values represented in bar graphs, especially with shorter bars and those farthest from the y axis (Graham, 1937)

2. The perceptual flattening of line graphs with respect to the y axis, resulting in larger underestimations of the represented data as the reader follows the line from its origin (Poulton, 1985)

3. Cyclic patterns of bias in estimations of part–whole relationships (e.g., reading pie charts) that are dependent on the number of available reference points on the graphs (Hollands and Dyre, 2000)

4. Distance distortions between cursor locations and target/icon locations induced by the shape of the cursor (overestimations and overshooting occur when the cursor is an arrow pointed toward the target; Phillips et al., 2003)

These distortions in distance and size may be special cases of geometric illusions such as those reviewed by Gillam (1980) and Gregory (1980). Poulton (1985), for example, has ascribed the perceptual flattening of lines in graphs to the Poggendorf illusion (Figure 3a). Phillips et al. (2003) refer to the Muller–Lyer illusion (Figure 3b) when describing the overestimations of distances between cursor and target. A variety of information-processing mechanisms have been proposed for such illusions, including misallocation of attention, and our tendency to misapply depth cues to wo-dimensional (2D) images. Whatever the cause, design modifications can reduce the impact of these illusions. Poulton (1985), for example, found that simply adding a redundant y axis to the right side of his graphs effectively reduced point-reading errors. Lowlighted horizontal gridlines can have the same beneficial effect. Kosslyn (2006) describes more ways to minimize the likelihood of these graphical illusions as well as ways to reduce graphical miscommunication more generally.

Although there are many salient examples of the use of perceptual illusions to misrepresent data in graphical displays, the presence of illusions is not always harmful. Some designers have even used illusions of size to increase traffic safety. Shinar et al. (1980) painted a pattern similar to that used to induce the Wundt illusion (see Figure 3c) on a roadway leading to a dangerous obscured curve. After the roadway was painted, drivers tended to reduce their speed before encountering the curve, presumably because the pattern made the road seem narrower.

The most systematic work on the size and nature of errors in the perception of graphical displays has focused
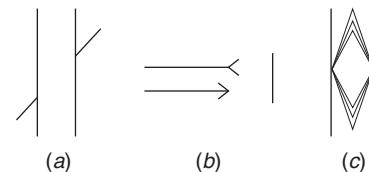


(a)          (b)          (c)

**Figure 3** Three perceptual illusions, influencing the perception of location and spatial extent: (a) Poggendorf illusion, in which two diagonal line segments that are actually collinear do not appear so; (b) Muller–Lyer illusion, in which the distances between the horizontal line segment and the tips of the two arrowheads appear to be different, even though they are not; (c) Wundt illusion, in which two parallel vertical lines appear to curve inward.
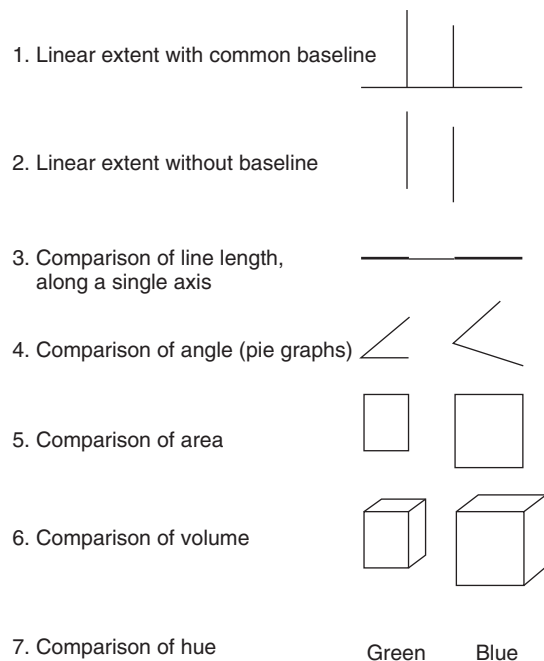
1. Linear extent with common baseline

2. Linear extent without baseline

3. Comparison of line length, along a single axis

4. Comparison of angle (pie graphs)

5. Comparison of area

6. Comparison of volume

7. Comparison of hue     Green     Blue

**Figure 4** Graphical dimensions for making comparative judgments. (From Wickens, 1992; used by Cleveland and McGill, 1985.)

on how precisely we can make comparisons *between* data values. Cleveland and McGill (1984; Cleveland, 1985) developed a list of the physical dimensions that are commonly used to represent data in graphs and maps. These dimensions were ordered, as shown in Figure 4, in terms of the accuracy with which they could be used to make relative-magnitude judgments (e.g., "What percentage is point *A* of point *B*?"). Using this list as a guide, designers are advised, for example, to avoid using variation in volume or area to represent data. Thus, pictograms should usually be avoided in favor of bar charts or point displays. Similar lists have also been proposed for other sensory modalities. For example, Wall and Brewster (2003) have suggested that haptic graphs should use friction to represent data values rather than stiffness or the spatial period of sinusoidal textures.

Although a sensible first step in designing displays, lists of this type do not ensure the eventual efficacy of graphs. The Cleveland–McGill list of preferred graphical dimensions, for example, predicts performance in graph-reading tasks less well when users move from making simple comparisons to performing more integrative tasks such as describing trends (Carswell, 1992). Furthermore, each step down the Cleveland–McGill list is not equally detrimental to performance. Position, length, and angle judgments are associated with small differences from one another, but all three are used much more accurately than either area or volume. As we will see next, the misperception of volume and other spatial

stimulus dimensions may reflect ambiguities in our perception of three-dimensional (3D) form and distance.

Additional comprehensive coverage of information-processing factors and biases in graph interpretation can be found in Kosslyn (2006), Gillan et al. (1998), and Wickens and Hollands (2000).

### 4.4 Judgments of Distance and Size in Three-Dimensional Space

Judgments of extent and position, discussed in the preceding section, are also made in 3D space, a space that can be either true space (e.g., judging whether there is adequate room to pass a car on a two-lane road) or a display-synthesized 3D space (e.g., comparing the volume of cubes in Figure 5). When making judgments in either real or synthesized virtual spaces, human perception depends on a host of cues to provide information about the absolute or relative distance from the viewer (Cutting and Vishton, 1995). Many of these depth cues are called *pictorial cues* because they can be used to generate the impression of depth in 2D pictures such as Figure 6. Here the convergence of the edges of the roadway at the horizon, the *linear perspective*, suggests that the roadway is receding. The decreasing grain of the texture of the cornfield moving
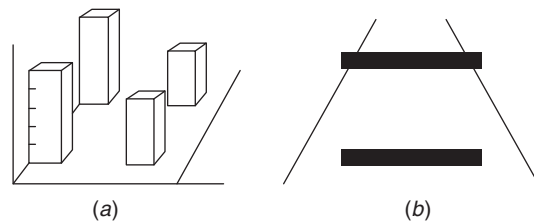


**Figure 5** Role of size constancy in depth perception in creating illusions: (*a*) distorted overestimation of the size of the more distant bar graphs; (*b*) Ponzo illusion, illustrating the greater perceived length of the more distant bar.
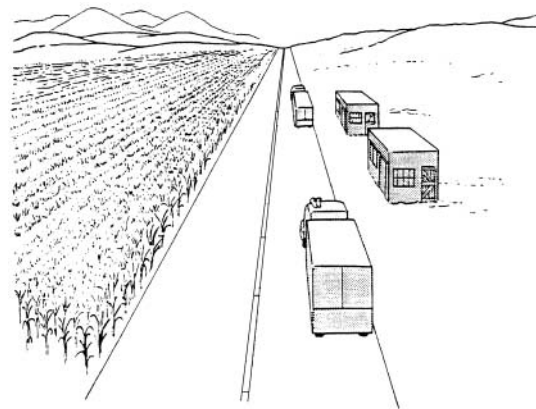


**Figure 6** Perceptual cues for depth perception. (From Wickens, 1992.)

from the bottom to the top of the figure, the *textural gradient*, also informs us that the surface is receding, as well as revealing the slant angle of the surface and the vantage point of the viewer above the landscape. Three additional pictorial cues allow us to judge which building is closer. The building closer to the top of the image is signaled to be farther away (*height in the plane*), as is the building that captures the smaller retinal image (*relative size*); the building that obscures the contours of the other is seen to be closer (*interposition or occlusion*).

In addition to the pictorial cues, which are part of the image itself, there are five cues that result from characteristics of the viewer. *Motion parallax* results whenever the viewer moves, with nearer objects producing greater relative motion across the visual field than distant objects. *Binocular disparity* refers to the difference in viewpoint of the two eyes, a difference that diminishes exponentially as objects are farther from the viewer. *Stereopsis*, the use of binocular disparity to perceive depth, can occur in 3D displays when slightly different images are presented to the two eyes (Patterson, 1997). Some depth information is also obtained from *accommodation* and *binocular convergence*. These cues result from the natural adjustments of the eyes needed to focus on specific objects at different distances. Accommodation is the response of the lens required to bring very close objects into focus, and convergence is the "cross-eyed" viewing of the two eyes, also necessary to bring the image of closer objects into focus on the back of both retinas.

In viewing real 3D scenes, most of these depth cues operate redundantly and fairly automatically to give us very precise information about the relative distance of objects in the visual scene and adequate information about the absolute distance (particularly of nearby objects). Such distance judgments are also a necessary component of judgments of 3D shape and form. A host of research studies on depth perception reveal that the depth cues respond in a generally additive fashion to convey a sense of distance (see Wickens et al., 1989). Thus, as the viewer looks at a 3D display, the more depth cues that are available to the viewer, the more perceived separation there is between objects in depth (i.e., along the $Z$ axis in 3D space).

Although all depth cues contribute to the sense of distance, there is evidence that three of those cues (i.e., motion parallax, binocular disparity, and interposition) are the most powerful and will dominate other cues when they are placed in opposition (e.g., Braunstein et al., 1986; Wickens et al., 1989, 2000). Hence, in designing a 3D display to synthesize the 3D spatial world, it is a good idea to try to incorporate at least one, if not two, of these dominant cues.

In constructing a 3D perceptual representation from a displayed or real 3D image, people may often be guided by knowledge-driven expectancies when interpreting the bottom-up distance cues. For example, the use of relative size as an effective cue depends on the viewer's assumption about the *true size* of the objects to be compared. If the assumptions are inaccurate, the use of relative size to judge distance can lead to illusions,

sometimes dangerous ones. For example, Eberts and MacMillan (1985) concluded that the high rate of rear-end collisions suffered by small cars resulted from drivers' misuse of relative size. The trailing driver, seeing the smaller-than-expected retinal image of the small car, perceives the car to be of normal size and farther away. The perception of greater distance between cars, in turn, may lead to dangerously delayed braking.

Recent research indicates that reliance on other pictorial depth cues may lead to equally poor distance estimations on the part of drivers. Buchner et al. (2006) found that cars with higher backlights are at an increased risk of being perceived as more distant than they actually are. This illusion may be due to the misapplication of the depth cue that associates height in the picture plane with increased distance. Naturally, such illusions, based on inappropriate application of knowledge and expectancies, are more likely to occur when there are fewer depth cues available. This is why pilots are particularly susceptible to illusions of depth and distance (Previc and Ercoline, 2004), since many of the judgments they must make are removed from the natural coordinate framework of Earth's surface and must sometimes be made in the degraded conditions of haze or night.

Besides the problems of impoverished cues, a second problem with 3D displays is *line-of-sight ambiguity*. This is illustrated in Figure 7, which depicts a viewer looking into a volume containing three objects, A, B, and C, as we see the views from the side (top view). The view of these objects on the screen is shown below. Here we see that when the viewer makes judgments of position along the viewing axis into the 3D world, a given distance in the world is represented by a smaller visual angle than when judgments are made parallel to the display surface, a phenomenon known as *compression* (Stelzer and Wickens, 2006). In Figure 7 judgment of the distance $AB$ along the $Z$ axis is compressed, whereas the judgment of the distance $AC_1$ along the $X$ axis is not. As a consequence of this reduced resolution along the viewing axis, it is harder to tell exactly how distant things are. This distance ambiguity, in turn, has repercussions for the ability to make precise spatial comparisons between objects in the plane orthogonal to the line of sight. Returning to Figure 5*a*, note how difficult it is to tell if the difference in height between the two distant bars is the same or different from the closer ones.

To make matters worse, it is often difficult in 3D displays to resolve the extent to which an object displaced to a new location is receding in depth or is moved to a higher location at the same depth, a further form of ambiguity. For example, the various movements of point $C$ in Figure 7 to points $C_1$, $C_2$, and $C_3$ would all appear nearly equivalent to the viewer of the 3D display with few depth cues, since they all would occupy the same position along the line of sight into the display; the relative contribution of altitude to distance change would be difficult to resolve.

Of course, as we have noted, some of this ambiguity can be resolved in display viewing if the designer incorporates progressively more depth cues. Yet in many
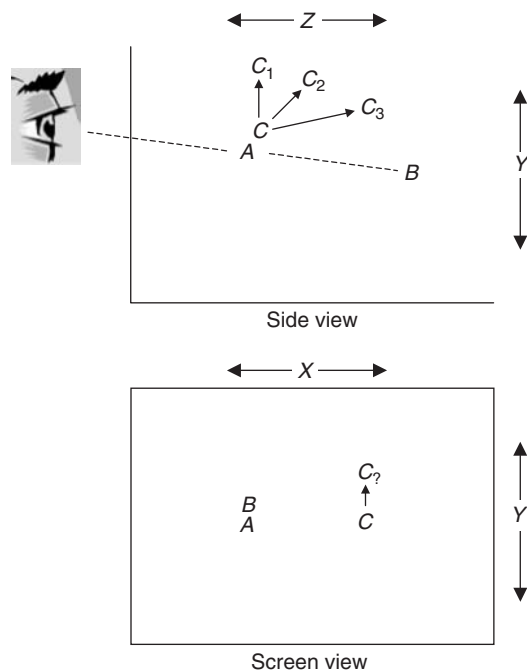
**Figure 7** Ambiguity in 3D displays. The actual distance of objects A, B, and C from the observer is shown from the side, in the top view. The screen viewed by the observer is shown below. From the 3D screen view it may be ambiguous whether A or B is closer. Movement from $C$ to $C_1$, $C_2$ and $C_3$ may all look identical.

circumstances it may be cumbersome or computer intensive to incorporate the most powerful cues of stereo and relative motion realistically; furthermore, there are certain tasks, such as those involved in air traffic control, precise robotic control, and some forms of minimally invasive surgery, in which the requirement for very precise spatial judgments with no ambiguity on any axis is so strong that a set of 2D displays, from orthogonal viewing axes, may provide the best option, even if they present a less natural or realistic view of the world (Wickens, 2000, 2003a).

Concerns for 3D ambiguity notwithstanding, the power of computers to create stereo and motion parallax rapidly and effectively continues to grow, thereby supporting the design of *virtual environments* that capture the natural 3D properties of the world. Such environments have many uses, as discussed in Chapter 40, and their creation must again make effective use of an understanding of the human's natural perception of depth cues. Further discussion of the benefits of different kinds of 3D displays for active navigation is provided in Section 5.4.2.

### 4.5 Motion Perception and Dynamic Displays

Although many of the displays discussed above are static, many other analog displays are frequently or continuously updated. Such displays have long been a part of air and ground transportation, process control, and manufacturing systems, and our ability to process motion cues is a requirement for our interaction with objects in both natural and virtual worlds. However, as with the perception of depth and 3D shape, the visual perception of motion can fall prey to a variety of illusions and ambiguities (Blake and Sekuler, 2006). Sometimes these perceptual distortions predispose operators to accidents, for example, the alarming tendency of drivers to cross railroads when trains are drawing dangerously close. Leibowitz (1985) has argued that this behavior may, in part, be due to a size-contingent distortion of speed. Specifically, we tend to see small objects as approaching faster than large ones, perhaps causing drivers to underestimate the speed at which trains, a large object by most vehicular standards, are approaching. Fortunately, errors of motion perception can sometimes be corrected by intentionally employing some of these same illusions, as we will see below. And most of us would agree that at least one motion illusion is of particular value as it is at the very core of many forms of communication and entertainment, that is, media that involve "moving" pictures.

*Apparent motion* is our ability to perceive seamless motion of displayed objects from a series of still images—as long as the individual images are presented within certain spatial–temporal bounds. Given over a century of experience with apparent motion, many of the parameters that affect this illusion are well known among animators and videographers as well as scientists (Hochberg, 1998). For example, an object is seen as moving smoothly from one location to another when presented in two successive pictures that are separated by approximately 20 m and when the displacement of the image is no more than approximately 15° of arc. Under these circumstances, the movement of an object usually appears to take the shortest path between its positions in the two frames. Although this process is mediated by relatively early, bottom-up processing, there is evidence that when intervals between images are long enough, apparent motion appears to be influenced by top-down factors such as the plausibility of the movement (e.g., a person's hand will be seen going over rather than through her head when the two successive images used to create apparent motion show a hand first on one side and then on the other side of her head; Shiffrar and Freyd, 1990).

Of importance to a variety of simulation and gaming applications is another motion illusion—*vection*, or the perception of self-motion by a stationary operator. The illusion of self-motion can be induced by visual displays that imitate the regularities in the pattern of flow of textures across the visual field as we change speed and trajectories in the natural world (Hettinger, 2002), thus taking advantage of some of the types of scene statistics we introduced when discussing context effects on recognition. In addition to their use to create perceptions of self-motion in virtual environments, relevant scene properties can be manipulated in the real world to alter operators' perceptions of self-motion in ways that can enhance safety. In order to encourage

drivers to slow down when approaching a roadwork site, for example, researchers have found that placing a series of cones by the roadway can be helpful, especially when the spacing between the cones systematically decreases, giving drivers the illusory impression that they are accelerating (Allpress and Leland, 2010; also see Denton, 1980).

Further sources of uncertainty in motion judgments are *framework effects.* As a general rule, the visual perception of motion is greatly influenced by the framework within which it occurs. A very sparse display with few fixed reference elements, for example, can make it appear that a small, slowly moving display element is actually stationary. At the opposite end of the spectrum are situations where the target object is too big to be fully seen through the viewing framework. If there is uncertainty about the overall shape of the target object, the target may appear to move in a direction other than its true course. DeLucia et al. (2006) demonstrate the potential importance of this *aperture effect* during minimally invasive surgeries, where the view through the endoscope reveals only a very limited part of the surgical field, and the shape and location of various anatomical landmarks may be distorted by disease. The surgeon may make incorrect judgments about the location of the endoscope inside the patient because of an incorrect perception of its direction of movement based on the apparent flow of organs under the endoscope. This aperture effect can be reduced by using rectangular rather than round viewing windows, but its existence is one reason designers are pursuing the development of augmented displays that generate a computer model of the entire surgical field surrounding the immediate, detailed view from the endoscope (e.g., Wang et al., 2009).

Even when the motion of target objects in dynamic displays can be accurately perceived, operators may still have difficulty linking the motion or changes they perceive to the appropriate actions they should take. Thus, moving displays need to consider the compatibility of their changing elements with the mental models of their users (Gentner and Stevens, 1983; Norman, 1988). In the remainder of this section, we describe two general principles that can help designers match the elements of dynamic displays to users' expectations. The first of these is clearly applicable to both static and dynamic displays, although it may be more important in situations where users must quickly react to changing information. The second principle, however, deals with the meaning of motion itself in the context of the user's goals.

*Code congruence* requires that designers take into consideration users' expectations about the codes (i.e., stimulus dimensions) that should represent critical information. If a display is being designed to represent temperature, for example, then the use of a vertically moving bar might be an appropriate code because of mental models associated with mercury-based thermometers that often show temperatures "rising" and "falling." If we do choose this code, we must also be careful to map the values of the to-be-represented variable (temperature) to the display code in a way that conforms to users' experience. Thus, an increase in bar height should indicate hotter rather than colder temperatures.

In the preceding example, temperature is associated with vertical position through experience with thermometers rather than because there is anything inherently spatial about temperature. Many other variables, however, *are* spatial to begin with. In these cases, code congruence is maintained simply by using rescaled versions of the original spatial dimensions, making sure to orient the spatial model in a manner consistent with the typical viewpoint of the user. Thus, the temperature of different rooms in a building could be represented by superimposing our individual temperature displays on an actual plan view, elevation, or 3D view of the building. When the spatial metaphor is direct, as in this case, the display fulfills Roscoe's (1968) *principle of pictorial realism*.

Finally, in dynamic displays, the designer must be concerned with whether motion represented in a display is compatible with the movement expected by the operator. These expectations in turn are often driven by the *frame of reference* of display motion. As an example, a common design decision in systems involving robotics control is whether the camera generating a display of the moving robotics arm or vehicle should be mounted to the moving element itself (viewing the target to be reached) or to an external frame viewing the moving element. The difference between these two views is that when a control moves the element to the right, in the first case the display of the world slews to the left, whereas in the second case, the world is stable and the rightward control depicts rightward motion on the display. These two frames of reference are referred to respectively as (1) inside-out, moving world, or "ego referenced," and (2) outside-in, moving object, or world referenced.

There is some evidence that the second frame is better and more compatible (Johnson and Roscoe, 1972) in that it corresponds to what we naturally see when we view our hand as well as to the ecological perspective that the world is stable and objects move within it. However, it is also acknowledged that in many complex systems both views are needed, often the outside-in view for global positioning and the inside-out view for fine vernier control. What is most vital is to understand that it is possible for an operator to confuse the two if both are offered (simultaneously or sequentially). An operator viewing an inside-out view moves the control to the right with the intention of moving the object to the right; he or she suddenly sees a left movement on the display, perceives (incorrectly) it to be a control error, and quickly reverses the movement, now inadvertently creating such an error. This problem can be amplified if views are often switched between different modes.

## 4.6 Perceptual Organization, Display Organization, and Proximity Compatibility

Our discussion of motion highlighted how the *relationship* among displayed elements, in both time and space, can produce psychologically important effects such as apparent motion from still images. In this section, we look at how the relationship among component elements of multielement displays, such as those found

in emergency management information systems, scientific visualizations, and process control workstations, can influence the effectiveness of these display systems, sometimes in dramatic ways.

The challenge for the designer of multielement displays is to determine how best to organize the various displays so that the natural laws of *perceptual organization* (e.g., Pomerantz and Kubovy, 1986; Hochberg, 1998) may support rather than hinder the user's acquisition of information. That is, we must take into account the laws by which our perceptual systems parse raw sensory data into potential targets for attention (e.g., perceptual parts, objects, and configurations). As noted in Section 3, multielement displays have the potential to create a variety of problems for the viewer, including increases in search time, increased information access effort, similarity-based confusion, and challenges to focused attention. Such concerns are becoming even more critical with the rapid evolution of expansive, high-definition displays (e.g., "data walls") to represent information for multiple concurrent tasks and users (Yost and North, 2006).

Many of the principles of display organization can be accounted for by the *proximity compatibility principle* (*PCP*) (Carswell and Wickens, 1987, 1996; Barnett and Wickens, 1988; Wickens and Andre, 1990; see Wickens and Carswell, 1995, for review). This is a broad-ranging principle with two parts: (1) when two (or more) elements of a display must be integrated (e.g., compared, multiplied, subtracted) in the service of a single task—a feature we describe as "close (or high) mental proximity"—they should be placed in close physical proximity on the display. However, (2) when one element must be processed by itself, without distraction from others (distant or low mental proximity requiring focused attention), its processing is

best served when its display is more distant from its neighbors (low physical proximity). Hence there is a sort of "compatibility" between proximity in the mind and proximity on the display. Of course, what this means is that no single display layout serves all tasks, for what may be best for integration tasks ("how does my speed compare with a maximum speed allowable?") may not serve focused attention ("what is my current speed?").

The previous broad overview of the PCP is complicated by the fact that there are at least two kinds of "mental integration" (simple comparison and arithmetic combination, such as computing and amount from the product of rate and time) and a large number of ways of creating or manipulating "display proximity," as illustrated in Figure 8. The most obvious source of physical proximity is simple distance, as illustrated by rows (a) and (b) of the figure. Moving things closer together can minimize visual scanning; in a cluttered display, it can also minimize visual search for the two items to be compared, as attention must move from one to the other to compare them. From an information-processing perspective, the problem is often that information from one display must be held in vulnerable working memory as the other is sought. At a minimum, this dual-task requirement may produce a delay— a decay in the representation of the first item until the second is found. More severely, it may produce concurrent task interference (see Section 7) as an effort-consuming search must be carried out while information from the first display is rehearsed.

To cite extreme examples of low physical proximity, consider the challenge of comparing a picture on one page of a textbook to the text describing that picture on the back side of the same page. Would it not be better if text and picture were side by side so that repeated page flipping would not be needed? Instructional designers
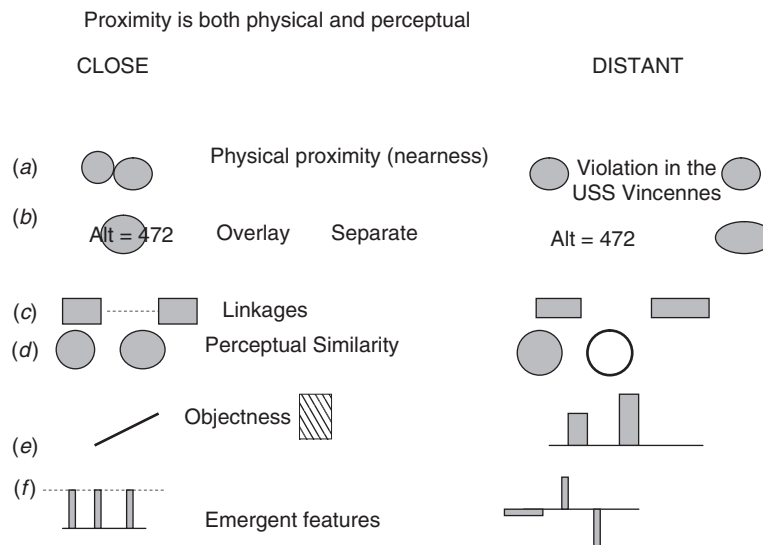


**Figure 8** Six examples of close (left side) vs. distant (right side) display proximity between information sources influencing different perceptual mechanisms

need to consider the role of cognitive load of information integration in preparing their material (Paas et al, 2003). Or consider the problem of comparing a box legend on a graph labeling the multiple graph lines with the graph itself. It would be better if each graph line had its own label attached to it (Huestegge et al, in press) In this sense, display proximity can not only be conceptualized as physical distance affecting visual scanning but also includes the manual (and sometimes cognitive) effort required to "move" attention from one source to the other.

Row (b) of Figure 8 also illustrates the occasional downside of close physical proximity for focused attention. On the right, the oval indicator and the digital altitude reading are separated. It is easy to focus on one and ignore the other. On the left, they are overlaid. The very close physical proximity can create clutter, making it hard to focus on one and ignore the other. Such overlay is often created when viewing the world beyond through a head-mounted display (C. D. Wickens et al., 2004); such clutter of close proximity also may be created by overlaying computer windows (Mori and Hayashi, 1995). Of course, the reader may ask, "Can't we get the best of both worlds by placing the two indicators adjacent, and not overlapping?" and the answer here is "yes." Adjacency will greatly reduce the costs of integration, but if there is at least $1°$ of visual angle separating an item from its neighbors, this will pretty well minimize the costs to focusing attention.

Naturally, there are some circumstances in which displays simply cannot be moved to create adjacency or close proximity. In a cluttered demographic map, for example, two cities may be located at different places whose demographics (portrayed in icons, or text boxes) cannot be "relocated." As shown in row (c), linkages can achieve such physical proximity or "connectedness" (Jolicoeur and Ingleton, 1991). Indeed, in the simple line graphs, the lines connecting the data points are not essential to understanding the trends. But those lines, linking points belonging to the same condition, greatly assist graph interpretation. (As we will see below, the lines have another benefit for integration by creating emergent features.) Thus, if rendered appropriately, often in lower intensity, linkages can increase benefits to integration without imposing costs on focused attention.

When displays cannot be moved, an alternative to linkages, as shown in row (d), is to exploit perceptual similarity; for example, two items to be compared may be highlighted in the same color (as this *word* and the word in the paragraph below; L. D. Wickens, Alexander, et al., 2004) or may both be flashed or "jittered" in synchrony. This technique can be quite advantageous for an air traffic controller (ATC) who needs to consider the trajectories of two planes on the same altitude and converging but still far apart. On an ATC display cluttered with many other aircraft symbols, a highlighting tool (Remington et al., 2000) that can illuminate each in a common color can greatly reduce the demands of such a comparison.

In row (e) is depicted a conceptually different way of creating display proximity. Here a single object (in this case a line connecting two points on a line graph) is created to foster integration, whereas the more separated display shown on the right contains two objects (two bar graphs, or even three, if the baseline is considered). Basic research in attention describes how processing of all attributes of a single object can take place more or less in parallel, whereas processing two objects requires serial processing (and less efficiency; Treisman et al., 1983; Duncan, 1984; Scholl, 2007). In the graph shown in (e), not only is the single line more perceptually efficient to process, but also the slope of the line provides a direct indicator of the strength of the effect or trend, an indication that is easier to extract than by comparing the height of the two bars in the more separated display on the right (Carswell and Wickens, 1996; Hollands and Spence, 1992). We refer to this slope as one example of an *emergent feature* of the object display, a concept elaborated below.

There are of course many other ways of creating an object than just "connecting the dots." Shown just to the right of the word "objectness" in row (e) is a rectangle. This object has four features, its height, width, color, and texture, all conveyed within the confines of a single object. It would require four bar graphs (four objects) to convey all that information in separate displays. And object displays, if carefully configured, can be shown to greatly improve the facility of information integration (Barnett and Wickens, 1988). As a simple example, consider a demographic map. If each city were represented by four bar graphs, when several cities are closely packed together, the display would appear extremely cluttered, in contrast to the case where each is represented by a single rectangle.

Finally, row (f) of the figure illustrates a concept of display proximity that can also facilitate integration. Emergent features of the multiple elements of a display are created when the display is "configured" in such a way that a feature emerges from the combination of the elements (Buttigieg and Sanderson, 1991; Pomerantz and Pristach, 1989). In the close-proximity version on the left, the bar graphs are all aligned to the same baseline. Hence the emergent feature of "colinearity" is created when all bars are the same height (e.g., all of three engines of a machine are operating at the same power setting). In this configuration, it is quite easy to perform the mental integration task: "Is the power equally distributed?" Any break from this equality will be rapidly and easily perceived. In contrast, the configuration on the right will make the precise integration judgment of equality very challenging.

Referring back to row (e), we can see how, for example, the simple connections of endpoints to form a single object as in the line graph creates an emergent feature (line slope) that directly serves the integration task (how strong is the trend from left to right?). Also, creating an object like a rectangle display creates an emergent feature from its height and width: its area (the product of height and width) as well as its shape (tall skinny, short fat), which can often carry important information (Barnett and Wickens, 1988). But it is important to realize that not all dimensions of an object will configure in some geometric or spatial pattern: For example, color and height do not form an emergent

feature (Carswell and Wickens, 1996). And of course, as row (f) illustrates, emergent features can readily be formed in the absence of the object display.

Regardless of whether emergent features are part of an object display or result from the configuration of separate objects, they can have potent effects on performance. We must caution that these effects are not always positive. Emergent features are *only* useful if they directly represent integrated variables or system states of importance to the operator. If they are irrelevant or, worse, cause patterns that cue responses inconsistent with those that are appropriate for safe and efficient system operation, the user may be better off with separable formats. Using the terms of Section 3, emergent features are salient and capture focused attention, whether wanted or not. *Representation aiding* is a display design approach in the cognitive engineering tradition (Section 2) that provides guidance on how to optimize the design of displays, including the use of emergent features (Vicente and Rasmussen, 1992; Smith et al., 2006). The focus of this approach is directly on understanding the physical constraints of dynamic systems and matching these to the geometric properties of configural formats in a way that is perceptually salient and meaningful to the operator. An example is shown in Figure 9.

Our discussion above has focused most heavily on the way in which close proximity supports information integration. But while we have identified separation (lower proximity) as a tool to support focused attention on a single element, lower proximity has a second benefit. That is, the close proximity and particularly

the "binding" in an object display represent a designer-imposed solution to encourage the user to integrate particular subsets of information sources. However, such binding or grouping is not a "free lunch" for the viewer. If the viewer's task is not appropriately characterized by the designer, or if the task demands change, the viewer may need to "unbind" one or more elements from the perceptual group to use in a new way. The extra effort and reduced performance efficiency that occur in this situation may be considered a "parsing cost." This is what we experience when a graph's designer has not organized individual data points in a way that allows us to make the comparisons that are of most interest to us, instead forcing us to compare points across different perceptual groups (e.g., points on different lines or bars in different data series).

Clearly, the specific way in which a designer uses proximity tools in any display will make some integration tasks easy while imposing parsing costs on others. However, the designer's choices can also be used to infer their communication goals (or *graphical pragmatics*). Following this logic, there have been attempts to apply the PCP to the development of software that "reverses engineers" displays. That is, the software provides inferences to the user about the designer's underlying communication goals based on the way in which proximity is applied (Elzer et al., 2003). This information may, in turn, be used by viewers to help them decide if the display is likely to suit their information needs and to help automate the textural summary of graphical information for those individuals who cannot access the information visually. More generally, the
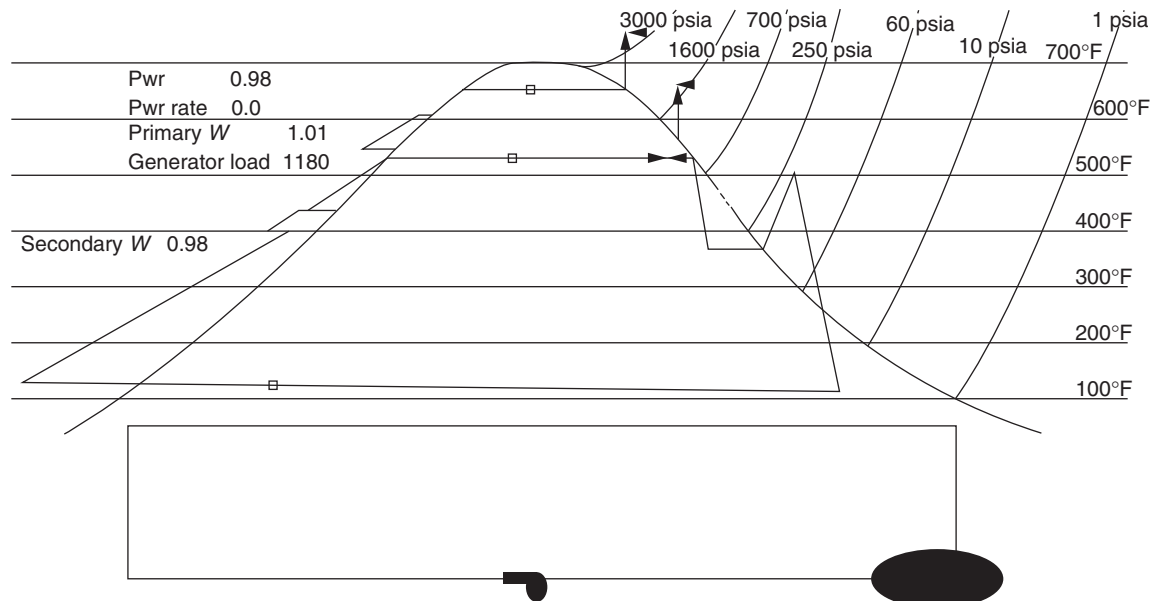


**Figure 9** Rankine cycle display for monitoring the health of a nuclear power generating plant. The jagged line indicates the trajectory of the plant parameters (steam pressure and temperature) as they follow the constraints of the thermodynamic laws (proposed but not yet implemented for operational evaluation). (From Moray et al., 1994.)

study of graphical pragmatics highlights the extent to which individual, noninteractive visualizations, no matter how carefully designed, can impose on the viewer the designer's values and assumptions about what is meaningful in the displayed data. In particular, one may argue that in scientific or interpretive endeavors one of the most critical information-processing challenges is integrating data in a variety of different ways in order to obtain unique insights. Hence the importance of flexible data visualization tools that are not constrained by a particular designer's choices (Robertson et al., 2009; North, 2006), which can configure data in different ways, encouraging the use of different aspects of proximity.

In conclusion, forming effective multielement displays can sometimes be as much a creative "art" as it is a science. But effective use of different proximity tools can greatly facilitate information integration without necessarily compromising focused attention.

## 5 COMPREHENSION AND COGNITION

In our discussion of perception and display design, we have treated many of the operator's perceptual tasks as decision-making, problem-solving, or reasoning tasks. Detection involves decisions about criterion setting. Identification involves estimations of stimulus probabilities. Size and distance judgments in 3D space involve the formulation of perceptual hypotheses. For the most part, however, these processes occur rapidly and automatically, and as a result, we are generally not aware of them. In this sense, perceptual reasoning is a far cry from the effortful, deliberate, and often time-consuming process that we are very aware of when trying to troubleshoot a malfunctioning microwave, find our way through an unfamiliar airport, understand a legal document, or choose among several product designs. Before discussing such higher order cognitive tasks, we describe the critical limits of working memory. As we will see, the parameters of working memory constrain, sometimes severely, the strategies we can deploy to understand and make choices in many types of tasks.

### 5.1 Working Memory Limitations

Working memory refers to the limited number of ideas, sounds, or images that we can maintain and manipulate mentally at any point in time. The concept has its roots in William James's (1890) *primary memory* and Atkinson and Shiffrin's (1971) *short-term store*. All three concepts share the distinction between information that is available in the conscious here and now (working, short-term, or primary memory) and information that we are not consciously aware of until it is retrieved from a more permanent storage system (long-term or secondary memory).

Unlike items in long-term memory, items in working memory are lost rapidly if no effort is made to maintain them through rehearsal (Brown, 1959; Peterson and Peterson, 1959). For example, *decay rates* of less than 20 s have been obtained for verbally delivered navigation information (Loftus et al., 1979) as well as for visuospatial radar information (Moray, 1986). However,

even when tasks require minimal delays before recall, working memory is still severely limited in its capacity. Miller (1956) suggested that this capacity, the *memory span*, is limited to about five to nine independent items. The qualifier "independent" is critical, however, because physically separate items that are stored together as a unit in long-term memory may be rehearsed and maintained in working memory as a single entity: a *chunk*. Thus, for most people, the letter string "H–T–E" contains three items to remember, whereas the rearranged string "T–H–E" contains only one.

Baddeley (2003) has integrated information about the limits described above with evidence from neuropsychological and psychometric investigations to develop a four-part model of working memory. First, there are two temporary storage systems, the *phonological loop* and *visuospatial sketchpad*. These subsystems are used by a *central executive* that manipulates information from these stores. The central executive also integrates information from the component storage systems with long-term memory to create more complex, multimodal representations of coherent objects. These integrated representations, in turn, are held in an *episodic buffer*. It should be noted that the proposed episodic buffer is a relatively recent addition to Baddeley's model (2000), which was originally conceptualized as having only three components (Baddeley and Hitch, 1974). Although relatively little is known about the limits of the episodic buffer compared to those of the phonological loop and the visuospatial sketchpad, the notion of a repository for objects of interest that is not limited to a single processing code is relevant to our discussions of virtually all higher order cognitive tasks.

Most research on the limits of working memory have focused on the *phonological loop*, so named because it is associated with our silent repetition or rehearsal of words, letters, and numbers. The phonological loop stores a limited number of sounds for a short period of time; thus, the number of items that can be held in working memory is related to the length of time it takes to pronounce each item (Gathercole and Baddeley, 1993; Gathercole, 1997). This implies that our memory span is slightly lower for words with many syllables.

The *visuospatial sketchpad* holds visual and spatial information as well as visualizations of information acquired verbally (Logie, 1995; Baddeley, 2003). The information held in the sketchpad may be in the form of mental images, and as with the phonological loop, the contents will be lost rapidly if not rehearsed. Research suggests that rehearsal in the visuospatial sketchpad involves repeated switching of selective attention to different positions across these images (Awh et al., 1998). The proposed maintenance function of attention switching may explain why our control of eye movement apparently disrupts some contents of the sketchpad (Postle et al., 2006).

The central executive is aptly named because its functions can be compared to those of a business executive. The central executive's role is not to store information but to coordinate the use of information. This information may come from the phonological loop, the visuospatial sketchpad, or the episodic buffer. The

central executive is presumed to be involved in integrating information from these different stores, and it is also involved with selecting information, suppressing irrelevant information, coordinating behavior, and planning (see Section 5.5 for the relevance of these functions for problem solving). It is important to note that there are limits on how many of these operations the central executive can execute at one time.

The concept of working memory has a number of implications for design. We describe some of these implications below, and in subsequent sections we discuss other implications for higher order tasks such as decision making, problem solving, and creative thinking.

1. The capacity of working memory's short-term storage systems are easily exceeded, resulting in a loss of information that may be necessary to perform an ongoing task (consider rehearsing a 10-digit phone number and area code). The design implication is to avoid, whenever possible, codes that infringe on the limits of these systems.

2. When it is necessary to use codes that exceed the limits of working memory capacity, there are several ways to reduce memory loss. For example, parsing material into three- or four-item units may increase chunking and subsequent recall (Wickelgren, 1964). Thus, 3546773 is more difficult to recall than 354–6773. In addition, information for different tasks may be split between storage systems so that a single system, for example, the visuospatial sketchpad, is not overwhelmed. More will be said about such interventions when we turn to the discussion of multitask performance (see Section 7). Finally, designers should prefer easily pronounced verbal codes. For example, numerical codes that make frequent use of the two-syllable number "seven" will be more prone to loss from the phonological loop than codes that make frequent use of other numbers. It may also suggest that the functional number span for some languages may be larger than those for others.

3. Information from working memory may be lost if there are delays longer than a few seconds between receiving the information and using it. Thus, systems should not be designed so that the user must perform several operations before being able to perform a "memory dump." For example, voice menu systems should always allow users to select a menu option as soon as it is presented rather than forcing them to wait until all the options have been read to make their choice. Methods of responding should be simplified as well, so that users do not have to retain their choice for long periods of time while trying to figure out how to execute it. One aspect of the proximity compatibility principle (Section 4.6) emphasizes working memory limitations on the need to seek a second source of information to be integrated while rehearsing the contents of the first source.

4. The need to scan should be minimized if a person must hold spatial information in the sketchpad. Thus, a first responder who has found the likely location of a victim in cluttered, unfamiliar, and distorted terrain should not have to scan long lists of information to find appropriate communication codes.

5. Avoid the need to transfer information from one subsystem into the other before further transformations or integrations can be made. This reduces the resources available for the primary processing goal. Wickens et al. (1983, 1984) have provided evidence that the display format should be matched to the working memory subsystem that is used to perform the task. Specifically, visual-analog displays are most compatible with tasks utilizing the visuospatial sketchpad (e.g., air traffic controllers' maintenance of a model of the spatial relations among aircraft) and auditory–verbal displays are most compatible with tasks utilizing the phonological loop (e.g., a nurse keeping track of which medications to administer to a patient).

6. If working memory subsystems are updated too rapidly, old information may interfere with the new. For alphanumerical information, Loftus et al. (1979) found that a 10-s delay was necessary before information from the last message no longer interfered with the recall of the current material. As we will see below, when such updating occurs nearly continually, the capacity of working memory is greatly reduced to around its "$7 - 2 = 5$" value.

7. Interference in working memory is most likely when to-be-remembered information is similar in either meaning or sound. Thus, an air traffic controller might have particular difficulties remembering a series of aircraft with similar call signs (UAL 235, UAL 325). Interference will also be greater if there is similarity between material to be remembered and other competing tasks (i.e., listening, speaking) (Banbury et al., 2001).

8. The capacity of working memory varies between people and has been associated with differences in the fluency of the central executive and hence with success in multitasking and general intelligence (Engle, 2002; Engle et al., 1999).

## 5.2 Dynamic Working Memory, Keeping Track, and Situation Awareness

Much of the research devoted to working memory has examined tasks in which information is delivered in discrete batches and the goal is to remember as much of the information as possible. However, there are many other tasks in which the operator must deal with continuous information updates with little expectation of perfect retention. Moray (1981) studied several *running memory tasks* that simulated the demands of a more continuous input stream, and he found the typical

memory span to be less than five chunks. In some cases it was difficult for subjects to keep track of items more than two places back in the queue. Yntema (1963) demonstrated that the way information is organized has a direct impact on supervisors' abilities to keep track of values of multiple attributes of several objects (e.g., status and descriptions of several aircraft). Supervisors had greater success keeping track of a few objects that varied on many different (and discriminable) attributes than in keeping track of variation in a few attributes for many objects. In the former case there are fewer opportunities for confusion than in the latter case, and confusion is a major source of disruption in working memory (Hess and Detweiler, 1995).

This earlier research on running memory anticipates the rapid growth of interest over the last two decades in *situation awareness* (SA) or, colloquially, our understanding and use of information about "what's happening" during dynamic tasks (Wickens, 2008; Tenney and Pew, 2006; Banbury and Treselian, 2004; Endsley and Garland, 2000; Durso et al., 2007; see also Chapter 19). Endsley (1995) provides a more formal definition of SA, one that has been adopted by many current researchers: SA is "...the perception of the elements of the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future." This definition suggests that SA has three stages or "levels": (1) perception or "noticing," (2) understanding or comprehending, and (3) projecting or prediction. These three levels of situation awareness can be tied directly to different aspects of information processing and, therefore, failures at different levels will require different types of training or design interventions.

Stage 1 SA, noticing, traces directly to issues of selective attention and attentional capture, discussed in Section 2. Indeed, Jones and Endsley (1996) found that a majority of aircraft accidents attributable to loss of SA were related to breakdowns at this first stage. Likewise, Durso et al. (2007) found failures of stage 1 SA to be responsible for many SA-related problems encountered in air traffic control. This is not surprising given our previous discussion of how easily we can fail to notice significant changes in dynamic systems (e.g., the "change blindness" phenomenon; see Section 3.1). In general, failures of stage 1 SA typically indicate the need for interventions involving display design, especially the use of alerts and attentional cueing. However, because sampling of information in dynamic environments also involves long-term memory in the form of knowledge about "where and when to look," training interventions may also be considered (e.g., Hoffman et al., 1998).

Stage 2 SA, understanding the implications of events noticed in stage 1, depends heavily on the limits of working memory as they apply to keeping track of evolving situations (e.g., the pilot asks: "Where was that traffic aircraft the last time that I looked?"). The episodic buffer of working memory proposed by Baddeley (2003), with its connections to long-term memory, may be necessary to explain the ability of skilled performers to hold more information for longer periods than would be expected on the basis of the decay rates established

for other working memory stores. Likewise, Ericsson and Kintsch (1995) have used the concept of *long-term working memory* with its automatic activation of relevant knowledge structures in long-term memory to explain the unusual ability of experts to maintain and integrate relatively large amounts of information across time. For operators who may lack extensive experience, and even for experienced operators who may be interrupted by other tasks, displays of system history may also be helpful in understanding the implications of current events in light of prior observations (St. John et al., 2005).

The stage 3 component, prediction and projection, is perhaps the most complex and may depend to a greater extent than the other two stages on the expertise and training of the operator. Accurate prediction of an evolving situation certainly depends on current perception and understanding (stages 1 and 2), but it also requires a well-calibrated *mental model* of the dynamic process under supervision (Gentner and Stevens, 1983; Wilson and Rutherford, 1989), a mental model that can be "played" in response to the current data, in order to predict the future state. For example, an air traffic controller with a good mental model of aircraft flight characteristics can examine the display of the current state and turn the rate of an aircraft and project when (and whether) that aircraft will intersect a desired approach course, thereby attaining a satisfactory separation. A well-calibrated mental model resides in long-term memory, but to play the model with the current data requires perception of those data as well as the active cognitive operations carried out in working memory. In some cases, prediction can be approximated by using an expert acquired *script* of the way a typical situation unfolds. However, unless active processing (stage 1) of incoming perceptual information is carried out, there is a danger that projection will be based totally on expectancies of typical situations and that unusual or atypical events will be overlooked. Naturally, stage 3 SA can benefit greatly from accurate predictive displays (Wickens et al., 2000).

It should be noted, finally, that situation awareness is a construct that is resident within the perceptual–cognitive operations of the brain. It is not itself a part of the action (other than the actions chosen to acquire new information).

## 5.3 Text Processing and Language Comprehension

Comprehension of language, whether written or spoken, shares many of the processes described for situation awareness. Noticing relevant information, understanding its implications, and to varying degrees projecting the content of upcoming messages are all part of the active process of language comprehension. The constraints relevant to information processing at each of these stages helps determine why we find some conversations, lectures, journal articles, instructions, and warnings easier to understand than others.

Of course, factors influencing the detectability and discriminability of the individual speech sounds

(phonemes) and written symbols (letters) will limit the extent to which language can be processed meaningfully. However, recall that easily comprehended phrases or sentences can also influence the detectability of the individual words. See Section 3.2 for a discussion of the effect of context on identification. We are typically able to understand sentences such as the last one despite the absence of many of the correct letters in the words, because of the profound effects of context, expectancies, and the redundancies inherent in language. It is worth noting that just as context can help us recognize familiar words; it can also help us understand the meanings of words that we have never encountered before (Sternberg and Powell, 1983).

As our discussion of context suggests, the comprehensibility of text depends on many factors, from the reader's experience, knowledge, and mental models that drive expectations to the structuring of text so as to make maximum use of these expectancies. It is not surprising, then, that *readability metrics* that attempt to estimate the difficulty of text passages, generally based on average word and sentence length, are not altogether satisfactory. Although it may be true that longer words are generally less familiar and longer sentences place greater demands on our working memory capacities, many other factors influence comprehensibility. Kintsch and Vipond (1979), for example, used traditional readability indices to compare the speeches of candidates in the 1952 presidential campaign. Eisenhower's speeches were generally reputed to be simpler than those of Stevenson, yet formal readability indices indicated that Stevenson used shorter words and sentences. This contradiction between public opinion and the formal metrics corresponds to our experience that some sentences with a few short words can still be very confusing. We now discuss some additional factors that determine comprehensibility and have implications for message design.

Kintsch and colleagues (e.g., Kintsch and Keenan, 1973; Kintsch and Van Dijk, 1978) argue that the complexity of a sentence is actually determined by the number of underlying ideas, or *propositions*, that it contains rather than by the number of words. Although a few specific words may be carried forward in working memory for brief periods, it is the underlying propositions that are used to relate information in different phrases and sentences. Just as Moray (1981) estimates that running memory carries forward less than five chunks of information, Kintsch and Van Dijk (1978) estimate that only four propositions can be held in working memory at one time. There are some exceptions to this general rule, as when a highly skilled reader reads text on a very familiar topic. As we saw in our discussion of situation awareness, the memory effects of such expertise have led some to argue for the existence of a long-term working memory in which long-term memory associations are automatically activated and used during ongoing comprehension at little additional processing cost. However, as a general rule, readers must be very selective in their choices of propositions to retain, usually favoring the most recent propositions and those they believe to be most central to the overall text message.

Problems arise in comprehension when newly encountered propositions cannot easily be related to the propositions active in working memory. Such problems often occur when readers attempt to integrate information across sentence boundaries. Consider, for example, the following sentences:

1.  When the battery is weak, a light will appear.
2.  You will see it at the top of the display panel.

Readers must make the *bridging inference* that the second sentence is telling them where to look for the light rather than where to find the battery. This inference, in turn, depends on their general knowledge of displays: Lights rather than batteries tend to be visible on display panels. A second type of integration failure occurs when a concept introduced earlier in the text is not actually used again until some sentences, paragraphs, or pages later. In fact, even relatively minor delays, such as the need to scroll in order to see additional text on a web page, can lead to comprehension decrements, especially in readers with smaller working memory capacities (Sanchez and Wiley, 2009). Such challenges to text comprehension are often explained by the need to conduct a *reinstatement search*, which requires an effortful search of long-term memory or a rereading of earlier text, in order to clarify the meaning of a proposition in current working memory.

One general goal in striving for comprehensibility is to avoid the need to make bridging inferences or perform reinstatement searches. However, it is clearly impossible to remove the need to make some inferences, and it is probably undesirable given that such elaborations may make the information more memorable. One goal of the text designer is simply to assist the reader in making the *appropriate* inferences. One important way that this can be done is by providing adequate context immediately prior to the presentation of target information (McKoon and Ratcliff, 1992). Because inferences draw on the reader's knowledge of particular topics, it is useful to allow the reader to access the relevant knowledge structures in long-term memory at the outset. Bransford and Johnson (1972) provide a powerful demonstration of the importance of providing context in the form of pictures or descriptive titles presented just prior to textual material. A series of instructions on how to wash clothes was presented with and without the prior context of a title, "washing clothes." When the title was removed, the reduction in readers' abilities to understand and recall the instructions was dramatic.

Other factors that increase the processing demands of verbal material include the use of negations and lack of congruence between word orders and logical orders. With regard to negations, research indicates that it takes longer to verify a sentence such as "the circle is not above the star" compared to "the star is above the circle" (Clark and Chase, 1972; Carpenter and Just, 1975). Results suggest further that the delay is due to something other than the time necessary to process an additional word (i.e., "not"). Instead, it appears that listeners or readers first form a representation of the objects in the sentence based on the order of presentation (e.g.,

circle-before-star in the sentence "the circle is not above the star." However, to make their mental representation congruent with the meaning of the negation, they must perform a transformation of orders (i.e., to end up with a circle that is *not* before/above the star). Similar logic is used to explain why we have trouble processing statements in which there is a mismatch between the underlying logical order and the actual, physical order of the words (DeSoto et al., 1965). Returning to our battery of instructions once again, the underlying causal sequence assumed by most people would be that a weak battery would trigger a warning light. To be consistent with this causal order, it would be better to state that "If the battery is weak, the light will come on" rather than "If the light comes on, the battery is weak."

Finally, the physical parsing of sentences on a page, sign, or computer screen can also influence the comprehensibility of verbal messages. Just and Carpenter (1987) have argued that readers pause and integrate propositions at the end of each phrase in a sentence. This idea may explain why comprehension of sentences that must be split over multiple lines is better when the end of each line of text corresponds to the end of a constituent phrase (Graf and Torrey, 1966). Thus, instructions or warnings that must appear on several different lines (or as a few words on several successive screens) should be divided by phrases rather than, for example, on the basis of the number of letters. "Watch your step . . . when exiting . . . the bus" will be understood more quickly than "Watch your . . . step when . . . exiting the bus."

## 5.4   Spatial Awareness and Navigation

Language comprehension sometimes taxes working memory, particularly the phonological rehearsal loop and central executive. However, as we saw when discussing problems with negation, people may use text to generate representations of spatial relations. This spatial facet of text and language comprehension has been particularly prominent in recent discussion of the "situation models" that we develop when reading or listening to a story (e.g., knowing where in a room all the characters are sitting). We now turn to a task that relies more heavily, for many people, on the capacity limits of the visuospatial sketchpad (Logie, 1995) or, more generally, spatial working memory and spatial cognition (Shah and Iyiri, 2005)—navigating through our worlds, both real (finding our way through a maze of looping suburban streets and cul-de-sacs; Taylor et al., 2008) and virtual (searching a complex computer-displayed multidimensional database; Wickens et al., 2005).

### 5.4.1   Geographical Knowledge

Thorndyke (1980) has studied the knowledge that people use when finding their way about. Of particular interest is Thorndyke's claim that increased familiarity with an area causes changes in more than the amount of detail contained in our mental representation of that area stored in long-term memory. In addition, the actual *type* of mental representation (analog versus verbal/symbolic), as well as its frame of reference, may evolve

in a predictable way. After an initial encounter with a city, neighborhood, or building, we may develop *landmark knowledge*. If told that his or her destination is beside the "telephone tower," a person with landmark knowledge will scan the environment visually until spotting something that appears to be the tower and will then strike off in its direction. Thus, the newcomer has the knowledge necessary to recognize the landmark but has no knowledge about its location. For the person with landmark knowledge alone, wayfinding would be impossible if the landmarks were obscured. This problem has become commonplace as once-salient landmarks have become obscured by new and often taller structures. The problem for urban planners, then, is to ensure that landmarks (both natural and designed) remain easily visible and distinctive in order to serve their navigational function for years to come.

With more experience traveling about an area, we typically develop an ordered series of steps that will get us from one location to another. These sets of directions, called *route knowledge*, tend to be verbal in nature, stated as a series of left–right turns (e.g., "Go left on Woodland until you get to the fire station. Then take a left . . . "). Navigation along these routes may be rapid and very automatic; however, limited knowledge of the higher order relations among different routes and landmarks still limits navigational decision making, making it difficult, for example, to figure out shortcuts and particularly difficult to recover when lost. With still more extensive wayfinding experience, or with specific map study, *survey knowledge* may be acquired. Survey knowledge is an integrated representation of the various routes and landmarks that preserves their spatial distance relations. This analog representation is often referred to as a *cognitive map*.

The type of representation—route versus survey—that best supports performance in various wayfinding tasks, like so many other aspects of mental (and display) representation, depends on the nature of the task or problem. Thorndyke and Hayes-Roth (1982) compared *route training* (actual practice navigating between specific points in a large building) to *survey training* (study of the building map). Route training appeared to facilitate people's estimates of route distance and orientation, while survey training appeared to facilitate judgments of absolute (Euclidean) distance and object localization.

### 5.4.2   Navigational Aids

Although we can often navigate through environments on the basis of our acquired knowledge stored in long-term memory, whether route, survey, or even landmark, there are many other circumstances in which we require displayed *navigational aids* which are perceived. These aids may take on a wide variety of forms, ranging in the degree to which guidance to a target is supported: from tightly guided flight directors in aircraft and turn signs on highways to route lists to electronic maps that highlight one's current position to simple paper maps. Furthermore, electronic maps can vary in the extent to which they rotate so that the direction of travel is "up" on the map, and both electronic and paper maps can vary

in terms of whether they present the world in planar or 3D perspective view as the latter is seen in many in-car navigation displays (see Section 4.4).

To understand which forms of maps support the best spatial information processing to accomplish navigation, it is important to consider briefly the stages involved in this process. The navigator must engage in some form of visual search of both the navigational aid (to locate the final destination, intermediate goals, and current location) and the environment or a displayed representation thereof (to locate landmarks that establish the current location and orientation). The navigator must then establish the extent to which the former and the latter are congruent, determining the extent to which "where I am" (located and oriented) agrees with the intermediate goal of "where I want to be." Finally, the traveler must choose an action (e.g., turn right) to move from the current location toward the goal. Establishing this congruence as well as choosing the action may require any number of different *cognitive transformations* that add both time and effort to the navigational task (Aretz, 1991; Hickox and Wickens, 1999; Gugerty and Brooks, 2001; Wickens et al., 2005, 2010).

An example of two of these transformations is represented in Figure 10, which represents the information processing of a pilot flying south through an environment depicted on a north-up contour map. To establish navigational congruence, the pilot must rotate the map mentally to a track-up orientation and then envision the contour representation of the 3D terrain to determine its congruence with the forward view. Both of these information transformations are effortful and time consuming and provide sources for error. In particular, those sources involved with mental rotation of maps have been well documented (Levine, 1982; Eley, 1988; Warren et al., 1990; Aretz, 1991; Olmos et al., 1997; Gugerty and Brooks, 2001; Macedo et al., 1998.

Different transformations may be required when other navigational aids than the 2D map are provided. For example, verbal descriptions of landmarks will also require some transformations to evaluate against their visible 3D spatial counterparts. Transformations may also be required to "zoom in" to a large-scale map (Kosslyn, 1987) in order to establish its congruence with a close-in view of a small part of the environment. Modeled in terms of processing operations such as visual search and spatial transformations, one can then determine the form of navigational aids that would be of benefit for certain tasks. For example, electronic maps are beneficial if they highlight the navigator's current location, thus obviating visual search of the map. Highlighting landmarks on the map, which are salient in the visual world, will correspondingly reduce search.
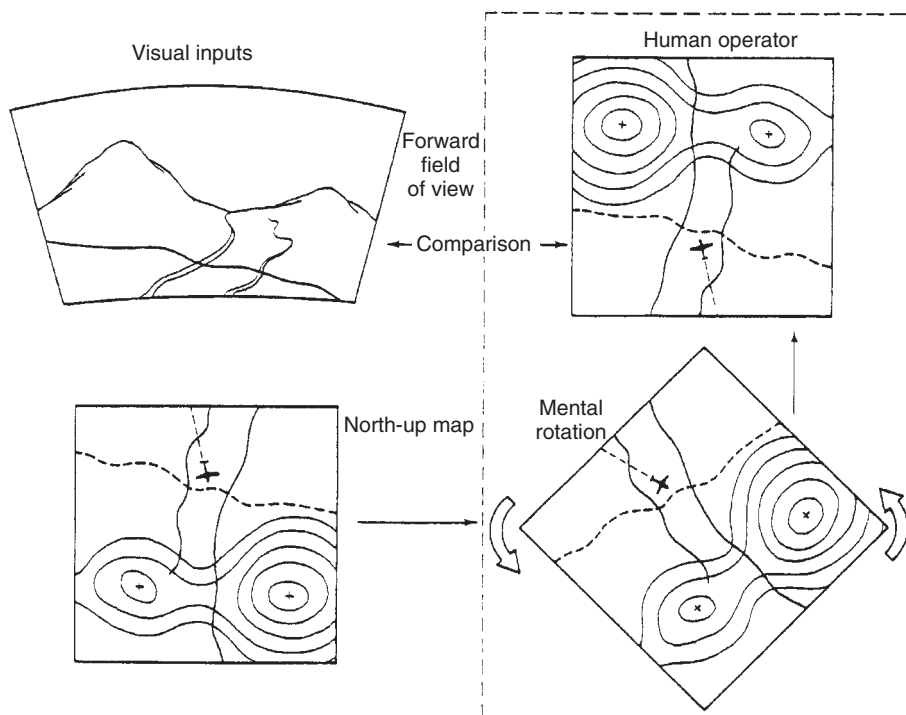


**Figure 10** Mental rotation required to compare the image seen in an ego-referenced forward field of view (top) with a world-referenced north-up map (below) when the aircraft is heading south. The map image is mentally rotated (right) to bring it into lateral congruence with the forward field of view. It is then envisioned in three dimensions to compare with the forward field of view.

Rotating maps in a track-up orientation will help navigation by eliminating mental rotation (Aretz, 1991; Olmos et al., 1997; Wickens, 1999). Presenting guidance information in a 3D format (Section 4.4), like one would see looking ahead into the environment itself, will also reduce the magnitude of any sort of transformations and considerably improve navigational performance (Wickens and Prevett, 1995; Wickens et al., 2005). The benefits of a 3D view will be enhanced if the viewpoint of the display corresponds to the same zoom-in viewpoint as that occupied by the navigator, looking forward, rather than a viewpoint that is behind and from the outside (Wickens and Prevett, 1995; Prinzel and Wickens, 2008–2009). These viewpoint relationships are shown in Figure 11, which depicts the viewpoint location (top) and the view seen by a pilot (bottom row) in an immersed or *egocentric view* (a and b). (These two views differ in terms of their geometric field of view.) Panel (c) represents an external or exocentric view. Panel (d) represents a 2D coplanar view, which was discussed in Section 4.4.

Expressing navigational guidance in terms of command *route lists* (e.g., "turn left at X; go three blocks until Y") will also eliminate the need for many of the spatial cognitive transformations that may be imposed when spatial maps are used, since the language of command is thereby expressed directly in the language of action. Such congruence can account for the benefits of route lists over spatial maps in certain ground navigation tasks (Streeter et al., 1985). A second advantage to such route lists is that they can be presented verbally and represented in working memory mentally in a phonetic or verbal code, thus reducing competition for the spatial processing resources involved in many aspects of environmental scanning and vehicle navigation (Section 7). The verbal descriptions inherent in route lists are well suited for some navigational environments, particularly human-designed environments (cities) with objects that are easily labeled, have distinct unambiguous landmarks,

and can easily be discriminated by counting (the "fourth house"). In many naturalistic environments, however, where features are defined by continuous, not categorical, properties and may be confusable, route lists are more problematic and should be coupled with redundant spatial or pictorial guidance.

The most direct levels of navigational guidance that eliminate most or all levels of mental transformations (e.g., a flight director, the 3D forward-looking display shown in Fiigure 11a, or a verbal route list) will provide for effective navigation while en route. However, such displays may do a disservice to the navigator who suddenly finds himself lost, disoriented, or required to make a spontaneous departure from the planned route. Also, those features that make a navigational display best for guidance will harm its effectiveness to support the spatial situation awareness (Section 5.2) that is necessary for a successful recovery from a state of geographical disorientation (Wickens and Prevett, 1995; Wickens, 1999). This is an important trade-off between the immersed 3D view of Fiigure 11a, which is good for guidance, but because of its "keyhole view" of the world, it is poor for maintaining global situation awareness, a task better supported by the exocentric view of Fiigure 11c. Finally, we note that the immersed 3D view makes a poor tool for route planning, an activity that we turn to in the following section.

## 5.5 Planning and Problem Solving

Our previous discussion has focused on cognitive activities that were heavily and directly driven by information in the environment (e.g., text, maps, or material to be retained in working memory). In contrast, the information-processing tasks of planning and problem solving are tied much less directly to perceptual processing and are more critically dependent on the interplay between information available in (and retrieved from) long-term memory and information-processing transformations carried out in working memory.
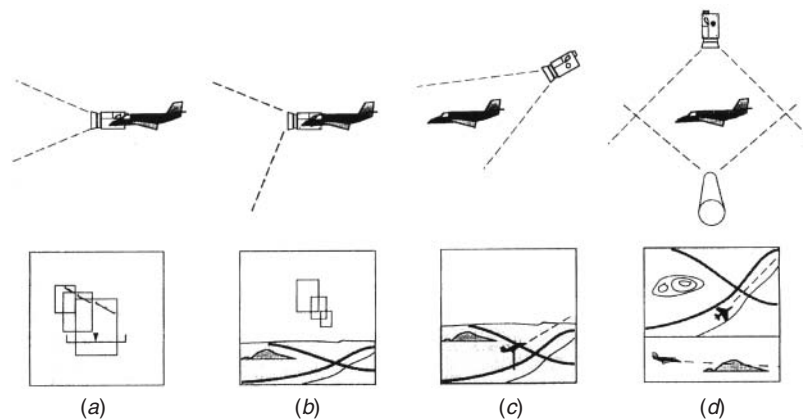


**Figure 11** Display viewpoints in an aircraft display that require varying degrees of transformations to compare with a pilot's direct view forward from the cockpit, looking at a virtual "highway in the sky" The lower figures illustrate schematically what would be seen by the pilot with the viewpoint shown above. The transformation in (a) is minimal; in (b) and (c), modest; and in (d), large. Views (a) and (b), however, reduce more global situation awareness.

### 5.5.1 Planning

The key to successful operation in many endeavors (Miller et al., 1960) is to develop a good plan of action. When such a plan is formulated, steps toward the goal can be taken smoothly without extensive pauses between subgoals. Furthermore, developing contingency plans will allow selection of alternative courses of actions should primary plans fail. As an example, pilots are habitually reminded to have contingency flight plans available should the planned route to a destination become unavailable because of bad weather.

Planning can typically depend on either of two types of cognitive operations (or a blend of the two). Planners may depend on scripts (Schank and Abelson, 1977) of typical sequences of operations that they have stored in long-term memory on the basis of past experience. In essence, one's plan is either identical to or involves some minor variation on the sequence of operations that one has carried out many times previously. Alternatively, planning may involve a greater degree of guess work, and some level of mental simulation of the intended future activities (Klein and Crandall, 1995; see Chapter 37). For example, in planning how to attack a particular problem, one might play a series of "what-if" games, imagining the consequences of action, based again on some degree of past experience. Hence a surgeon, in planning how to manage a potential future operation, might mentally simulate the future body conditions and reactions of the patient under different proposed surgical procedures to see if the intended commands would resolve the conflict and would stay clear of other aircraft.

Consideration of human performance issues and some amount of experimental data reveals three characteristics of planning activities. First, they place fairly heavy demands on working memory, particularly as plans become less script based and more simulation based. Hence, planning is a task that is vulnerable to competing demands from other tasks. Under high-workload conditions, planning is often the first task to be dropped, and operators become less proactive and more reactive (Hart and Wickens, 1990). The absence of planning is often a source of poor decision making (Orasanu, 1993; Orasanu and Fischer, 1997). Second, perhaps because of the high-working-memory demands of planning, in many complex settings, people's *planning horizon* tends to be fairly short, working no more than one or two subgoals into the future (Tulga and Sheridan, 1980). To some extent, however, this characteristic may be considered as a reasonably adaptive one in an uncertain world, since many of the contingency plans for a long time horizon in the future would never need to be carried out and hence are probably not worth the workload cost of their formulation. Finally, given the dependency of script-based planning on long-term memory, many aspects of planning may be biased by the *availability heuristic* (Tversky and Kahneman, 1974; Schwarz and Vaughan, 2002), discussed in more detail in Chapter 8. That is, one's plans may be biased in favor of trajectories that have been tried with success in the past and therefore are easily recalled.

Consideration of such vulnerabilities leads inescapably to the conclusion that human planning is a cognitive information-processing activity that can benefit from automated assistance, and indeed, such planning aids have been well received in the past for activities such as flight route planning (Layton et al., 1994) and industrial scheduling (Sanderson, 1989). Such automated planners provide assistance that need not necessarily replace the cognitive processes of the human operator but merely provide redundant assistance to those processes in allowing the operator to keep track of plausible courses of future action.

### 5.5.2 Problem Solving, Diagnosis, and Troubleshooting

The three cognitive activities of problem solving, diagnosis, and troubleshooting all have similar connotations, although there are some distinctions between them. All have in common the characteristic that there is a goal to be obtained by the human operator; that actions, information, or knowledge necessary to achieve that goal is currently missing; and that some physical action or mental operation must be taken to seek these entities (Mayer, 1983; Levine, 1988). To the extent that these actions are not easy or not entirely self-evident, the processes are more demanding.

Like planning, the actual cognitive processes underlying the diagnostic troubleshooting activities can involve some mixture of two extreme approaches. On the one hand, situations can sometimes be diagnosed (or solutions to a problem reached) by a direct match between the features of the problem observed and patterns experienced previously and stored in long-term memory. Such a *pattern-matching technique*, analogous to the role of scripts in planning, can be carried out rapidly, with little cognitive activity, and is often highly accurate (Rasmussen, 1981). This is a pattern of behavior often seen in the study of *naturalistic decision making* (Zsambok and Klein, 1997; Kahneman and Klein, 2009; see Chapter 8).

At the other extreme, when solving complex and novel problems that one has never experienced before, a series of diagnostic tests must often be performed, their outcomes considered, and based on these outcomes, new tests or actions taken, until the existing state of the world is identified (diagnosis) or the problem is solved. Such an iterative procedure is typical in medical diagnosis (Shalin and Bertram, in press). The updating of belief in the state of the world, on the basis of the test outcomes, may or may not approach prescriptions offered by guidelines for *optimal information integration*, such as Bayes's theorem (Yates, 1990; see Chapter 8).

In between these two extremes are hybrid approaches that depend to varying degrees on information already stored in long-term memory on the basis of experience. For example, the sequence of administering tests (and the procedures for doing so) may be well learned in long-term memory even if the outcome of such tests is unpredictable and must be retained or aggregated in working memory. Furthermore, the sequence and procedures may be supported by (and therefore directly perceived from) external *checklists*, relieving cognitive demands still further. The tests themselves might be physical tests, such as the blood tests carried out by

medical personnel, or they may involve the same mental simulation of what-if scenarios that was described in the context of planning (Klein et al., 1993).

As with issues of planning, so also with diagnosis and problem solving, there are three characteristics of human cognition that affect the efficiency and accuracy of such processes. First, as these processes become more involved with mental simulation and less with more automatic pattern matching, their cognitive resource demands grow and their vulnerability to interference from other competing tasks increases in a corresponding fashion (see also Chapter 9). Second, as we noted, past experience, reflected in the contents of long-term memory, can often provide a benefit for rapid and accurate diagnosis or problem solutions. But at the same time, such experience can occasionally be hazardous, by trapping the troubleshooter to consider only the most *available* hypotheses: often those that have been experienced recently or frequently and hence are well represented in long-term memory (Tversky and Kahneman, 1974; Schwarz and Vaughn, 2002). In problem solving, this dependence on familiar solutions in long-term memory has sometimes been described as *functional fixedness* (Adamson, 1952; Levine, 1988).

Third, the diagnostic/troubleshooting process is often thwarted by a phenomenon referred to alternatively by such terms as *confirmation bias* and *cognitive tunneling* (Levine, 1988; Nickerson, 1998; Woods et al., 1994; Wickens and Hollands, 2000). These terms describe a state in which the troubleshooter tentatively formulates one hypothesis of the true state of affairs (or the best way to solve a problem) and then continues excessively on that track even when it is no longer warranted. This may be done by actively seeking only evidence to confirm that the hypothesis chosen is correct (the confirmation bias) or simply by ignoring competing and plausible hypotheses (cognitive tunneling).

Collectively, then, the joint cognitive processes of planning and problem solving (or troubleshooting), depending as they do on the interplay between working memory and long-term memory, reflect both the strengths and the weaknesses of human information processing. The output of each process is typically a decision: to undertake a particular course of action, to follow a plan, to choose a treatment based on the diagnosis, or to formulate a solution to the problem. The cognitive processes involved in such decision making are discussed extensively in Chapter 8, as are some of the important biases and heuristics in diagnosis discussed more briefly above.

### 5.5.3 Creativity

In general, creativity involves human problem solving that is relatively free from the confirmation bias, cognitive tunneling, and functional fixedness, each of which restricts the number of problem solutions we consider. For most theorists, creativity refers to the production of *effective novelty* (Cropley, 1999; Mayer, 1999). This is a process that involves thinking of a variety of previously untried solutions *and* judging their probable effectiveness. Finke et al. (1992) argue that *generating*

novel cognitive structures involves retrieving, associating, synthesizing, and transforming information, while *evaluating* novel structures involves inferring, hypothesis testing, and context shifting, among other strategies. It is clear from this analysis that the cognitive load imposed by creative tasks can be immense and that working memory, including the storage systems and the central executive, will be taxed.

Novelty production may be particularly difficult to maintain for long periods of time for at least two reasons. First, the cognitive load imposed by creative problem solving, as we have described above, is high from the outset. Second, because novel stimuli often increase arousal levels, it is likely that the *production* of novelty will create a cycle of upward-spiraling arousal in the problem solver. This, in turn, will cause some degree of cognitive tunneling, making continued novelty production and evaluation difficult (Cropley, 1999). This may suggest that, unlike some other tasks, where higher levels of arousal may be desirable to maintain performance (e.g., long-duration search tasks for low-probability targets), creativity may be fostered by low initial levels of arousal.

The idea that novelty production may cause spiraling levels of arousal also provides one explanation for the often-discussed benefits of *incubation* for creative problem solving. Smith (1995) describes incubation in terms of the general finding that people are more likely to solve a problem after taking a break rather than working on a solution without interruption. In controlled trials, incubation effects are not invariably found (Nickerson, 1999); however, research continues to focus on the conditions under which incubation works. It is possible that a break from the act of novelty generation may serve to reduce arousal levels to more task-appropriate levels. Another explanation is that the probability of a new *problem representation* being put into action (e.g., the mental image or list of procedural steps being manipulated to generate solutions) is greater when a person disrupts his or her own processing. The person may simply be more likely to have forgotten components of a previous, ineffective representation upon returning to the task.

The importance of the cognitive representation of problems, and the different display formats that support these representations, has been demonstrated for a variety of problem-solving tasks (Davidson and Sternberg, 1998). Flexible scientific, information, and design visualization tools may prove to be particularly valuable for creative problem solving, because changing the orientation, color scheme, format, or level of focus will change the salience of different aspects of the problem. For example, when designers were asked to generate a design for a new lamp, Damle (2010) found that the use of design software that permitted monochromatic viewing of the otherwise multicolored designs helped the designers avoid fixating prematurely on design details. Presumably, this relatively simple change in the design software influenced designers' self-evaluations of their evolving designs, shifting their attention to global characteristics such as symmetry and

balance. Developing software tools that reduce problems like functional fixedness, that instead encourage the perception of different aspects of the problem, are an important focus of current work on the design of *creativity support tools* (Schneiderman, 2009).

## 5.6 Metacognition and Change Blindness Blindness

We end our discussion of higher order cognitive processes by discussing a type of knowledge that may have a profound impact on the successful performance of any task, but especially on tasks involving problem solving, comprehension, and the maintenance of situation awareness. The term *metacognition* was introduced by Flavell (1979) to indicate a person's knowledge about his or her own cognitive processes and, further, the use of this information to regulate performance (Reder, 1996). The most active area of research on metacognition has been in education, where researchers have looked at how students' beliefs about their own information-processing capabilities influence learning strategies and ultimate academic success (Veenman et al., 2006; Bjork, 1999). However, the concept has important implications for the practice of human factors and ergonomics as well.

When describing metacognition, most researchers distinguish between *metacognitive knowledge* and *metacognitive control processes*. In general, metacognitive knowledge includes beliefs about one's own processing capacity, about potential strategies that enhance performance (or minimize capacity limits), and about when and why such strategies are appropriate (Schraw and Moshman, 1995). For example, an expert operator may come to the conclusion that it is more difficult to notice critical signals when they occur in a particular region of a visual display. Further, the operator may believe that he or she can compensate by oversampling that region but may also believe that oversampling will come at the cost of much greater mental effort. In contrast, Levine et al. (2000) have noted that most people are unaware of their strong tendencies toward change blindness, a metacognitive phenomenon referred to as "change blindness blindness." Clearly, such beliefs can influence the strategies an operator chooses. And it is important to realize that these beliefs may not be accurate and, further, that individuals, especially experts, may sometimes be unaware of the assumptions they are making (Chi et al., 1988).

Metacognitive control processes include planning, monitoring, and evaluating one's own performance (Schraw and Moshman, 1995). We might associate many of these processes with the functions of Baddeley's (2003) central executive in working memory. Planning includes determining the appropriate allocation of attention and time to different parts of a task as well as decisions about what aspects of performance to sacrifice if capacity limits create mandatory trade-offs (e.g., would a fast but inaccurate strategy be better than an accurate but slow one?). Monitoring involves keeping track of the quality of performance as a task progresses, for example, our ability to know whether or not we comprehend some written instructions well enough to act on them; or for the learner, whether we have studied enough on one lesson to have mastered the material and move on to the next. Finally, evaluation looks at the products of task performance and the control processes used to obtain them. This allows an assessment of how effectively we deployed our limited cognitive resources and may further elaborate our metacognitive knowledge in long-term memory.

Metacognitive knowledge and control processes are important for understanding human performance in a variety of domains. One important application is to determine how metacognitive knowledge may shape users' preferences for and use of specific interface and product designs. Users' choices may sometimes seem less than optimal based on existing performance data or predictions of formal performance models (Andre and Wickens, 1995). Cases in point are users' "intuitive" judgments of product usability. Payne (1995; see also Vu and Proctor, 2003), for example, explored users' judgments about the compatibility of different arrangements of multielement displays and controls. These judgments were not accurate predictors of actual performance using the different display–control arrangements, and the naïve judgments seemed to be based on the average "goodness" of the individual matches between each display and its associated control rather than on the match between the configuration of displays and controls. In short, the participants undervalued (or, perhaps, did not understand) the importance of configural properties on memory and performance.

Smallman and St. John (2005) propose that many users and designers alike fall prey to naïve realism, a tendency to prefer more realistic-looking displays, even when simpler formats would support better performance. In general, these authors attribute users' preferences to faulty metacognitive knowledge. For example, most individuals seem unaware of the inherent ambiguities of size and depth found in more realistic-looking 3D displays (see Section 4.4). They were also unaware of the relatively small portion of complex displays that they could process in a single glance, and they seemed not to consider the visual search costs that could result from increased display complexity.

Other examples of the impact of both metacogntive knowledge and control processes are found in drivers' overconfidence in their driving abilities in a variety of situations. For example, nighttime driving, as well as the use of devices such as cell phones, selectively degrades focal (central) vision which is critical for detecting hazards. Peripheral vision, which provides sufficient information for lane keeping, is relatively unimpaired by such factors. However, drivers may not appreciate the distinctive functions of each visual subsystem and may take satisfactory lane keeping as evidence that they have suffered no impairment from nighttime viewing conditions or in-vehicle distractions. This leads to dangerous overconfidence in their ability to detect and identify hazards, a task using a very different visual system and resources (see Section 7.2). In this case, metacognitive monitoring and performance evaluation are impaired, in part because the drivers have continuous feedback about lane keeping but only infrequent feedback about hazard detection. In addition, the

driver's knowledge about their own perceptual systems may not lead them to suspect that one visual subsystem can be spared while another is degraded. Similar sorts of overconfidence may also be implicated when operators fail to use automation or decision support tools in circumstance where their own performance is deficient without these aids (e.g., McGuirl and Sarter, 2006).

When correctly diagnosed, metacognitive errors can be addressed in a variety of ways. Training is one approach, for example, providing information about the types of displays that best support different types of tasks (Milkman et al., 2009). Shen et al. (in press) found that the failure of users to select the appropriate perspective views to use when performing different emergency management tasks could be partially corrected with very simple verbal guidance about the best use of 2D versus 3D displays. Alternatively, poor strategy choices (or poor use of technology ) might be addressed by providing better performance feedback to help users recalibrate their perceptions of their own skill levels, knowledge, and capabilities and to allow them to learn which performance strategies are most effective.

In this section we have discussed information-processing tasks that are time consuming, require extensive cognitive resources, and for which the "correct" response is poorly defined and multiple responses are possible, even desirable. We turn now to characteristics of actions that are typically selected rapidly, sometimes without much effort, and often without great uncertainty about their outcome.

## 6  ACTION SELECTION

### 6.1  Information and Uncertainty

In earlier sections we discussed different stages at which humans process information about the environment. When we turn to the stage of action selection and execution, a key concern addresses the *speed* with which information is processed from perception to action. How fast, for example, will the driver react to the unexpected roadway hazard or pedestrian or how rapidly will the help desk access computer information in response to a client's question. Borrowing from terminology in communications, we describe information-processing speed in terms of the *bandwidth*, the amount of information processed per unit time. In this regard, a unit of information is defined as a *bit*. One bit can be thought of as specifying between one of two possible alternatives, two bits as one of four alternatives, three bits as one of eight, or, in general, the number of bits (conveyed by an event) $= \log_2 N$, where $N$ is the number of possible environmental events that could occur in the relevant task confronting the operator. In the following pages, after we describe a taxonomy of human actions, we will see how information influences the bandwidth of human processing.

The speed with which people perform a particular action depends jointly on the uncertainty associated with the outcome of that action and the skill of the operator in the task at hand. Rasmussen (1986; Rasmussen et al., 1995) has defined a behavior-level continuum

that characterizes three levels of action selection and execution that is characterized by both uncertainty and skill. *Knowledge-based behavior* describes the action selection of the unskilled operator or of the skilled operator working in a highly complex environment facing a good deal of uncertainty. In the first case, we might consider a vehicle driver trying to figure out how to navigate through an unfamiliar city; in the second case, we consider the nuclear reactor operator trying to diagnose an apparent system failure. This is the sort of behavior discussed in Section 5.5.

*Rule-based behavior* typically characterizes actions that are selected more rapidly based on certain well-known rules. These rules map environmental characteristics (and task goals) to actions, and their outcomes are fairly predictable: "If the conditions exist, do $X$, then $y$, then $z$." The operator response in executing rule-based behavior is fairly rapid but is still "thought through" and may be carried out within the order of a few seconds. Working memory is required. Finally, *skilled-based behavior* is very rapid and nearly automatic in the sense that little working memory is required, performance of concurrent tasks is possible, and the action may be initiated within less than a second of the triggering event. Skill-based behavior, for example, characterizes movement of the fingers to a key to type a letter, the sequence of steering wheel turns used to back out of a familiar driveway or compensate for a wind gust, or the response of the pilot to an emergency ground proximity warning that says "pull up, pull up."

Human factors designers are quite interested in the system variables that affect the speed and accuracy of behavior of all three classes. Typically, those variables affecting knowledge-based behavior are discussed within the realm of problem solving and decision making (see Section 7 and Chapter 8). We discuss below the variables that influence rule- and skill-based behavior [see Wickens and Hollands (2000) for a more detailed discussion].

### 6.2  Complexity of Choice

Response times for either rule- or skill-based behavior become longer if there are more possible choices that could be made and therefore more information transmitted per choice (Hick, 1952; Hyman, 1953). The rule-based decision to go left or right at a Y fork in the road is simpler (i.e., 1 bit) and made more rapidly than at an intersection where there are four alternative paths (i.e., 2 bits). Menu selections take longer on a page where there are more menu options, and each stroke on a typewriter (26-letter options) takes longer to initiate than each depression of a Morse code key (two options). Indeed, the time to select an option is roughly proportional to the number of bits in the choice (Hick, 1952). As a guideline, designers should not give users more choices of action than are essential, particularly if time is critical. Long menus, with lots of rarely chosen options, may not be desirable. The consequences of offering many choices are not only longer response time but also an increased possibility that the wrong option will be chosen by mistake. More items typically lead to greater similarity between items and hence an invitation for confusion.
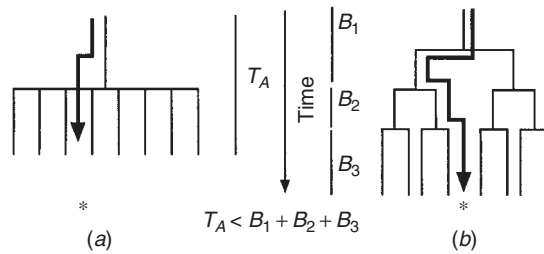
**Figure 12** Decision complexity advantage: (*a*) total time required for three "simple" (low-complexity) choices; (*b*) time required for a single high-complexity choice. The total amount of information transmitted is the same in both cases.

The guidance for avoiding very complex choices presented above does not necessarily mean that very simple choices (e.g., 1 bit per choice) are necessarily best. Indeed, generally *an operator can transmit more total information per unit time with a few complex* (*information-rich*) *choices than several simple* (*information-poor*) *choices*. This conclusion, referred to as the *decision complexity advantage* (Wickens and Hollands, 2000), can be illustrated by two examples: First, an option provided by a single computer menu with eight alternatives (one complex decision) can be selected faster than an option provided by three consecutive selections from three two-item menus (three simple decisions; see Figure 12). Second, voice input, in which each possible word is a choice from a potentially large vocabulary (high complexity), can transmit more information per unit time than typing, with each letter indicating 1 of only 26 letters (less complex); and typing in turn can transmit more information per unit time than can Morse code. The general conclusion of the decision complexity advantage drawn from these examples and from other studies (Wickens and Hollands, 2000) points to the advantage of incorporating keys or output options that can select from a larger number of possible options, such as special service "macro" keys, keys that represent common words, or "chording" keyboard devices (Baber, 1997) that allow a single action selection (a chord depression using several fingers simultaneously) to select from one of several options.

In conclusion, it may seem that two contradictory messages were offered in the paragraphs above: (1) Keep choices simply, but (2) use a small number of complex choices. In resolving these two guidelines in design, it is best to think that the first guideline pertains to not providing a lot of *rarely* used options, particularly in time-stressed situations and when errors of choice can have high-risk consequences. The second guideline pertains to how to structure a choice among a large number of remaining and *plausible* options. A single choice among a larger list is often better than multiple sequential choices among smaller lists.

### 6.3 Probability and Expectancy

People respond more slowly (and are more likely to respond erroneously) to signals and events that they do not expect. Generally, such events are unexpected (or surprising) because they occur with a low probability in a particular context. This is consistent with our discussion of context effects on object identification in Section 4.2. Low-probability events, such as events with a greater number of alternatives, are also said to *convey more information*. The information in bits, conveyed by a single event that occurs with probability $P$, is $\log_2(1/P)$. As we noted above, greater information content requires more time for processing (Fitts and Peterson, 1964). For example, system failures usually occur rarely and as such are often responded to slowly or inappropriately. A similar status may characterize a driver's response to the unexpected appearance of a pedestrian on a freeway or to a traffic light that changes sooner than expected. The maximum expected response times to truly unexpected events provide important guidance to traffic safety engineers in determining issues related to speed limits and roadway characteristics (Evans, 1991; Summala, 2000). Often more serious than the slower response to the unexpected event is the potential failure to detect that event altogether (see Section 4.1). It is for this reason that designers ensure that annunciators of rare events are made salient and obtrusive or redundant (to the extent that the rare event is also one that is important for the operator's task; see Section 3).

### 6.4 Practice

Practice has two benefits to action selection. First, practice can move knowledge-based behavior into the domain of rule-based behavior and sometimes move rule-based actions into the domain of skill-based ones. The novice pilot may need to think about what action to take when a stall warning sounds, whereas the expert will respond automatically and instinctively. In this sense, practice increases both speed and accuracy. Second, practice will provide the operator with a sense of expectancy that is more closely calibrated with the actual probabilities and frequencies of events in the real world. Hence, frequent events will be responded to more rapidly by the expert; but ironically, expertise may lead to *less* speedy processing of the rare event than would be the case for the novice, for which the rare event is not perceived as unexpected.

### 6.5 Spatial Compatibility

The compatibility between a display and its associated control has two components that influence the speed and accuracy of the control response. One relates to the *location* of the control relative to the display, the second to how the display reflects (or commands) control *movement*. In its most general form, the *principle of location compatibility* dictates that the location of a control should correspond to the location of a display. There are several ways of describing this correspondence. Most directly, this correspondence is satisfied by the *principle of colocation*, which dictates that each display should be located adjacent to its appropriate control. But this is not always possible in systems when the displays themselves may be closely grouped (e.g., closely clustered on a display panel) or

may not be reached easily by the operator because of other constraints (e.g., common visibility needed by a large group of operators on a group-viewed display or positioning the control for a display cursor on a head-mounted display).

When colocation cannot be maintained, the spatial compatibility *principle of congruence* takes over, which states that the spatial arrangement of a set of two or more displays should be congruent with the arrangement of their controls. One example of congruence is that left controls should be associated with left displays and right associated with right. In this regard, the distinction between "left" and "right" in designing for compatibility can be expressed either in relative terms (indicator A is to the left of indicator B) or in absolute terms relative to some prominent axis. This axis may be the body midline (i.e., distinguishing left hand from right hand) or it may be a prominent visual axis of symmetry in the system, such as that bisecting the cockpit on a twin-seat airplane design. When left–right congruence is violated such that a left display is matched to a right response, the operator may have a tendency to activate the incorrect control, particularly in times of stress (Fitts and Posner, 1967).

Sometimes an array of controls is to be associated with an array of displays (e.g., four-engine indicators). Here, congruence can be maintained (or violated) in several ways. Compatibility will best be maintained if the control and display arrays are parallel. It will be reduced if they are orthogonal (Figure 13; i.e., a vertical display array with a horizontal left–right or fore–aft control array). But even where there is orthogonality, compatibility can be improved by adhering to two guidelines: (1) The left end of a horizontal array should map to the near end of a fore–aft array (Figure 13b) and (2) the particular display (control) at the end of one array should map to the control (display) at the end of the other array to which it is closest (Andre and Wickens, 1990). It should be noted in closing, however, that the association of the top (or bottom) of a vertical array with the right (or left) level of a horizontal array is not a strong one. Therefore, ordered compatibility effects with orthogonal arrays will not be strong if one of those arrays is vertical (Chan and Hoffmann, 2010). Hence, some augmenting cue should be used to make sure that the association between the appropriate ends of the two arrays is clearly articulated (e.g., a common color code on both, or a painted line between them; Osborne and Ellingstad, 1987).

The movement aspect of SR compatibility may be defined as *intention–response–stimulus* (*IRS*) *compatibility*. This characterizes a situation in which the operator formulates an intention to do something (e.g., increase, activate, set, turn something on, adjust a variable). Given that intention, the operator makes a response or an adjustment. Given that response, some stimulus is (or should be) displayed as feedback from what has been done (Norman, 1988). There is a set of rules for this kind of mapping between an intention to respond, a response, and the display signal. The rules are based on the idea that people generally have a conception of how a quantity is ordered in space. As we noted in Section 4, when we think about something increasing, such as temperature, we think about a movement of a display that is upward (or from left to
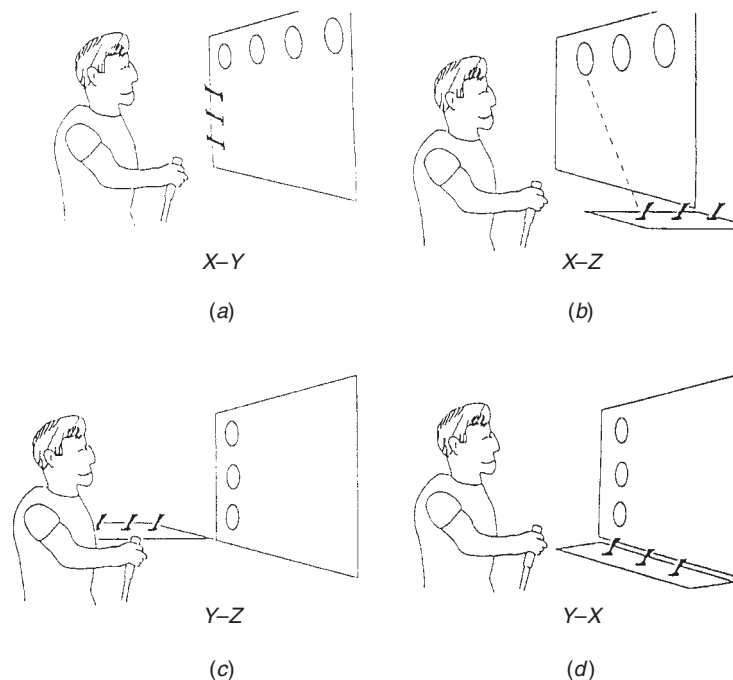


<center>
X–Y

(a)

X–Z

(b)

Y–Z

(c)

Y–X

(d)
</center>

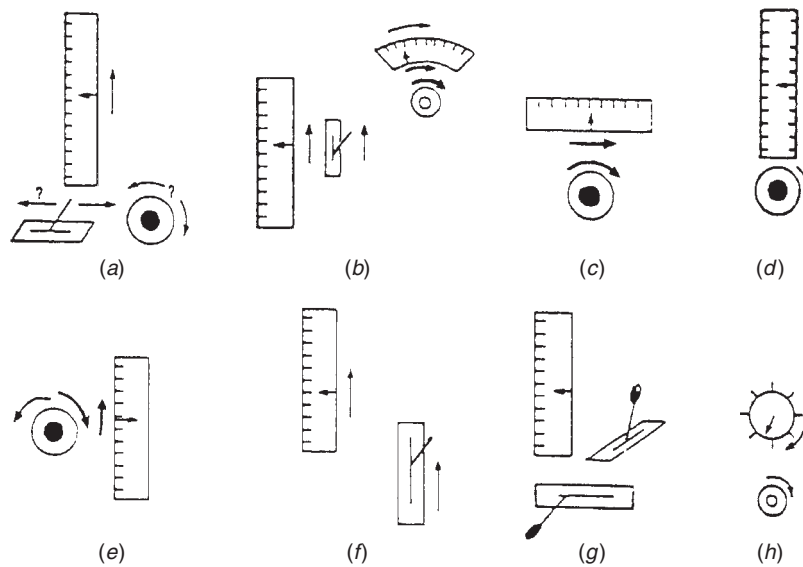**Figure 13**  Different possible orthogonal display–control configurations

**Figure 14** Examples of population stereotypes in control–display relations. (From Wickens, 1984.)

right, or clockwise). Both control and display movement should then be congruent in form and direction with this ordering. These guidelines are shown in Figure 14. Whenever one is dealing, for example, with a rotary control, people have certain expectations (a mental model) about how the movement of that control will be associated with the corresponding movement of a display. These expectancies may be defined as *stereotypes*, and there are three important stereotypes.

The first stereotype is the *clockwise increase stereotype*: A clockwise rotation of a control or display signals an increasing quantity (Figures 14*c* and *d*). The *proximity of movement stereotype* says that with any rotary control the arc of the rotating element that is closest to the moving display is assumed to move in the same direction as that display. In panel (*c*) of Figure 15, rotating the control clockwise is assumed to move the needle to the right, while rotating it counterclockwise is assumed to move the needle to the left (Chan and Hoffmann, 2010). It is as if the human's mental model is one that assumes that there is a mechanical linkage between the rotating object and the moving element, even though that mechanical linkage may not really be there.

Designers may sometimes develop control display relations that conform to one principle and violate another. Panel (*e*) shows a moving vertical-scale display with a rotating indicator. If the operator wants to increase the quantity, he or she rotates the dial clockwise. That will move the needle on the vertical scale up, thus violating the proximity-of-movement stereotype. The conflict may be resolved by putting the rotary control on the right side rather than the left side of a display. We have now created a display–control relationship that conforms to *both* the proximity-of-movement stereotype *and* the clockwise-to-increase stereotype.

The third stereotype of movement compatibility relates to *global congruence*. Just as with location compatibility, movement compatibility is preserved when controls and displays move in a congruent fashion: linear controls parallel to linear displays [(*f*), but not (*g*)] and rotary controls congruent with rotary displays [(*b*) and (*h*)]. Note, however, that (*h*) violates proximity of movement. When displays and controls move in orthogonal directions, as in (*g*), the movement relation between
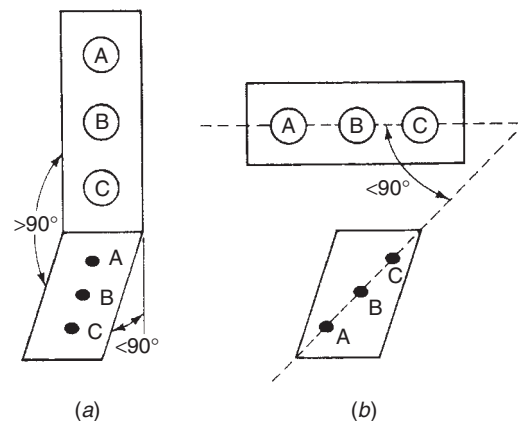


**Figure 15** Solutions of location compatibility problems by using cant. (*a*) The control panel slopes downward slightly (an angle greater than 90°), so that control A is clearly above B and B is above C, just as they are in the display array. (*b*) The controls are slightly angled from left to right across the panel, creating a left–right ordering that is congruent with the display array. (From Wickens and Hollands, 2000.)

them is ambiguous. Such ambiguity, however, can often be reduced by placing a modest "cant" on either the control or display surface, so that some component of the movement axes are parallel, as shown in Figure 15.

## 6.6   Modality

Skilled responses in most human–machine systems are typically executed by either the hands or the voice. With increasingly sophisticated automated voice recognition systems, the latter option is becoming progressively more feasible. Although the particulars of voice control are addressed in more detail in Chapter 24, at least three characteristics of voice control are relevant here in the context of information processing:

1.   Voice options allow more possible responses to be given in a shorter period of time without imposing added time-consuming finger movement components (i.e., keys), although this requires more sophisticated software in the voice recognition algorithms. Providing more options, enabling more complex decisions to be selected, is a positive benefit because it exploits the decision complexity advantage, as we saw in Section 6.2.

2.   Voice options represent more compatible ways of transmitting symbolic or verbal information than are possible with spatially guided manual options (Wickens et al., 1984), including sequential keypresses. In contrast, voice responses make relatively poor candidates for transmitting continuous analog–spatial information, particularly in dynamic situations (e.g., tracking; Wickens et al., 1985), since spoken vocabulary is better equipped to generate categorical commands (e.g., "left," "right") than continuously modulated closed-loop commands (e.g., "a little more to the left").

3.   Voice options are valuable in environments when the eyes, and in particular the hands, are otherwise engaged; but, conversely, voice options can be problematic in environments in which a large amount of other verbal activity is required, either by the user or by other people in the nearby workspace. The former causes competition for processing resources within the operator (see Section 7.2.3), while the latter creates the possibility of confusion on the part of the voice recognizer.

## 6.7   Response Discriminability

Whenever a set of manual responses are specified, any increases in the similarity between them (decreases in discriminability) will increase the likelihood of confusion. Thus, movement of the control stick to either one of two forward positions is a response choice that has greater opportunity for confusion than movement in either a forward *or* backward direction. Correspondingly, two buttons that look alike are more confusable (and hence error prone) than are two that are differently colored or shape coded. Although making controls physically distinct from each other may sometimes

destroy a sense of aesthetics in design, such distinctions will generally lead to improved human reliability (Norman, 1988). Incidentally, increased similarity between voice control options (Section 6.6) will also produce the same increase in error likelihood, although here the mediating agent is the voice recognition agent (whether human or computer) rather than the human responder. Thus, the vocabulary selected for use in an application should be chosen with a mind to avoiding confusable-sounding articulations, such as "to" and "through."

## 6.8   Feedback

The quality of feedback provided by control manipulation (or action expression) is often critical to the speed of information transmission (Norman, 1988). Indeed, sometimes the problems of poor response discriminability discussed in Section 6.7 can be addressed and at least partially remedied by providing clear, salient, and immediate feedback as to which (of several confusable) response alternatives has been chosen. This feedback may be in the form of a visible light or an auditory or tactile "click" as the control reaches its appropriate destination.

It turns out, however, that salient feedback is not always necessary or even desirable. In particular, expert or highly skilled users rely far less on feedback than do novices (e.g., the skilled typist, when transcribing, rarely looks at the keyboard or the screen). Thus, if the feedback is salient (and hence intrusive), it may be distracting to the expert, even as it is valuable for the novice. This will be particularly true whenever the feedback is *delayed*, a quality that is especially disruptive for relatively continuous tasks such as data transcription or voice translation (Smith, 1962).

## 6.9   Continuous Control

Our discussion in Section 6.8 focused on the selection of *discrete* actions, such as a keypress or lever movement. Equally important are the *continuous* movements of some controls to reach targets in space. These movements may refer, for example, to the movement of the hand to a point on a touch screen, the movement of a cursor to an icon or word on a computer screen, or the movement of a pointer to a set point on a meter. Generically, then, we can speak of these skills involving movement of a *cursor* to a *target*.

To an even greater extent than the discrete movements discussed in Section 6.8, performance of these continuous-movement skills depends greatly on *visual feedback*, depicting the difference between the current cursor location and the desired target. Performance on control tasks in which a cursor is moved a certain distance into a target is well described by *Fitts's law* (Fitts, 1966; Jagacinski and Flach, 2003):

$$\text{Movement time} = a + b \log_2(2D/W)$$

where $a$ and $b$ are constants, $D$ is the distance to the target, and $W$ is the target width. This very robust law can accurately predict the movement of all sorts of devices, from microscopic pointers (Langolf et al., 1976) to cursor movements by mice (Card, 1981) to

foot movement around a set of pedals (Drury, 1975) to the manipulation of endoscopic surgical instruments (Zheng et al., 2003). The basis of Fitts's law lies in the processing of visual feedback such that movement toward a target is maintained at a rate that is inversely proportional to the momentary distance of the target from the cursor (or other controlled object).

Just as Fitts's law nicely describes continuous movement toward a static target, it can also characterize movement of a cursor toward a *continuously* moving target, a process typically described as *tracking* (Wickens, 1986; Jagacinski and Flach, 2003). When operators engage in continuous tracking, however, whether keeping a car in the center of the highway, flying an airplane down a glide path, or moving one's viewpoint through a virtual environment via some control device, interest is more focused on minimizing the deviation from the target than on the time required to reach the target. Also, concern is less with the amplitude of the required movement than it is with other variables, such as the frequency with which corrections must be made (the *input signal bandwidth*), the complexity and lag of the *system dynamics* mediating between hand movement and cursor or output movement, and the manner in which feedback is displayed. These issues extend well beyond the scope of the current chapter and are covered in more depth in Chapter 5. We also note that compatibility effects in continuous control are also addressed in Section 4.5.

## 6.10 Errors

The previous discussion has focused primarily on the *time* required to process and respond to various items of information. Yet in many systems the occurrence of *errors* is more critical than the occurrence of delays in processing. That is, the loss of information, rather than its transmission delay, is the factor of greatest concern. Although errors are treated extensively in Chapter 27, we wish here to highlight the manner in which different classes of errors can be categorized in the context of the flow of information as depicted in Figure 1*a* (Norman, 1981; Reason, 1990, 1997, 2008).

First, *mistakes* represent errors of the earlier stages of information processing, in which incorrect action is carried out as a result of a failure to *understand* the nature of a situation (i.e., a failure of stage 2 or stage 3 situation awareness, as discussed in Section 5.2). This may result from a breakdown in perception or working memory or from insufficient knowledge to interpret the available cues (i.e., *knowledge-based errors*). Second, while a situation may be diagnosed and understood correctly, *rule-based errors* may result from a failure to apply the correct rules appropriately for selection of a response (Reason, 1990). Third, errors may result from *slips* of action, when the correct response is intended but an incorrect action is actually released (i.e., an unintended response "slips" out of the hands or mouth) (Norman, 1981). Slips of this sort are typically the result of poor human factors design, such as incompatible control–display relationships (see Section 6.5), confusable displays (Section 1) or controls (Section 6.7), coupled with an operator who is well skilled and performing a task in a highly automated

mode, thereby not carefully monitoring his or her own action selections.

A particular version of the slip is called the *mode error*, often observed in multimodal systems, where the operator forgets that the system is in one mode, thinking it is in another one, and executes a series of actions which have a very different effect on the system than intended. A trivial case is when a typist, without using visual feedback, is unaware that the keyboard is in the caps mode. In safety-critical systems, however, mode errors can have major consequences, as, for example, when a speed control can be set to a digital value which controls kilometers per hour or meters per second, depending on the mode setting. Clearly, such multimode systems must be accompanied by salient feedback of the existing mode.

Errors can also be attributed directly to a breakdown of *memory*. As noted in Section 2, working memory breakdowns may lead to forgetting or confusion of material, whereas *errors of prospective memory* may lead operators to forget to perform some action that was previously intended (Loukopopolous et al., 2009). Often described as *errors of omission*, these are typified by leaving the last copied paper on the glass of a photocopier or failing to tighten the bolts after completing a maintenance task (Reason, 1997).

It is usually the case that the conditions that are associated with slower processing are those that also produce more errors, and hence design remediation based on measures of processing speed will be productive in improving overall system accuracy. However, in certain circumstances, a strategic adjustment in how an operator performs a task will lead to an inverse relationship between speed and accuracy: the *speed–accuracy tradeoff*. In this case, a "set" to respond rapidly will lead to more rather than fewer errors (Drury, 1994). An example here would be that of the effects of time stress in emergencies, which may lead to hasty but error-prone actions in the processing of information.

## 7 MULTIPLE-TASK PERFORMANCE

Many task environments require operators to process information from more than one source and to perform more than one task at a time (Damos, 1991; Loukopopolous et al., 2009; Wickens and McCarley, 2008; Salvucci and Taatgen, 2011; Regan et al., 2009). Such environments are as diverse as that confronting the secretary conversing with the supervisor while typing, the maintenance technician who performs and observes diagnostic tests while keeping active hypotheses about possible faults rehearsed in working memory, the vehicle driver placing a cell phone call while searching for a road sign and steering (Regan et al., 2009; Collet et al., 2010), or the basketball point guard dribbling, while scanning the defense, and looking for the cutting forward.

In such multiple-task environments requiring divided attention, we may distinguish between three qualitatively different *modes* of multiple-task behavior: *perfect parallel processing*, in which two (or more) tasks are performed concurrently as well as either is performed

alone; *degraded concurrent processing*, in which both tasks are performed concurrently but one or both suffers relative to its single task level; and *strict serial processing*, during which only one task is performed at a time. Each of these modes is observed under different circumstances and has somewhat different implications for design.

## 7.1 Serial Processing and Interruption Management

The concerns of serial processing result when performance of one task or the other is delayed undesirably because of sequential constraints. Such a delay might characterize the behavior of a pilot who fails to check the aircraft altimeter sufficiently often because he is engaged in other visual tasks, leading to dangerous "altitude busts" (Raby and Wickens, 1994; Dismukes, 2001). Typically, the interests of human factors in sequential task performance are in modeling the decision process whereby the operator chooses to perform one task (and, by necessity, neglects another) at any given moment in time. This choice process is often modeled by *queuing theory* (Kleinman and Pattipati, 1991; Meyer and Kieras, 1997; Liu et al., 2006) or variants thereof (Moray, 1986; Wickens et al., 2003), which specify when a task should be sampled (performed) as a function of that task's importance (cost of not performing it) and the frequency with which it should be carried out. When evaluated against these optimal benchmarks, human performance appears to be reasonably optimal subject to the constraints of working memory.

However, reasonably optimal is not the same thing as perfectly optimal, and others have focused interest on the occasional breakdowns in optimality that do occur. Thus, a different approach to human multiple-task performance is to focus on the accidents and incidents that have apparently resulted from failures of effective task management (Raby and Wickens, 1994; Chou et al., 1996; Loukopopolous et al., 2009; Schutte and Trujillo, 1996; Orasanu and Fischer, 1997; Wickens, 2003b); that is, what causes people to neglect a task.

Here the answers based on empirical research are not entirely clear, although two prominent factors do appear to emerge. First, visible, and in particular audible, reminders to do a task increase the likelihood that that task will be done, compared to circumstances in which task initiation must be based on prospective memory alone (Norman, 1988; Dismukes and Nowinski, 2007). The vulnerability of such memory highlights the value of checklists as visual reminders for people to carry out certain actions at certain times (Degani and Wiener, 1993; Herrmann et al., 1999; Wickens, 2003b). Second, heavy involvement (high workload) with one task may lead an operator to neglect a second task and perhaps fail to return to an activity at a time when that return should be critical. Such high workload can amplify the negative effects of change blindness discussed in Section 3.1, given that such environmental changes often announce a task to which attention should be redirected. This deficiency may be addressed through task or workload management training programs (Loukopopolous et al., 2009).

Of course, task switching is a two-way street. The desirable properties of having the waiting task call attention to itself may also have the undesirable properties of interrupting an ongoing task, to the detriment of the latter (McFarlane and Latorella, 2002). Recently, study has focused on the concept of *interruption management.* This domain of interruption management embodies the research on situations which represent an operator performing an ongoing task (OT), being interrupted by an interrupting task (IT) and then returning to the OT (Trafton and Monk, 2007; Iani and Wickens, 2007; Latorella, 1996). This sequence (OT–IT–OT) can be thought of as a "unit of interruption." If the cycle is repeated many times, it represents the more general paradigm of task switching (Monsell, 2003), or task interleaving (Wickens and McCarley, 2008). Because interruptions and task switching seem to be an inevitable part of the modern workplace (Gonzales and Mark, 2004; Wolf et al., 2006), we ask if there are any general rules, tendencies, or factors that make interruption management more or less fluid.

One approach to the study of IM has focused on the first attention switch, from the OT → IT, and the extent to which *preemption* of the OT by the IT is optimal or not. IT modality certainly makes a difference here, as auditory and tactile alerts are more preferred than visual, but the two nonvisual modalities can sometimes be sufficiently intrusive that they can lead to abandoning the OT at nonoptimal times; auditory interruptions, like a phone call, are often hard to ignore. A promising approach is one in which the IT event can signal its own degree of importance, so that the performer can establish whether a switch need be immediate (Ho et al., 2004) before the OT is abandoned.

A second approach, related to *cognitive tunneling*, has been to focus on properties of an OT that will prevent the switch when in fact it should have occurred. This is the study of "engagement." For example, studies of cell phone use in cars suggest that the more "engaging" is the conversational task, the less likely is the driver (or simulated driver) to switch attention from the task to address an unexpected event in the driving task (Horrey and Wickens, 2006; Collet et al., 2010). Failure or fault management in complex systems (the OT) has also been associated with such tunneling, where attention is not switched to deal with other high-priority events that may occur as the operator is trying to deal with the fault (Dismukes and Nowinski, 2007). Aviation data also suggest that compelling immersive 3D displays can lead to this sort of attentional tunneling in the OT, causing a failure to switch to unexpected IT events (Wickens and Alexander, 2009; Wickens et al., 2009a).

A third approach to interruption management is to examine the resumption of the OT after an interruption. Much of this work is based on a theory of "memory for goals" or "goal activation" (Trafton and Monk, 2007), in which the critical determinant of resumption is how well the goals of the OT are mentally preserved during the IT period. For example, actions taken prior to IT switching (i.e., active "placekeeping" or "bookmarking," or rehearsal) can serve to maintain the goal in a more active state at the time of the second (IT → OT)

switch and hence increase the fluency of restarting the OT (McDaniel and Einstein, 2007). The state of the OT at the time of the interruption can also be briefly rehearsed (Trafton et al, 2003). In a corresponding fashion, strategically postponing switch 1 until a subgoal in the OT has been completed ("let me just finish this paragraph") will better enable starting the OT where it was left off, rather than having to "start from scratch." A closely related factor is the modality of the OT. When this involves processing speech, because speech is transient and vulnerable and must be rehearsed, there is more of a tendency to delay the abandonment of the OT when it is in the vocal, rather than the text modality, because in the former case an instant switch may lead to forgetting what was just said and require a request for a repeat upon return to the OT; in the latter, with print, one can simply go back and read the still-present text (Latorella, 1996).

Designers of human–computer interfaces are considerintg ways in which intelligent automation can postpone interruptions of a user's ongoing task until the inference is made that a subgoal has been completed (Bailey and Konstan, 2006).

We also note that similar strategic factors in interruption management influence switching in dynamic control tasks; people do (or at least should) schedule switches when the system is in a more stable state. For example, in driving, it is more optimal to look downward to a secondary task, when the car is centered in the lane and heading straight, than when the car may be veering toward one side or the other. Switching in stable states during dynamic control is analogous to switching cognitive states after subgoal completion.

In closing, it should be noted that some concurrent task performance may be invoked during interruption management to the extent that the operator may be attempting to rehearse the goals or status of the OT while addressing the IT. We now turn to this issue of concurrent task performance.

## 7.2 Concurrent Processing

In contrast to sequential processing, an understanding of concurrent processing, whether in degraded mode or perfect parallel mode, depends on somewhat different mechanisms. These mechanisms are as closely related to the structure of the information-processing sequences within the tasks themselves as they are to the operator's knowledge of task importance and priority (although there are interactions between these two influences; Gopher, 1992). Here human factors interest is in the task features that can enable any sort of concurrent processing to emerge from serial processing and that can enable that concurrent processing to be perfect rather than degraded. Four characteristics appear to influence this degree of success: task similarity, task demand, and task structure and resource allocation (Wickens, 2002, 2007; Wickens and Hollands, 2000), although how these influences are exerted is somewhat complex.

### 7.2.1 Task Similarity

A high degree of similarity between two tasks may induce confusion, just as similarity between perceptual signals will cause confusion (Section 3.3), high similarity between items held in working memory will increase the degree of interference between them (Section 5.1), or high similarity between two response devices will increase the possibility of confusion of actions (Section 6.7). In the context of interruption management, high similarity between the OT and the IT will degrade performance, because the remembered status of the OT will get confused with material from the IT (Dismukes and Nowinski, 2007; Cellier and Eyrolle, 1992).

In contrast to similarity of material, making the *rules* governing two tasks more similar may allow the tasks to be better integrated, fostering *more* effective concurrent processing. This may involve using similar control dynamics on two axes of a tracking task (Chernikoff and LeMay, 1963; Fracker and Wickens, 1989) or using similar rules to map stimuli (events) to responses (actions) (Duncan, 1979). Rule similarity also facilitates the ability of an operator to *switch* attention between two tasks in sequential fashion when in the serial mode of processing. That is, it is easier to keep doing alternative versions of the same task than it is to switch between different tasks, as if there is some "overhead" penalty for switching rules (Rogers and Monsell, 1995).

### 7.2.2 Task Demand

Easier tasks are more likely to be performed concurrently (and perfectly) than are more difficult or demanding tasks, an intuitive effect well documented in the study of cell phone interference with driving (Collet et al., 2010). We argue that easier tasks are generally more automated and consume less mental effort or resources than do more difficult ones (see Chapter 9). Such automaticity can often be achieved by extensive practice on what are called *consistently mapped tasks* (Fisk et al., 1987). These are tasks in which in each encounter by the learner, certain properties of the relation between the displayed elements, cognitive operations, and responses remain constant. These mimic many properties of skill- and rule-based tasks, as described in Section 6.5. Such consistent mapping will lead to not only more rapid performance but also performance that is relatively attention free and hence will allow other tasks to be time shared successfully.

### 7.2.3 Task Structure

Certain structural differences between two time-shared tasks increase the efficiency of their concurrent processing, as if the two tasks demand entirely (or partially) separated resources within the human processing system, such that it is easier to distribute tasks across *multiple resources* than to focus them within a single resource (Wickens, 1991, 2002; Wickens et al., 2003). These resources appear to be defined by *processing code* (verbal or linguistic versus spatial), *processing stage* (perceptual–cognitive operations versus response operations), *perceptual modality* (auditory versus visual), and *visual subsystems* (focal vision, required for object recognition, versus ambient vision, required for orientation and locomotion; Previc, 1998).

To briefly illustrate these dichotomies of resources, it is because of the separate spatial and verbal code

resources that spatially guided manual responses may be more effective than vocal responses when operators must also rehearse verbal material, but vocal responses will be more effective than spatially guided manual ones when the operator must concurrently perform another spatial task, such as tracking (e.g., the disruption of driving caused by cell phone dialing, Collet et al., 2010). As another example of code interference, we saw in Section 5.4 that the spatial cognitive processes involved in navigation and vehicle control are better time shared with the verbal input of a memorized verbal route list than the spatial input of a memorized map. It is because of the separate stage-defined resources that we are often able effectively to time share responding operations (e.g., talking) with perceptual ones (e.g., scanning). It is because of the separate perceptual resources that designers have chosen to offload the heavy visual processing load of pilots and vehicle drivers with some information presented on auditory channels (Wickens et al., 2003). Finally, it is because of the separate visual channels that we can effectively keep a car centered in the lane (using ambient vision) while searching for and reading road signs (a task requiring focal vision).

### 7.2.4 Resource Allocation

Multiple resource demand (the converse of automaticity) and resource type can predict the total amount of interference between two tasks (Wickens, 2005), but the two constructs together say nothing about the extent to which one task or the other suffers the greater decrement or bears the brunt of that interference. This relative decrement is predicted by *task priority* or the *resource allocation policy* between the two, the third element of multiple resource model (Wickens, 2007). Such policy will be at least partially influenced by task importance, but other factors may come into play here, such as the difficulty of the task (greater emphasis is given to the harder task) or the task's intrinsic interest or *engagement* value, the latter describing the occasional interference of cell phone conversation with safe driving, despite the driving tasks' generally greater importance.

## 8 CONCLUSIONS

In conclusion, the systems with which people must interact vary vastly in their complexity, from the simple graph or tool to things like nuclear reactors or the physiology of a patient under anesthesia. As a consequence, they vary drastically in terms of the type and degree of demands imposed on the varying information-processing components we have discussed in this chapter. In some cases, systems will impose demands on components that are quite vulnerable: working memory, predictive capabilities, and divided attention, for example. At other times they may impose on human capabilities that are a source of great strength, particularly if these sources rely on the vast store of information that we retain in long-term memory, information that assists us in pattern recognition, top-down processing, chunking, and developing plans and scripts on the basis of past experience are examples. The

importance of practice and training in the development of this knowledge base cannot be overestimated.

In addition to facilitating the performance of experts in many ways, long-term memory has a second implication for the practice of human factors. This is that predictions of human performance in many systems can be based only partially on an understanding of the generic information-processing components described in this chapter. An equal and sometimes greater partner in this prediction is extensive *domain knowledge* regarding the particular system with which the human is interacting. As several of the chapters in this handbook address, the best prediction of human performance must be based on the intricate interaction between the information-processing components discussed here, the domain knowledge employed by the human operator, and the physical environment within and tools with which the operator works. The reader will find all of these issues covered from multiple perspectives in subsequent chapters of the handbook.

## REFERENCES

Adams, M. J., Tenney, Y. J., and Pew, R. W. (1991), "Strategic Workload and the Cognitive Management of Advanced Multi-Task Systems," SOAR CSERIAC 91-6, Crew System Ergonomics Information Analysis Center, Wright-Patterson AFB, OH.

Adams, M. J., Tenney, Y. J., and Pew, R. W. (1995), "Situation Awareness and the Cognitive Management of Complex Systems," *Human Factors*, Vol. 37, No. 1, pp. 85–104.

Adamson, R. E. (1952), "Functional Fixedness as Related to Problem Solving: A Repetition of Three Experiments," *Journal of Experimental Psychology*, Vol. 44, pp. 288–291.

Allpress, J. A., and Leland, L. S. (2010), "Reducing Traffic Speed within Roadwork Sites Using Obtrusive Perceptual Countermeasures," *Accident Analysis and Prevention*, Vol. 42, pp. 377–383.

Andre, A. D., and Wickens, C. D. (1990), "Display Control Compatibility in the Cockpit: Guidelines for Display Layout Analysis," University of Illinois Institute of Aviation Technical Report ARL-90-12/NASA-A3I-90-1, Aviation Research Laboratory, Savoy, IL.

Andre, A. D., and Wickens, C. D. (1995), "When Users Want What's NotBest for Them," *Ergonomics in Design*, Vol. 3, No. 4, pp. 10–14.

Aretz, A. J. (1991), "The Design of Electronic Map Displays," *Human Factors*, Vol. 33, No. 1, pp. 85–101.

Ariely, D. (2001), "Seeing Sets: Representation by Statistical Properties," *Psychological Science*, Vol. 12, pp. 157–162.

Atkinson, R. C., and Shiffrin, R. M. (1971), "The Control of Short-Term Memory," *Scientific American*, Vol. 225, pp. 82–90.

Awh, E., Jonides, J., and Reuter-Lorenz, P. A. (1998), "Rehearsal in Spatial Working Memory," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 24, pp. 780–790.

Baber, C. (1997), *Beyond the Desktop*, Academic, San Diego, CA.

Baddeley, A. (2000), "The Episodic Buffer: ANew Component of Working Memory?" *Trends in Cognitive Sciences*, Vol. 4, No. 11, pp. 417–423.

Baddeley, A. (2003), "Working Memory: Looking Back and Looking Forward," *Nature Reviews Neuroscience*, Vol. 4, pp. 829–839.

Baddeley, A. D., and Hitch, J. G. (1974), in *The Psychology of Learning and Motivation*, Vol. 8, G. A. Bower, Ed., Academic, New York, pp. 47–89.

Bailey, B. P., and Konstan, J. A. (2006), "On the Need for Attention-Aware Systems: Measuring Effects of Interruption on Task Performance, Error Rate, and Affective State," *Computers in Human Behavior*, Vol. 23, pp. 685–708.

Baker, C. H. (1961), "Maintaining the Level of Vigilance by Means of Knowledge of Results About a Secondary Vigilance Task," *Ergonomics*, Vol. 4, pp. 311–316.

Ballard, D. H., Hayhoe, M. M., and Pelz, J. B. (1995), "Memory Representation in Natural Tasks," *Journal of Cognitive Neuroscience*, Vol. 7, No. 1, pp. 66–86.

Banbury, S. P., and Treselian, S. (2004), *Cognitive Approaches to Situation Awareness*, Ashgate, Brookfield, VT.

Banbury, S. P., Macken, W. J., Tremblay, S., and Jones, D. M. (2001), "Auditory Distraction and Short-Term Memory: Phenomena and Practical Implications," *Human Factors*, Vol. 43, pp. 12–29.

Barnett, B. J, and Wickens, C. D. (1988), "Display Proximity in Multicue Information Integration: The Benefit of Boxes," *Human Factors*, Vol. 30, pp. 15–24.

Bellenkes, A. H., Wickens, C. D., and Kramer, A. F. (1997), "Visual Scanning and Pilot Expertise: The Role of Attentional Flexibility and Mental Model Development," *Aviation, Space, and Environmental Medicine*, Vol. 68, No. 7, pp. 569–579.

Bennett, K. B., and Flach, J. M. (1992), "Graphical Displays: Implications for Divided Attention, Focused Attention, and Problem Solving," *Human Factors*, Vol. 34, pp. 513–533.

Bennett, K. B., and Fritz, H. I. (2005), "Objects and Mappings: Incompatible Principles of Display Design" *Human Factors*, Vol. 47, No. 1, pp. 131–137.

Bennett, K. B., Toms, M. L., and Woods, D. D. (1993), "Emergent Features and Graphical Elements: Designing More Effective Configural Displays," *Human Factors*, Vol. 35, No. 1, pp. 71–98.

Beringer, D. B., and Chrisman, S. E. (1991), "Peripheral Polar-Graphic Displays for Signal/Failure Detection," *International Journal of Aviation Psychology*, Vol. 1, pp. 133–148.

Biederman, I. (1987). "Recognition-by-components: A theory of human image understanding." *Psychological Review*, *94*(2), 115–147.

Biederman, I., Mezzanotte, R. J., Rabinowitz, J. C., Francolin, C. M., and Plude, D. (1981), "Detecting the Unexpected in Photo Interpretation," *Human Factors*, Vol. 23, pp. 153–163.

Bizantz, A., and Roth, E. (2007), "Analysis of Cognitive Work," in *Reviews of Human Factors & Ergonomics*, Vol. 3, D. Boehm-David, Ed., Human Factors Society, Santa Monica, CA.

Bjork, R. A. (1999), "Assessing Our Own Competence: Heuristics and Illusions", in *Attention & Performance*, Vol. 17, D. Gopher and A. Koriat, Eds., MIT Press, Cambridge, MA.

Blake, R., and Sekuler, R. (2006), *Perception*, 5th ed., McGraw-Hill, Boston.

Bransford, J. D., and Johnson, M. K. (1972), "Contextual Prerequisites for Understanding: Some Investigations of Comprehension and Recall," *Journal of Verbal Learning and Verbal Behavior*, Vol. 11, pp. 717–726.

Braunstein, M. L., Andersen, G. J., Rouse, M. W., and Tittle, J. S. (1986), "Recovering Viewer Centered Depth from Disparity, Occlusion and Velocity Gradients," *Perception and Psychophysics*, Vol. 40, No. 4, pp. 216–224.

Breznitz, S. (1983), *Cry Wolf: The Psychology of False Alarms*, Lawrence Erlbaum Associates, Mahwah, NJ.

Broadbent, D. (1958), *Perception and Communications*, Permagon, New York.

Broadbent, D. (1972), *Decision and Stress*, Academic, New York.

Broadbent, D. E. (1982), "Task Combination and Selective Intake of Information," *Acta Psychologica*, Vol. 50, pp. 253–290.

Brown, J. (1959), "Some Tests of the Decay Theory of Immediate Memory," *Quarterly Journal of Experimental Psychology*, Vol. 10, pp. 12–21.

Buchner, A., Brandt, M., Bell, R., and Weise, J. (2006), "Car Backlight Position and Fog Density Bias Observer-Car Distance Estimates and Time-to-Collision Judgments," *Human Factors*, Vol. 48, pp. 300–317.

Burns, C. (2000), "Putting It All Together: Improving Integration in Ecological Displays," *Human Factors*, Vol. 42, pp. 226–241.

Burns, C. M., Bisantz, A. M., and Roth, E. M. (2004), "Lessons from a Comparison of Work Domain Models," *Human Factors*, Vol. 46, pp. 711–727.

Buttigieg, M. A., and Sanderson, P. M. (1991), "Emergent Features in Visual Display Design for Two Types of Failure Detection Tasks," *Human Factors*, Vol. 33, pp. 631–651.

Card, S. K. (1981), "The Model Human Processor: A Model for Making Engineering Calculations of Human Performance," in *Proceedings of the 25th Annual Meeting of the Human Factors Society*, R. Sugarman, Ed., Human Factors Society, Santa Monica, CA.

Carpenter, P. A., and Just, M. A. (1975), "Sentence Comprehension: A Psycholinguistic Processing Model of Verification," *Psychological Review*, Vol. 82, No. 1, pp. 45–73.

Carswell, C. M. (1992), "Reading Graphs: Interactions of Processing Requirements and Stimulus Structure," in *Precepts, Concepts, and Categories*, B. Burns, Ed., Elsevier Science, Amsterdam, pp. 605–647.

Carswell, C. M., and Wickens, C. D. (1987), "Information Integration and the Object Display: An Interaction of Task Demands and Display Superiority," *Ergonomics*, Vol. 30, pp. 511–527.

Carswell, C. M., and Wickens, C. D. (1990), "The Perceptual Interaction of Graphical Attributes: Configurality, Stimulus Homogeneity, and Object Integration," *Perception and Psychophysics*, Vol. 47, pp. 157–168.

Carswell, C. M., and Wickens, C. D. (1996), "Mixing and Matching Lower-Level Codes for Object Displays: Evidence for Two Sources of Proximity Compatibility," *Human Factors*, Vol. 38, pp. 61–72.

Cellier, J., and Eyrolle, H. (1992), "Interference between Switched Tasks," *Ergonomics*, Vol. 35, No. 1, pp. 25–36

Chan, A., and Hofmann, E. (2010), "2010" *Ergonomics*, Vol. 53, pp. 1403–1419

Chernikoff, R., and LeMay, M. (1963), "Effect of Various Display-Control Configurations on Tracking with Identical and Different Coordinate Dynamics," *Journal of Experimental Psychology*, Vol. 6, pp. 95–99.

Chi, M. T. H., Glaser, R., and Farr, M. J., Eds. (1988), *The Nature of Expertise*, Erlbaum, Hillsdale, NJ.

Chong, S. C., and Treisman, A. (2005), "Statistical Processing: Computing the Average Size in Perceptual Groups," *Vision Research*, Vol. 45, pp. 891–900.

Chou, C., Madhavan, D., and Funk, K. (1996), "Studies of Cockpit Task Management Errors," *International Journal of Aviation Psychology*, Vol. 6, No. 4, pp. 307–320.

Clark, H. H., and Chase, W. G. (1972), "On the Process of Comparing Sentences Against Pictures," *Cognitive Psychology*, Vol. 3, pp. 472–517.

Cleveland, W. S. (1985), *The Elements of Graphing Data*, Wadsworth, Monterey, CA.

Cleveland, W. S., and McGill, R. (1984), "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphic Methods," *Journal of the American Statistical Association*, Vol. 70, pp. 531–554.

Cleveland, W. S., and McGill, R. (1985), "Graphical Perception and Graphical Methods for Analyzing Scientific Data," *Science*, Vol. 229, pp. 828–833.

Cohen, H. H. (2003), "Why Do We Bump into Things While Walking? Revisited," *Ergonomicsin Design*, Vo. 11, No. 4, pp. 12–15.

Collet, C., Guillot, A., and Petit, C. (2010), "Phoning While Driving II: A Review of Driving Condition Influence," *Ergonomics*, Vol. 53, pp. 602–616.

Cropley, A. J. (1999), "Creativity and Cognition: Producing Effective Novelty," *Roeper Review*, Vol. 21, pp. 253–260.

Cutting, J. E., and Vishton, P. M. (1995), "Perceiving Layout and Knowing Distances: The Integration, Relative Potency, and Contextual Use of Different Information About Depth," in *Perception of Space and Motion*, W. Epstein and S. Rogers, Eds., Academic, San Diego, CA.

Damle, A. (2010), "Using Monochromatic Design Views to Avoid a Premature Fixation on Design Solutions," *Human Factors*, Vol. 52, No. 1, pp. 28–40.

Damos, D., Ed. (1991), *Multiple Task Performance*, Taylor & Francis, London.

Davidson, J. E., and Sternberg, R. J. (1998), "Smart Problem-Solving: How Metacognition Helps," in *Metacognition in Educational Theory and Practice*, D. J. Hacker, J. Dunlosky, and A. C. Graesser, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 47–65.

Degani, A., and Wiener, E. L. (1993), "Cockpit Checklists: Concepts, Design, and Use," *Human Factors*, Vol. 35, No. 4, pp. 345–360.

DeLucia, P. R., Mather, R. D., Griswold, J. A., and Mitra, S. (2006), "Toward the Improvement of Image-Guided Interventions for Minimally Invasive Surgery: Three Factors That Affect Performance," *Human Factors*, Vol. 48, No. 1, pp. 23–38.

Denton, G. G. (1980), "The Influence of Visual Pattern on Perceived Speed," *Perception*, Vol. 9, pp. 393–402.

DeSoto, C. B., London, M., and Handel, S. (1965), "Social Reasoning and Spatial Paralogic," *Journal of Personal and Social Psychology*, Vol. 2, pp. 513–521.

Dismukes, K. (2001), "The Challenge of Managing Interruptions, Distractions, and Deferred Tasks," in *Proceedings of the 11th International Symposium on Aviation Psychology*, Ohio State University, Columbus, OH.

Dismukes, K., and Nowinski, J. (2007). "Prospective Memory, Concurrent Task Management and Pilot Error," in, *Attention: From Theory to Practice*, A. Kramer, D. Wiegmann, and A. Kirlik, Eds., Oxford University Press, Oxford.

Dixon, S., Wickens, C. D., and McCarley, J. M. (2007), "On the Independence of Reliance and Compliance: Are False Alarms Worse Than Misses," *Human Factors*, Vol. 49, pp. 564–572.

Drury, C. (1975), "Application to Fitts' Law to Foot Pedal Design," *Human Factors*, Vol. 17, pp. 368–373.

Drury, C. (1994), "The Speed-Accuracy Tradeoff in Industry," *Ergonomics*, Vol. 37, pp. 747–763.

Drury, C. G., and Addison, S. L. (1973), "An Industrial Study of the Effects of Feedback and Fault Density on Inspection Performance," *Ergonomics*, Vol. 16, pp. 159–169.

Duncan, J. (1979), "Divided Attention: The Whole Is More Than the Sum of Its Parts," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 5, pp. 216–228.

Duncan, J. (1984), "Selective Attention and the Organization of Visual Information", *Journal of Experimental Psychology: General*, Vol. 119, pp. 501–517.

Durso, F., Rawson, K., and Girotto, S. (2007), "Comprehension and Situation Awareness," in *Handbook of Applied Cognition*, F. Durso, Ed., Wiley, Chichester.

Durso, F. T., Truitt, T. R., Hackworth, C., Crutchfield, J., and Manning, C. A. (1998), "En Route Operational Errors and Situation Awareness," *International Journal of Aviation Psychology*, Vol. 8, pp. 177–194.

Eberts, R. E., and MacMillan, A. G. (1985), "Misperception of Small Cars," in *Trends in Ergonomics/Human Factor*, Vol. 2, R. E. Eberts and C. G. Eberts, Eds., Elsevier Science, Amsterdam, pp. 33–39.

Egeth, H. E., and Yantis, S. (1997), "Visual Attention: Control, Representation, and Time Course," *Annual Review of Psychology*, Vol. 48, pp. 269–297.

Egly, R., Driver, J., and Rafal, R. D. (1994a), "Shifting Visual Attention between Objects and Locations: Evidence from Normal and Parietal Lesion Subjects," *Journal of Experimental Psychology: General*, Vol. 123, pp. 161–177.

Egly, R., Rafal, R. D., Driver, J., and Starrveveld, Y. (1994b), "Covert Orienting in the Split Brain Reveals Hemispheric Specialization for Object-Based Attention," *Psychological Science*, Vol. 5, pp. 380–383.

Eley, M. G. (1988), "Determining the Shapes of Land Surfaces from Topographical Maps," *Ergonomics*, Vol. 31, pp. 355–376.

Elzer, S., Green, N., and Carberry, S. (2003), "Exploiting Cognitive Psychology Research for Recognizing Intention in Information Graphics," paper presented at the Annual Conference of the Cognitive Science Society, Boston, August.

Endsley, M. R. (1995), "Measurement of Situation Awareness in Dynamic Systems," *Human Factors*, Vol. 37, No. 1, pp. 65–84.

Endsley, M. R., and Garland, D. J. (2000), "Situation Awareness Analysis and Measurement," Lawrence Erlbaum Associates, Mahwah, NJ.

Engle, R. W. (2002), "Working Memory Capacity as Executive Attention," *Current Directions in Psychological Science*, Vol. 11, pp. 19–23

Engle, R. W., Tuholski, S. W., Laughlin, J. E., and Conway, A. R. A. (1999), "Working Memory, Short-Term Memory, and General Fluid Intelligence: A Latent-Variable Approach," *Journal of Experimental Psychology: General*, Vol. 128, pp. 309–331.

Ericsson, K. A., and Kintsch, W. (1995), "Long-Term Working Memory," *Psychological Review*, Vol. 102, pp. 211–245.

Evans, L. (1991), *Traffic Safety and the Driver*, Van Nostrand Reinhold, New York.

Fadden, S., Ververs, P. M., and Wickens, C. D. (2001), "Pathway HUDS: Are They Viable?" *Human Factors*, Vol. 43, No. 2, pp. 173–193.

Finke, R. A., Ward, T. B., and Smith, S. M. (1992), *Creative Cognition*, MIT Press, Cambridge, MA.

Fisher, D. L., Coury, B. G., Tengs, T. O., and Duffy, S. A. (1989), "Minimizing the Time to Search Visual Displays: The Role of Highlighting," *Human Factors*, Vol. 31, No. 2, pp. 167–182.

Fisk, A. D., Ackerman, P. L., and Schneider, W. (1987), "Automatic and Controlled Processing Theory and Its Applications to Human Factors Problems," in *Human Factors Psychology*, P. A. Hancock, Ed., Elsevier Science, New York, pp. 159–197.

Fitts, P. M. (1966), "Cognitive Aspects of Information Processing, III: Set for Speed Versus Accuracy," *Journal of Experimental Psychology*, Vol. 71, pp. 849–857.

Fitts, P. M., and Peterson, J. R. (1964), "Information Capacity of Discrete Motor Responses," *Journal of Experimental Psychology*, Vol. 67, pp. 103–112.

Fitts, P. M., and Posner, M. A. (1967), *Human Performance*, Brooks/Cole, Pacific Palisades, CA.

Flach, J. M., Hancock, P. A., Caird, J., and Vicente, K. J., Eds. (1995), *Global Perspectives on the Ecology of Human–Machine Systems*, Lawrence Erlbaum Associates, Mahwah, NJ.

Flavell, J. H. (1979), "Metacognition and Cognitive Monitoring: A New Area of Cognitive Developmental Inquiry," *American Psychologist*, Vol. 34, pp. 906–911.

Fracker, M. L., and Wickens, C. D. (1989), "Resources, Confusions, and Compatibility in Dual Axis Tracking: Displays, Controls, and Dynamics," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 15, pp. 80–96.

Gathercole, S. E. (1997), "Models of Verbal Short-Term Memory," in *Cognitive Models of Memory*, M. A. Conway, Ed., MIT Press, Cambridge, MA, pp. 13–45.

Gathercole, S. E., and Baddeley, A. D. (1993), *Working Memory and Language*, Lawrence Erlbaum Associates Hove, East Sussex, England.

Gentner, D., and Stevens, A. L. (1983), *Mental Models*, Lawrence Erlbaum Associates, Mahwah, NJ.

Gibson, J. J. (1979), *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston.

Giffin, Gillam, B. (1980), "Geometrical Illusions," *Scientific American*, Vol. 242, pp. 102–111.

Gillan, D. J., Wickens, C. D., Hollands, J. G., and Carswell, C. M. (1998), "Guidelines for Presenting Quantitative Data in HFES Publications," *Human Factors*, Vol. 40, pp. 28–41.

Gold, J., Bennett, P. J., and Sekuler, A. B. (1999), "Signal But Not Noise Changes with Perceptual Learning," *Nature*, Vol. 402, pp. 176–178.

Gonzales, V. G., and Mark, G. (2004), "Constant Multi-Tasking Craziness: Managing Multiple Working Spheres," in *Human Factors of Computing Systems: CHI 04*, Association for Computing Machinery, New York, pp. 113–120.

Gopher, D. (1992), "The Skill of Attention Control: Acquisition and Execution of Attention Strategies," in *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience—A Silver Jubilee*, D. Meyer and S. Kornblum, Eds., MIT Press, Cambridge, MA.

Graf, R., and Torrey, J. W. (1966), "Perception of Phrase Structure in Written Language," in *Proceedings of the Annual Convention of the American Psychological Association APA*, Washington, DC, pp. 83–84.

Graham, J. L. (1937), "Illusory Trends in the Observation of Bar Graphs," *Journal of Experimental Psychology*, Vol. 20, pp. 597–608.

Gray, W. D., and Fu, W. T. (2004), "Soft Constraints in Interactive Behavior: The Case of Ignoring Perfect Knowledge In-The-World for Imperfect Knowledge In-The-Head", *Cognitive Science*, Vol. 28, pp. 359–382.

Green, C., and Hummel, J. E. (2006), "Familiar Interacting Object Pairs Are Perceptually Grouped," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 32, pp. 1107–1119.

Green, D. M., and Swets, J. A. (1989), *Signal Detection Theory and Psychophysics*, Wiley, New York.

Gregory, R. L. (1980), "The Confounded Eye," in *Illusion in Nature and Art*, R. E. Gregory and E. H. Gombrich, Eds., Charles Scribner's Sons, New York.

Gugerty, L., and Brooks, J. (2001), "Seeing Where You Are Heading: Integrating Environmental and Egocentric Reference Frames in Cardinal Direction Judgments," *Journal of Experimental Psychology: Applied*, Vol. 7, No. 3, pp. 251–266.

Hancock, P., Flach, J., Caird, J., and Vicente, K., Eds. (1995), *Local Applications of the Ecological Approach to Human–Machine Systems*, Lawrence Erlbaum Associates, Mahwah, NJ.

Harris, D. H., and Chaney, F. D. (1969), *Human Factors in Quality Assurance*, Wiley, New York.

Hart, S. G., and Wickens, C. D. (1990), "Workload Assessment and Prediction," in *MANPRINT: An Approach to Systems Integration*, H. R. Booher, Ed., Van Nostrand Reinhold, New York, pp. 257–296.

Herrmann, D., Brubaker, B., Yoder, C., Sheets, V., and Tio, A. (1999), "Devices That Remind," in *Handbook of Applied Cognition*, F. T. Durso, Ed., Wiley, New York, pp. 377–407.

Hess, S. M., and Detweiler, M. C. (1995), "The Effects of Response Alternatives on Keeping-Track Performance," in *Proceedings of the 39th Annual Meeting of the Human Factors and Ergonomics Society*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1390–1394.

Hettinger, L. J. (2002), "Illusory Self-Motion in Virtual Environment," in *Handbook of Virtual Environments: Design, Implementation, and Application*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 471–491.

Hick, W. E. (1952), "On the Rate of Gain of Information," *Quarterly Journal of Experimental Psychology*, Vol. 4, pp. 11–26.

Hickox, J. C., and Wickens, C. D. (1999), "Effects of Elevation Angle Disparity, Complexity, and Feature Type on Relating Out-of-Cockpit Field of View to an Electronic Cartographic Map," *Journal of Experimental Psychology: Applied*, Vol. 5, No. 3, pp. 284–301.

Ho, C. Y., Nikolic, M. I., Waters, M., and Sarter, N. B. (2004), "Not Now! Supporting Interruption Management by Indicating the Modality and Urgency of Pending Tasks", *Human Factors*, Vo. 46, pp. 399–410.

Hochberg, J. (1986), "Representation of Motion and Space in Video and Cinematic Displays," in *Handbook of Perception and Human Performance*, Vol. 1, K. Boff, R. Kaufman, and J. Thomas, Eds., Wiley Interscience, New York, pp. 22-1–22-64.

Hochberg, J. (1998), "Gestalt Theory and Its Legacy," in *Perception and Cognition at Century's End*, J. Hochberg, Ed., Academic, New York, pp. 253–306.

Hoffman, R., Crandall, B., and Shadbolt, N. (1998), "Use of the Critical Decision Method to Elicit Expert Knowledge". *Human Factors*, Vol. 40, pp. 254–276.

Hollands, J. G., and Dyre, B. P. (2000), "Bias in Proportion Judgments: The Cyclical Power Model," *Psychological Review*, Vol. 107, No. 3, pp. 500–524.

Hollands, J. G., and Spence, I. (1992), "Judgments of Change and Proportion in Graphical Perception," *Human Factors*, Vol. 34, No. 3, pp. 313–334.

Holohan, C. J., Culler, R. E., and Wilcox, B. L. (1978), "Effects of Visual Distraction on Reaction Time in a Simulated Traffic Environment," *Human Factors*, Vol. 20, No. 4, pp. 409–413.

Horrey, W. J., and Wickens, C. D. (2006), "The Impact of Cell Phone Conversations on Driving: A Meta-Analytic Approach," *Human Factors*, Vol. 48, No. 1, pp. 196–205.

Horrey, W. J., Wickens, C. D., and Consalus, K. P. (2006), "Modeling Drivers' Visual Attention Allocation While Interacting with In-Vehicle Technologies," *Journal of Experimental Psychology: Applied*, Vol. 12, No. 2, pp. 67–78.

Huestegge, L., and Phillipp, A. M, (in press), "Effects of Spatial Compatibility on Integration Processes in Graph Comprehension," *Attention, Perception & Psychophysics*.

Hutchins, E. L., Hollan, J. D., and Norman, D. A. (1985), "Direct Manipulation Interfaces," *Human–Computer Interaction*, Vol. 1, No. 4, pp. 311–338.

Hyman, R. (1953), "Stimulus Information as a Determinant of Reaction Time," *Journal of Experimental Psychology*, Vol. 45, pp. 423–432.

Iani, C., and Wickens, C. D. (2007), "Factors Affecting Task Management in Aviation," *Human Factors*, Vol. 49, No. 1, pp. 16–24

Jagacinski, R. J., and Flach, J. M. (2003), *Control Theory for Humans*, Lawrence Erlbaum Associates, Mahwah, NJ.

James, W. (1890), *The Principles of Psychology*, Vol. 1, Holt, New York.

Jenkins, D., Stanton, N., Salmon, P., and Walder, G. (2009), *Cognitive Work Analysis: Coping with Complexity*, Ashgate, Averbury VT.

Johnson, A., and Proctor, R. W. (2004), *Attention Theory and Practice*, Sage, Thousand Oaks, CA.

Johnson, S. L., and Roscoe, S. N. (1972), "What Moves, the Airplane or the World?" *Human Factors*, Vol. 14, pp. 107–129.

Jolicoeur, P., and Ingleton, M. (1991), "Size Invariance in Curve tracing," *Memory & Cognition*, Vol. 19, No. 1, pp. 21–36.

Jones, D. G., and Endsley, M. R. (1996), "Sources of Situation Awareness Errors in Aviation," *Aviation, Space, and Environmental Medicine*, Vol. 67, No. 6, pp. 507–512.

Just, M. A., and Carpenter, P. A. (1987), *The Psychology of Reading and Language Comprehension*, Allyn and Bacon, Boston.

Kahneman, D. (1973), *Attention and Effort*, Prentice-Hall, Englewood Cliffs, NJ.

Kahneman, D., and Klein, G. (2009), "Conditions for Intuitive Expertise," *American Psychologist*, Vol. 64, pp. 215–226

Kahneman, D., and Treisman, D. (1984), "Changing Views of Attention and Automaticity," in *Varieties of Attention*, R. Parasuraman, R. Davies, and J. Beatty, Eds., Academic, New York, pp. 29–61.

Kintsch, W., and Keenan, J. (1973), "Reading Rate and Retention as a Function of the Number of Propositions in the Base Structure of Sentences," *Cognitive Psychology*, Vol. 5, pp. 257–274.

Kintsch, W., and Van Dijk, T. A. (1978), "Toward a Model of Text Comprehension and Reproduction," *Psychological Review*, Vol. 85, pp. 363–394.

Kintsch, W., and Vipond, P. (1979), "Reading Comprehension and Readability in Educational Practice and Psychological Theory," in *Perspectives on Memory Research*, L. G. Nilsson, Ed., Lawrence Erlbaum Associates, Mahwah, NJ.

Kirlik, A., Miller, A., and Jagacinski, R. J. (1993), "Supervisory Control in a Dynamic and Uncertain Environment: A Process Model of Skilled Human–Environment Interaction," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 23, No. 4, pp. 929–952.

Klein, G., and Crandall, B. W. (1995), "The Role of Simulation in Problem Solving and Decision Making," in *Local Applications of the Ecological Approach to Human–Machine Systems*, P. Hancock, J. Flach, J. Caird, and K. Vicente, Eds., Lawrence Erlbaum Associates, Mahwah, NJ.

Klein, G. A., Orasanu, J., Calderwood, R., and Zsambok, E., Eds. (1993), *Decision Making in Action: Models and Methods*, Ablex, Norwood, NJ.

Kleinman, and Pattipati, (1991), "A Review of Engineering Models of Information Processing and Decision Making in Multitask Supervisory Control," in *Multiple Task Performance*, D. Damos, Ed., Taylor & Francis, London.

Kosslyn, S. M. (1987), "Seeing and Imagining in the Cerebral Hemispheres: A Computational Approach," *Psychological Review*, Vol. 94, pp. 148–175.

Kosslyn, S. M. (2006), *Graph Design for the Eye and Mind*, Oxford University Press, New York.

Kundel, H. L., and Nodine, C. F. (1978), "Studies of Eye Movements and Visual Search in Radiology," in *Eye Movements and Higher Psychological Functions*, J. Senders, D. Fisher, and R. Monty, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 317–328.

Langolf, C. D., Chaffin, D. B., and Foulke, S. A. (1976), "An Investigation of Fitts' Law Using a Wide Range of Movement Amplitudes," *Journal of Motor Behavior*, Vol. 8, pp. 113–128.

Latorella, K. A. (1996), "Investigating Interruptions—An Example from the Flightdeck" *Proceedings of the 40th Annual Meeting HFES*, Human Factors Society, Santa Monica, CA.

Layton, C., Smith, P. J., and McCoy, C. E. (1994), "Design of a Cooperative Problem-Solving System for En-Route Flight Planning: An Empirical Evaluation," *Human Factors*, Vol. 36, No. 4, pp. 94–119.

Lee, J. D. (2005), "Driving Safety,"in *Reviews of Human Factors and Ergonomics*, R. S. Nickerson, Ed., Human Factors and Ergonomics Society, Santa Monica, CA, pp. 172–218.

Leibowitz, H. W. (1985), "Grade Crossing Accidents and Human Factors Engineering," *American Scientist*, Vol. 73, pp. 558–562.

Levine, D. T., Momen, N., Drivdahl, S. B., and Simons, D. J. (2000), "Change Blindness Blindness: The Metacognitive Error of Overestimating Change-Detection Ability," *Visual Cognition*, Vol. 7, pp. 397–412.

Levine, M. (1982), "You-Are-Here Maps: Psychological Considerations," *Environment and Behavior*, Vol. 14, pp. 221–237.

Levine, M. (1988), *Effective Problem Solving*, Prentice-Hall, Englewood Cliffs, NJ.

Liu, Y., Feyen, R., and Tsimhoni, O. (2006). "Queueing Network-Model Human Processor (QN-MHP): A Computational Architecture for Multitask Performance in Human-Machine Systems," *ACM Transactions on Computer-Human Interaction*, Vol. 13, No. 1. pp. 37–70.

Loftus, G., Dark, V., and Williams, D. (1979), "Short-Term Memory Factors in Ground Controller/Pilot Communication," *Human Factors*, Vol. 21, pp. 169–181.

Logie, R. H. (1995), *Visuo-Spatial Working Memory*, Lawrence Erlbaum Associates, Hove, East Sussex, England.

Loukopopolous, L. Dismukes, K., and Barshi, I. (2009), *The Multi Tasking Myth*, Ashgate, Avebury, VT.

Lusted, L. B. (1976), "Clinical Decision Making," in *Decision Making and Medical Care*, D. Dombal and J. Grevy, Eds., North-Holland, Amsterdam.

Macedo, J., Kaber, D., Endsley, M., Powanusorn, P., and S. Myung (1998), "The Effect of Automated Compensation for Incongruent Axes on Teleoperator Performance," *Human Factors*, Vol. 40, pp. 541–553.

Macmillan, N. A., and Creelman, C. D. (2005), *Detection Theory: A User's Guide*. Lawrence Erlbaum, Mahwah, NJ.

Maltz, M., and Shinar, D. (2003), "New Alternative Methods of Analyzing Human Behavior in Cued Target Acquisition," *Human Factors*, Vol. 45, No. 2, pp. 281–295.

Massaro, D. W. (1979), "Letter Information and Orthographic Context in Word Perception," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 5, pp. 595–609.

Mayer, R. (2007), "Task Based Guidelines for Multi-Media Instructions," in *Reviews of Human Factors & Ergonomics*, Vol. 3, D. Boehm-Davis, Ed., Human Factors Society, Santa Monica, CA.

Mayer, R. E. (1983), *Thinking, Problem Solving, Cognition*, W. H. Freeman, San Francisco.

Mayer, R. E. (1999), "Fifty Years of Creativity Research," in *Handbook of Creativity*, R. J. Sternberg, Ed., Cambridge University Press, New York, pp. 449–460.

McCarley, J. M., Kromer, A. F., Wickens, C. D., Vidoni, E. D., and Boot, W. R. (2004), "Visual Skills in Airport Security Screening," *Psychological Science*, Vol. 15, pp. 302–306.

McClelland, J. L., and Rumelhart, D. E. (1981), "An Interactive Activation Model of Context Effects in Letter Perception, I: An Account of Basic Findings," *Psychological Review*, Vol. 88, pp. 375–407.

McDaniel, M. A., and Einstein, G. O. (2007), *Prospective Memory: An Overview and Synthesis of an Emerging Field*, Sage, Thousand Oaks, CA.

McFarlane, D. C., and Latorella, K. A. (2002), "The Score and Importance of Human Interruption in Human–Computer Interface Design," *Human–Computer Interaction*, Vol. 17, No. 1, pp. 1–61.

McGuirl, J. M., and Sarter, N. B. (2006), "Supporting Trust Calibration and the Effective use of Decision Aids by Presenting Dynamic System Confidence Information," *Human Factors*, Vol. 48, No. 4, pp. 656–665.

McKoon, G., and Ratcliff, R. (1992), "Inference during Reading," *Psychological Review*, Vol. 99, pp. 440–466.

McRuer, D. (1980), "Human Dynamics in Man–Machine Systems," *Automatica*, Vol. 16, pp. 237–253.

Meyer, D. E., and Kieras, D. E. (1997), "A Computational Theory of Executive Cognitive Processes and Multiple-Task Performance: Part 1: Basic Mechanisms," *Psychological Review*, Vol. 104, pp. 3–65.

Milkman, K., Clugh, D., and Bazerman, M. (2009), "How Can Decision Making Be Improved?" *Perspectives on Psychological Science*, Vol. 4, pp. 379–383.

Miller, G. A. (1956), "The Magical Number Seven Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psychological Review*, Vol. 63, pp. 81–97.

Miller, G. A., Galanter, E., and Pribram, K. H. (1960), *Plans and the Structure of Behavior*, Holt, Rinehart and Winston, New York.

Monsell, S. (2003), "Task Switching," *Trends in Cognitive Science*, Vol. 7, pp. 134–140.

Moray, N. (1981), "The Role of Attention in the Detection of Errors and the Diagnosis of Errors in Man–Machine Systems," in *Human Detection and Diagnosis of System Failures*, J. Rasmussen and W. Rouse, Eds., Plenum, New York.

Moray, N. (1986), "Monitoring Behavior and Supervisory Control," in *Handbook of Perception and Performance*, Vol. 2, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., Wiley, New York, pp. 40-1–40-51.

Moray, N., Lee, J., Vicente, K., Jones, B. G., and Rasmussen, J. (1994), "A Direct Perception Interface for Nuclear Power Plants," in *Proceedings of the 38th Annual Meeting of the Human Factors and Ergonomics Society*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 481–485.

Mori, H., and Hayashi, Y. (1995). Visual Interference with Users' Tasks on Multi-Window Systems, *International Journal of Human–Computer Interaction*, Vol. 7, pp. 329–340.

Neisser, U. (1967), *Cognitive Psychology*, Prentice-Hall, Englewood Cliffs, NJ.

Neisser, U., Novick, R., and Lazar, R. (1964), "Searching for Novel Targets," *Perceptual and Motor Skills*, Vol. 19, pp. 427–432.

Nickerson, R. (1998), "Confirmation Bias: AUbiquitous Phenomenon," *Review of General Psychology*, Vol. 2, pp. 175–220

Nickerson, R. S. (1999), "Enhancing Creativity," in *Handbook of Creativity*, R. J. Sternberg, Ed., Cambridge University Press, New York, pp. 392–430.

Norman, D. (1988), *The Psychology of Everyday Things*, Basic Books, New York.

Norman, D. A. (1981), "Categorization of Action Slips," *Psychological Review*, Vol. 88, pp. 1–15.

North, C. (2006), "Information Visualization,"in *Handbook of Human Factors & Ergonomics*, 3rd., G. Salvendy, Ed., Wiley, Hoboken, NJ.

Oliva, A., and Torralba, A. (2010), "The Role of Context in Object Recognition," *Trends in Cognitive Science*, Vol. 11, No. 12, pp. 620–627.

Olmos, O., Liang, C.-C., and Wickens, C. D. (1997), "Electronic Map Evaluation in Simulated Visual Meteorological Conditions," *International Journal of Aviation Psychology*, Vol. 7, No. 1, pp. 37–66.

Olmos, O., Wickens, C. D., and Chudy, A. (2000), "Tactical Displays for Combat Awareness: An Examination of Dimensionality and Frame of Reference Concepts and the Application of Cognitive Engineering," *International Journal of Aviation Psychology*, Vol. 10, No. 3, pp. 247–271.

Orasanu, J. M. (1993), "Decision-Making in the Cockpit," in *Cockpit Resource Management*, E. L. Wiener, B. G. Kanki, and R. L. Helmreich, Eds., Academic, San Diego, CA, pp. 137–173.

Orasanu, J., and Fischer, U. (1997), "Finding Decisions in Natural Environments: The View from the Cockpit," in *Naturalistic Decision Making*, C. Zsambok and G. Klein, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 343–357.

Osborne, D. W., and Ellingstad, V. S. (1987), "Using Sensor Lines to Show Control–Display Linkages on a Four Burner Stove," in *Proceedings of the 31st Annual Meeting of the Human Factors Society*, Human Factors Society, Santa Monica, CA, pp. 581–584.

Paas, F., Renkl, A., and Sweller, J. (2003), "Cognitive Load Theory and Instructional Design: Recent Developments," *Educational Psychologist*, Vol. 31, pp. 1–5.

Palmer, S. E. (1975), "The Effects of Contextual Scenes on the Identification of Objects," *Memory and Cognition*, Vol. 3, pp. 519–526.

Parasuraman, R., Davies, R., and Beatty, J., Eds. (1984), *Varieties of Attention*, Academic, New York.

Parasuraman, R., Hancock, P. A., and Obofinbaba, O. (1997), "Alarm Effectiveness in Driver Centered Collision Warning Systems," *Ergonomics*, Vol. 40, pp. 390–399.

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000), "A Model for Types and Levels of Human Interaction with Automation," *IEEE Transactions on Systems, Man, & Cybernetics: Part A: Systems and Humans*, Vol. 30, No. 3, pp. 286–297.

Parasuraman, R., Sheridan, T., and Wickens, C. D. (2008), "Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs," *Cognitive Engineering and Decision Making*, Vol. 2, pp. 140–160.

Pashler, H. E. (1998), *The Psychology of Attention*, MIT Press, Cambridge, MA.

Patterson, R. (1997), "Visual Processing of Depth Information in Stereoscopic Displays," *Human Factors*, Vol. 17, pp. 69–74.

Payne, S. J. (1995), "Naive Judgments of Stimulus-Response Compatibility," *Human Factors*, Vol. 37, No. 3, pp. 495–506.

Penrose, J. M. (2008), "Annual Report Graphic Use: A Review of the Literature," *Journal of Business Communication*, Vol. 45, No. 2, pp. 158–180.

Peterson, L. R., and Peterson, M. J. (1959), "Short-Term Retention of Individual Verbal Items," *Journal of Experimental Psychology*, Vol. 58, pp. 193–198.

Pew, R. W., and Baron, S. (1978), "The Components of an Information Processing Theory of Skilled Performance Based on an Optimal Control Perspective," in *Information Processing in Motor Control and Learning*, G. E. Stelmach, Ed., Academic, New York.

Phillips, J. G., Triggs, T. J., and Meehan, J. W. (2003), "Conflicting Directional and Locational Cues Afforded by Arrowhead Cursors in Graphical User Interfaces," *Journal of Experimental Psychology: Applied*, Vol. 9, pp. 75–87.

Pomerantz, J. R. (1981), "Perceptual Organization in Information Processing," in *Perceptual Organization*, M. Kubovy and J. R. Pomerantz, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 141–180.

Pomerantz, J. R., and Kubovy, M. (1986), "Theoretical Approaches to Perceptual Organization," in *Handbook of Perception and Human Performance*, Vol. II, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., Wiley, New York, pp. 36-1–36-46.

Pomerantz, J. R., and Pristach, E. A. (1989), "Emergent Features, Attention, and Perceptual Glue in Visual Form Perception," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 15, pp. 635–649.

Posner, M. I. (1978), *Chronometric Explorations of the Mind*, Lawrence Erlbaum Associates, Mahwah, NJ.

Postle, B. R., Idzikowski, C., Salac, S. D., Logie, R. H., and Baddeley, A. D. (2006), "The Selective Disruption of Spatial Working Memory by Eye Movements," *Quarterly Journal of Experimental Psychology*, Vol. 59, No. 1, pp. 100–120.

Poulton, E. C. (1985), "Geometric Illusions in Reading Graphs," *Perception and Psychophysics*, Vol. 37, pp. 543–548.

Previc, F., and Ercoline, W. (2004), *Spatial Orientation in Aviation*, American Institute of Aeronautics and Astronautics, Reston, VA.

Previc, F. H. (1998), "The Neuropsychology of 3-D Space," *Psychological Bulletin*, Vol. 124, pp. 123–164.

Prinzel, L., and Wickens C. D., Eds. (2008–2009), Special Issue on Synthetic Vision Displays, *International Journal of Aviation Psychology*, Vol. 38, No. 4; Vol. 39, No. 1.

Raby, M., and Wickens, C. D. (1994), "Strategic Workload Management and Decision Biases in Aviation," *International Journal of Aviation Psychology*, Vol. 4, No. 3, pp. 211–240.

Rasmussen, J. (1981), "Models of Mental Strategies in Process Control," in *Human Detection and Diagnosis of System Failures*, J. Rasmussen and W. Rouse, Eds., Plenum, New York.

Rasmussen, J. (1986), *Information Processing and Human–Machine Interaction: An Approach to Cognitive Engineering*, North-Holland, New York.

Rasmussen, J., Pejtersen, A.-M., and Goodstein, L. (1995), *Cognitive Engineering: Concepts and Applications*, Wiley, New York.

Reason, J. (1990), *Human Error*, Cambridge University Press, New York.

Reason, J. (1997), *Managing the Risks of Organizational Accidents*, Ashgate, Brookfield, VT.

Reason, J. (2008), *The Human Contribution*, MIT Press, Cambridge, MA.

Reder, L. M. (1996), *Implicit Memory and Metacognition*, Lawrence Erlbaum Associates, Mahwah, NJ.

Regan, M., Lee, J. D., and Young, K. (2009), *Driver Distraction: Theories, Effects and Mitigation*, CRC Press, Boca Raton, FL.

Reicher, G. M. (1969), "Perceptual Recognition as a Function of Meaningfulness of Stimuli Materials," *Journal of Experimental Psychology*, Vol. 81, pp. 275–280.

Remington, R. W., Johnston, J. C., Ruthruff, E., Gold, M., and Romera, M. (2000), "Visual Search in Complex Displays: Factors Affecting Conflict Detection by Air Traffic Controllers," *Human Factors*, Vol. 42, No. 3, pp. 349–366.

Rensink, R. A. (2002), "Change Detection," *Annual Review of Psychology*, Vol. 53, pp. 245–277.

Robertson, G., Czerwinski, M., Fisher, D., and Lee, B. (2009), "Human Factors of Information Visualization," *Reviews of Human Factors & Ergonomics*, Vol. 5, pp. 41–81.

Rogers, D., and Monsell, S. (1995), "Costs of a Predictable Switch between Simple Cognitive Tasks," *Journal of Experimental Psychology: General*, Vol. 124, pp. 207–231.

Roscoe, S. N. (1968), "Airborne Displays for Flight and Navigation," *Human Factors*, Vol. 18, pp. 321–332.

Saks, M. J., Risinger, D. M., Rosenthal, R., and Thompson, W. C. (2003), "Context Effects in Forensic Science: A Review and Application of the Science of Science to Crime Laboratory Practice in the United States," *Science & Justice*, Vol. 43, pp. 77–90.

Salvucci, D., and Taatgen, N. A. (2011), *The Multi-Tasking Mind*, Oxford University Press, Oxford.

Sanchez, C. A., and Wiley, J. (2009), "To Scroll Or Not to Scroll: Scrolling, Working Memory Capacity, and Comprehending Complex Texts," *Human Factors*, Vol. 51, No. 5, pp. 730–738.

Sanders, M. S., and McCormick, E. J. (1993), *Human Factors in Engineering and Design*, 7th ed., McGraw-Hill, New York.

Sanderson, P. M. (1989), "The Human Planning and Scheduling Role in Advanced Manufacturing Systems: An Emerging Human Factors Domain," *Human Factors*, Vol. 31, pp. 635–666.

Sarter, N. (2007), "Multiple Resource Theory as a Basis for Multi-Modal Interface Design," in *Attention: From Theory to Practice*, A. F. Kramer, D. A. Wiegmann, and A. Kirlik, Eds., Oxford University Press, Oxford.

Schank, R. C., and Abelson, R. (1977), *Scripts, Plans, Goals, and Understanding*, Lawrence Erlbaum Associates, Mahwah, NJ.

Schneider, W., and Shiffrin, R. (1977), "Controlled and Automatic Human Information Processing, I: Detection, Search, and Attention," *Psychological Review*, Vol. 84, pp. 1–66.

Schneiderman, B. (2009), "Creativity Support Tools: A Grand Challenge for HCI Researchers," in *Engineering the User Interface*, M. Redondo, Ed., Springer, London, pp. 1–9.

Scholl, B. (2007), *Object Perception*, MIT Press, Cambridge, MA.

Schraw, G., and Moshman, D. (1995), "Metacognitive Theories," *Educational Psychology Review*, Vol. 7, No. 4, pp. 351–371.

Schutte, P. C., and Trujillo, A. C. (1996), "Flight Crew Task Management in Nonnormal Situations," in *Proceedings of the 40th Annual Meeting of the Human Factors and Ergonomics Society*, Human Factors and Ergonomics Society, Santa Monica, CA.

Schwaninger, A., and Wales, A. W. J. (2009), "One Year Later: How Screener Performance Improves in X-ray Luggage Search with Computer-Based Training," in *Contemporary Ergonomics*, P. D. Bust, Ed., Taylor and Francis, London, pp. 381–389.

Schwarz, N., and Vaughan, L. (2002), "The Availability Heuristic Revisited," in *Heuristics and Biases*, T. Gilovich, D. Griffin, and D. Kahneman, Eds., Cambridge University Press, Cambridge.

Senders, J. W. (1980), "Visual Scanning Processes," unpublished doctoral thesis, University of Tilburg, Tilburg, The Netherlands.

Shah, P., and Iyiri, A. (2005), *Handbook of Visual Spatial Thinking*, Cambridge University Press, Cambridge.

Shalin, V. L., and Bertram, D. A. (in press), "Functions of Expertise in a Medical Intensive Care Unit," *Journal of Experimental and Theoretical Artificial Intelligence: Special Issue on Expertise*.

Shen, M., Carswell, M., Santhanam, R., and Bailey, K. (in press),"Can Decision Makers Be Advised to Choose Appropriate Display Formats?"

Shiffrar, M., and Freyd, J. J. (1990), "Apparent Motion of the Human Body," *Psychological Science*, Vol. 1, No. 4, pp. 257–264.

Shinar, D., Rockwell, T. H., and Malecki, J. A. (1980), "The Effects of Changes in Driver Perception on Road Curve Negotiation," *Ergonomics*, Vol. 23, pp. 263–275.

Simons, D., and Levin, D. (1997), "Change Blindness," *Trends in Cognitive Science*, Vol. 1, pp. 261–267.

Smallman, H. S., and St. John, M. S. T. (2005), "Naive Realism: Misplaced Faith in Realistic Displays," *Ergonomics in Design*, Vol. 13, No. 3, pp. 6–13.

Smith, K. U. (1962), *Delayed Sensory Feedback and Balance*, Saunders, Philadelphia.

Smith, P., Bennett, K., and Stone, R. (2006), "Representational Aiding to Support Problem Solving Tasks,"in *Reviews of Human Factors and Ergonomics*, Vol. 2, R. C. Williges, Ed., Human Factors Society, Santa Monica, CA.

Smith, S. M. (1995), "Getting Into and Out of Mental Ruts: A Theory of Fixation, Incubation, and Insight," in *The Nature of Insight*, R. J. Sternberg and J. F. Davidson, Eds., MIT Press, Cambridge, MA, pp. 229–251.

Sorkin, R. (1989), "Why Are People Turning Off Our Alarms?" *Human Factors Bulletin*, Vol. 32, pp. 3–4.

Sorkin, R. D., and Woods, D. D. (1985), "Systems with Human Monitors, a Signal Detection Analysis," *Human–Computer Interaction*, Vol. 1, pp. 49–75.

St. Johns, M., and Manes, D. I. (2002), :"Making Unreliable Automation Useful," in*Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society*, Human Factors and Ergonomics Society, Santa Monica, CA.

St. John, M., Smallman, H. S., and Manes, D. I. (2005), "Recovery from Interruptions to a Dynamic Monitoring Task: The Beguiling Utility of Instant Replay," in *Human Factors and Ergonomics Society Annual Meeting*

*Proceedings*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 473–477.

Stanton, N., Ed. (1994), *Human Factors in Alarm Design*, Taylor & Francis, Bristol, PA.

Stelzer E. M., and Wickens, C. D. (2006). "Pilots Strategically Compensate for Display Enlargements in Surveillance and Flight Control Tasks," *Human Factors*, Vol. 48, No. 1, pp. 166–181.

Sternberg, S. (1969), "The Discovery of Processing Stages: Extension of Donders' Method," *Acta Psychologica*, Vol. 30, pp. 276–315.

Sternberg, R. J., and Powell, J. S. (1983), "Comprehending Verbal Comprehension," *American Psychologist*, Vol. 38, pp. 878–893.

Strayer, D. L., Wickens, C. D., and Braune, R. (1987), "Adult Age Differences in the Speed and Capacity of Information Processing, II: An Electrophysiological Approach," *Psychology and Aging*, Vol. 2, pp. 99–110.

Streeter, L. A., Vitello, D., and Wonsiewicz, S. A. (1985), "How to Tell People Where to Go: Comparing Navigational Aids," *International Journal on Man–Machine Studies*, Vol. 22, pp. 549–562.

Summala, H. (2000), "Brake Reaction Times and Driver Behavior Analysis," *Transportation Human Factors*, Vol. 2, pp. 217–226.

Taleb, N. (2007), *The Black Swan*.

Tanner, W. P., and Swets, J. A. (1954), "A Decision-Making Theory of Visual Detection," *Psychological Review*, Vol. 61, pp. 401–409.

Taylor, H. Brunye, T., and Taylor, S. (2008), "Spatial Mental Representation: Implications for Navigation System Design," in) *Reviews of Human factors & Ergonomics*, Vol.4, C. M. Carswell, Ed., Human Factors and Ergonomics Society, Santa Monica, CA.

Tenney, Y., and Pew, R. W., (2006), "Situation Awareness Catches On: What? So What? Now What?" in *Reviews of Human Factors & Ergonomics*, Vol. 3, R. Williges, Ed., Human Factors and Ergonomics Society, Santa Monica, CA.

Thorndyke, P. W. (1980, December), "Performance Models for Spatial and Locational Cognition," Technical Report R-2676-ONR, Rand, Washington, DC.

Thorndyke, P. W., and Hayes-Roth, B. (1982), "Differences in Spatial Knowledge Acquired from Maps and Navigation," *Cognitive Psychology*, Vol. 14, pp. 560–589.

Trafton, G., and Monk, C. (2007), "Task Interruptions," in *Reviews of Human Factors & Ergonomics*, Vol. 3, D. Boehm-Davis, Ed., Human Factors Society, Santa Monica, CA.

Trafton, G., Altman, E., Brock, D., and Mintz, F. E. (2003), "Preparing to Resume an Interrupted Tasks: Effects of Prospective Goal Encoding and Retrospective Rehearsal," *International Journal of Human Computer Studies*, Vol. 58, pp. 582–602

Travis, D. (1991), *Effective Color Displays: Theory and Practice*, Academic, New York.

Treisman, A. (1986), "Properties, Parts, and Objects," in *Handbook of Perception and Human Performance*, Vol. II, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., Wiley, New York, pp. 31-1–35-70.

Treisman, A. (1993), "The Perception of Features and Objects," in *Attention, Selection, Awareness, and Control: A Tribute to Donald Broadbent*, A. D. Baddeley and L. Weiskrantz, Eds., Clarendon, Oxford.

Treisman, A., Kahneman, D., and Burkell, J. (1983), "Perceptual Objects and the Cost of Filtering," *Perception and Psychophysics*, Vol. 33, pp. 527–532.

Tulga, M. K., and Sheridan, T. B. (1980), "Dynamic Decisions and Workload in Multitask Supervisory Control," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 10, pp. 217–232.

Tulving, E., Mandler, G., and Baumal, R. (1964), "Interaction of Two Sources of Information in Tachistoscopic Word Recognition," *Canadian Journal of Psychology*, Vol. 18, pp. 62–71.

Tversky, A., and Kahneman, D. (1974), "Judgment under Uncertainty: Heuristics and Biases," *Science*, Vol. 185, pp. 1124–1131.

Veenman, M. V. J., Hout-Wolters, van B. H. A. M., and Afflerbach, P. (2006), "Metacognition and Learning: Conceptual and Methodological Considerations," *Metacognition and Learning*, Vol. 1, pp. 3–14.

Vicente, K. J. (1999), *Cognitive Work Analysis*, Lawrence Erlbaum Associates, Mahwah, NJ.

Vicente, K. J. (2002), "Ecological Interface Design: Progress and Challenges," *Human Factors*, Vol. 44, pp. 62–78.

Vicente, K., and Rasmussen, J. (1992), "Ecological Interface Design: Theoretical Foundations," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 22, pp. 589–606.

Vicente, K. J., Christofferson, K., and Pereklita, A. (1995), "Supporting Operator Problem-Solving through Ecological Interface Designs," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 25, No. 4, pp. 529–545.

Vu, K., and Proctor, R. W. (2003), "Naive and Experienced Judgments of Stimulus-Response Compatibility: Implications for Interface Design," *Ergonomics*, Vol. 46, No. 3, pp. 169–187.

Wall, S. A., and Brewster, S. A. (2003),"Assessing Haptic Properties for Data Representation," paper presented at theConference on Human Factors in Computing Systems: CHI '03 extended abstracts on Human factors in computing systems, April 5–10, Ft. Lauderdale, FL, pp. 858–859.

Wallsten, T. S., Ed. (1980), *Cognitive Processes in Choice and Decision Behavior*, Lawrence Erlbaum Associates, Mahwah, NJ.

Wang, X., Zhang, Q., Han, Q., Yang, R., Carswell, M., Seales, B., and Sutton, E. (2009), "Endoscopic Video Texture Mapping on Pre-Built 3D Anatomical Objects without Camera Tracking," *IEEE Transactions on Medical Imaging*, Vol. 7, No. 7, pp. 1–12.

Warm, J., Ed. (1984), *Sustained Attention in Human Performance*, Wiley, New York.

Warm, J., Parasuraman, R., and Matthews, G. (2008), "Vigilance Requires Hard Mental Work and Is Stressful," *Human Factors*, Vol. 50, No. 3, pp. 433–441.

Warren, D. H., Rossano, M. J., and Wear, T. D. (1990), "Perception of Map–Environment Correspondence: The Roles of Features and Alignment," *Ecological Psychology*, Vol. 2, pp. 131–150.

Wetherell, A. (1979), "Short-Term Memory for Verbal and Graphic Route Information," in *Proceedings of the 23rd Annual Meeting of the Human Factors Society*, Human Factors Society, Santa Monica, CA.

Whitaker, L. A., and CuQlock-Knopp, V. G. (1995), "Human Exploration and Perception in Off-Road Navigation," in *Local Applications of the Ecological Approach to Human–Machine Systems*, P. A. Hancock, J. M. Flach,

J. Caird, and K. J. Vicente, Eds., Lawrence Erlbaum Associates, Mahwah, NJ.

Wickelgren, W. A. (1964), "Size of Rehearsal Group in Short-Term Memory," *Journal of Experimental Psychology*, Vol. 68, pp. 413–419.

Wickelgren, W. A. (1977), "Speed–Accuracy Tradeoff and Information Processing Dynamics," *Acta Psychologica*, Vol. 41, pp. 67–85.

Wickens, C. D. (1984), *Engineering Psychology and Human Performance*, 1st ed., Charles Merrill, Columbus, OH.

Wickens, C. D. (1986), "The Effects of Control Dynamics on Performance," in *Handbook of Perception and Performance*, Vol. II, K. Boff, L. Kaufman, and J. Thomas, Eds., Wiley, New York, pp. 39-1–39-60.

Wickens, C. D. (1991), "Processing Resources and Attention," in *Multiple Task Performance*, D. Damos, Ed., Taylor & Francis, London.

Wickens, C. D. (1992), *Engineering Psychology and Human Performance*, 2nd ed., HarperCollins, New York.

Wickens, C. D. (1993), "Cognitive Factors in Display Design," *Journal of the Washington Academy of Sciences*, Vol. 83, No. 4, pp. 179–201.

Wickens, C. D. (1999), "Frames of Reference for Navigation," in *Attention and Performance*, Vol. 16, D. Gopher and A. Koriat, Eds., Academic, Orlando, FL, pp. 113–144.

Wickens, C. D. (2000), "The When and How of Using 2-D and 3-D Displays for Operational Tasks," in *Proceedings of the IEA2000/HFES2000 Congress*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 3-403–3-406.

Wickens, C. D. (2002), "Multiple Resources and Performance Prediction," *Theoretical Issues in Ergonomic Science*, Vol. 3, No. 2, pp. 159–177.

Wickens, C. D. (2003a), "Aviation Displays," in *Principles and Practices of Aviation Psychology*, P. Tsang and M. Vidulich, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 147–199.

Wickens, C. D. (2003b), "Pilot Actions and Tasks: Selections, Execution, and Control," in *Principles and Practices of Aviation Psychology*, P. Tsang and M. Vidulich, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 239–263.

Wickens, C. D. (2005), "Multiple Resource Time Sharing Model," in *Handbook of Human Factors and Ergonomics Methods*, N. A. Stanton, E. Salas, H. W. Hendrick, A. Hedge, and K. Brookhuis, Eds., Taylor & Francis, London, pp. 40-1–40-7.

Wickens, C. D. (2007), "Attention to Attention and Its Applications: A Concluding View," in *Attention: From Theory to Practice*, A. F. Kramer, D. A. Wiegmann, and A. Kirlik, Eds., Oxford University Press, Oxford.

Wickens, C. D. (2008), "Situation Awareness: Review of Mica Endsley's 1995 Articles on Situation Awareness Theory and Measurement," *Human Factors*, Vol. 50, No. 3, pp. 397–403.

Wickens, C. D. (2009), "Keynote Address: The Psychology of Aviation Surprise," in *Proceedings 15th International Symposium on Aviation Psychology*, P. Tsang and M. Vidulich, Eds., Wright State University, Dayton, OH.

Wickens, C. D., and Alexander, A. L. (2009), "Attentional Tunneling and Task Management in Synthetic Vision Displays," *International Journal of Aviation Psychology*, Vol. 19, pp. 182–189.

Wickens, C. D., and Andre, A. D. (1990), "Proximity Compatibility and Information Display: Effects of Color, Space, and Objectness on Information Integration," *Human Factors*, Vol. 32, pp. 61–77.

Wickens, C. D., and Carswell, C. M. (1995), "The Proximity Compatibility Principle: Its Psychological Foundation and Its Relevance to Display Design," *Human Factors*, Vol. 37, No. 3, pp. 473–494.

Wickens, C. D., and Hollands, J. (2000), *Engineering Psychology and Human Performance*, 3rd ed., Prentice-Hall, Upper Saddle River, NJ.

Wickens, C. D., and McCarley, J. M. (2008), *Applied Attention Theory*, CRC Press, Boca Raton, FL.

Wickens, C. D., and Prevett, T. (1995), "Exploring the Dimensions of Egocentricity in Aircraft Navigation Displays: Influences on Local Guidance and Global Situation Awareness," *Journal of Experimental Psychology: Applied*, Vol. 1, No. 2, pp. 110–135.

Wickens, C. D., Sandry, D., and Vidulich, M. (1983), "Compatibility and Resource Competition between Modalities of Input, Output, and Central Processing," *Human Factors*, Vol. 25, pp. 227–248.

Wickens, C. D., Vidulich, M., and Sandry-Garza, D. (1984), "Principles of S–C–R Compatibility with Spatial and Verbal Tasks: The Role of Display-Control Location and Voice-Interactive Display-Control Interfacing," *Human Factors*, Vol. 26, pp. 533–543.

Wickens, C. D., Zenyuh, J., Culp, V., and Marshak, W. (1985), "Voice and Manual Control in Dual Task Situations," in *Proceedings of the 29th Annual Meeting of the Human Factors Society*, Human Factors Society, Santa Monica, CA.

Wickens, C. D., Todd, S., and Seidler, K. S. (1989), "Three-Dimensional Displays: Perception, Implementation, and Applications," CSERIAC SOAR 89-001, Crew System Ergonomics Information Analysis Center, Wright-Patterson AFB, OH.

Wickens, C. D., Vincow, M. A., Schopper, A. W., and Lincoln, J. E. (1997), "Computational Models of Human Performance in the Design and Layout of Controls and Displays," CSERIAC SOAR 97-22, Crew System Ergonomics Information Analysis Center, Wright Patterson AFB, OH.

Wickens, C. D., Gempler, K., and Morphew, M. E. (2000), "Workload and Reliability of Predictor Displays in Aircraft Traffic Avoidance," *Transportation Human Factors Journal*, Vol. 2, No. 2, pp. 99–126.

Wickens, C. D., Goh, J., Helleberg, J., Horrey, W., and Talleur, D. A. (2003), "Attentional Models of Multitask Pilot Performance Using Advanced Display Technology," *Human Factors*, Vol. 45, No. 3, pp. 360–380.

Wickens, C. D., Ververs, P., and Fadden, S. (2004), "Head-Up Display Design," in *Human Factors for Civil Flight Deck Design*, D. Harris, Ed., Ashgate, Avebury, VT, pp. 103–140.

Wickens, C. D., Vincow, M., and Yeh, M. (2005), "Design Applications of Visual Spatial Thinking: The Importance of Frame of Reference," in *Handbook of Visual Spatial Thinking*, A. Miyaki and P. Shah, Eds., Oxford University Press, Oxford.

Wickens, C. D., McCarley, J. S., Alexander, A. L., Thomas, L. C., Ambinder, M., and Zheng, S. (2008), "Attention-Situation Awareness (A-SA) Model of Pilot Error," in *Human Performance Models in Aviation*, D. Foyle and B. Hooey (Eds.), Taylor & Francis, Boca Raton, FL.

Wickens, C. D. Hooey, B., Gore, B. F., Sebok, A, and Koenicke, C. (2009a), "Identifying Black Swans in NextGen: Predicting Human Performance in Off-Nominal Conditions," *Human Factors*, Vo. 51, pp. 638–651.

Wickens, C., Rice, S., Keller, M. D., Hutchins, S., Hughes. J., and Klayton, K. (2009b), "False Alerts in Air Traffic Control Conflict Alerting System: Is There A Cry Wolf Effect"? *Human Factors.*, Vol. 51, pp. 446–452.

Wickens, C. D., Keller, J., and Small, R. (2010), "Development of the Frame of Reference Transformation Tool (FORT)," in *Proceedings of the 54th Annual Meeting of the Human Factors and Ergonomics Society*, Human Factors Society, Santa Monica, CA.

Wickens, L. D., Alexander, A. L., Martens, M., and Ambinder, M. (2004), "The Role of Highlighting in Visual Search through Maps," *Spatial Vision*, Vol. 37, pp. 373–388.

Wickens, T. D. (2002), *Signal Detection Theory*, Freeman, San Francisco.

Wilkinson, R. T. (1964), "Artificial 'Signals' as an Aid to an Inspection Task," *Ergonomics*, Vol. 7, pp. 63–72.

Wolf, L. D., Potter, P., Sedge, J., Bosserman, S., Grayson, D., and Evanoff, B. (2006), "Describing Nurses' Work: Combining Quantitative and Qualitative Analysis," *Human Factors*, Vol. 48, pp. 5–14.

Wolfe, J. M. (2007), "Guided Search 2.0. A Revised Model of Visual Search," *Psychonomics Bulletin and Review*, Vol. 1, No. 2, pp. 202–238.

Wolfe, J. M., and Horowitz, T. S. (2004), "What Attributes Guide the Deployment of Visual Attention and How Do They Do It?" *Nature Neuroscience*, Vol. 5, pp. 1–7

Wolfe, J. M., Horowitz, T. S., and Kenner, N. M. (2005), "Rare Items Often Missed in Visual Searches," *Nature*, Vol. 435, pp. 439–440.

Woods, D., Wise, J., and Hanes, L. (1981), "An Evaluation of Nuclear Power Plant Safety Parameter Display Systems," in *Proceedings of the 25th Annual Meeting of the Human Factors Society*, Human Factors Society, Santa Monica, CA.

Woods, D. D., Johannesen, L. J., Cook, R. I., and Sarter, N. B. (1994), "Behind Human Error: Cognitive Systems, Computers, and Hindsight," CSERIAC #SOAR 94-01, Crew System Ergonomics Information Analysis Center, Wright-Patterson AFB, OH.

Yantis, S. (1993), "Stimulus Driven Attentional Capture," *Current Directions in Psychological Science*, Vol. 2, pp. 156–161.

Yates, J. F. (1990), *Judgment and Decision Making*, Prentice-Hall, Englewood Cliffs, NJ.

Yeh, M., and Wickens, C. D. (2001), "Attentional Filtering in the Design of Electronic Map Displays: A Comparison of Color-Coding, Intensity Coding, and Decluttering Techniques," *Human Factors*, Vol. 43, No. 4, pp. 543–562.

Yeh, M., Merlo J. L., Wickens, C. D., and Brandenburg, D. L. (2003), "Head Up Versus Head Down: The Costs of Imprecision, Unreliability, and Visual Clutter on Cue Effectiveness for Display Signaling," *Human Factors*, Vol. 45, No. 3, pp. 390–407.

Yntema, D. (1963), "Keeping Track of Several Things at Once," *Human Factors*, Vol. 6, pp. 7–17.

Yost, B., and North, C. (2006), "The Perceptual Scalability of Visualization," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 12, No. 5, pp. 837–844.

Zheng, B., Janmohamed, Z., and MacKenzie, C. L. (2003), "Reaction Times and the Decision Making Process in Endoscopic Surgery," *Surgical Endoscopy*, Vol. 17, pp. 1475–1480.

Zsambok, C., and Klein, G. (1997), *Naturalistic Decision Making*, Lawrence Erlbaum Associates, Mahwah, NJ.

# CROSS-CULTURAL DESIGN

**Tom Plocher**
**Honeywell International**
**Golden Valley, Minnesota**


**Pei-Luen Patrick Rau**
**Tsinghua University**
**Beijing, China**


**Yee-Yin Choong**
**National Institute of Standards and Technology**
**Gaithersburg, Maryland**

## 1 INTRODUCTION

Internet and ubiquitous mobile access to information and services, together with the globalization of companies, has transformed our world into a truly international marketplace. Users cross international boundaries to access an enormous range of information. Products and services are developed in one country and traded in another. In this new international marketplace, the cultural characteristics of users have become increasingly important. There is an expectation that products and services marketed globally will be sufficiently compatible with local cultures to ensure a highly usable and satisfying user experience.

People with different cultural backgrounds think and behave differently. Examined closely, these differences go way beyond speaking and writing in different languages. Cultural differences are present in values and attitudes, social relationships, communication styles, visual preferences, and cognitive styles. All of these potentially affect the design of highly usable and satisfying user interfaces for users from different cultures. Anthropometry also varies across different world populations, which needs to be considered in the design of physical interfaces and devices for global use. Finally, we are just beginning to discover that users in one culture may emphasize and place value on

different characteristics of usability than those in another. Understanding those differences is essential to designing and bringing a product to market successfully in different world cultures.

Designing user interfaces and products for different cultures also affects human factors design methodology. Careful user needs research conducted in the target cultures at the very beginning of a development will ensure that the product, service, or application concept and requirements will support the tasks and lifestyle of the intended users and be compatible with their environment. The methodology for user needs research must be adapted to the local customs, attitudes, and behaviors of the culture being studied. Recent research on cross-cultural usability testing has revealed a host of issues inherent in evaluating user interfaces across cultures. Discovering the most important usability problems during an evaluation requires an understanding of these issues and willingness to adapt the test methodology to local cultures.

This chapter provides human–computer interaction (HCI) researchers and human factors design practitioners a comprehensive perspective on cross-cultural issues in user interface design. It first presents a definition of culture and various dimensions that are relevant to human information processing and hence user interface design. It then reviews the literature on cross-cultural

usability within this framework of cultural dimensions and human information processing. Where supported by the literature, design guidelines are presented for anthropometry, languages and format, presentation, graphics, cultural preferences, information architecture, searching, and interaction. Recent cross-cultural design research for a variety of paradigms—Web and mobile services design and tangible interfaces—is presented. Suggestions for conducting international user needs research and usability testing are also provided.

## 2 THEORY AND METHODOLOGY

### 2.1 Cross-Cultural Psychology

#### 2.1.1 Framework

The metamodels of culture (Trompenaars, 1993; Hoft, 1996; Stewart and Bennett, 1991; del Galdo and Nielsen, 1996) almost universally conclude that a significant portion of what can be called "culture" is embodied in the psychology of people. This includes (1) their values and attitudes, (2) preferred communication style, and (3) cognitive style. A look at research into these core dimensions of culture provides some insight into how these dimensions might affect user interactions with technology and hence the design of user interfaces. The role played by values and attitudes and the preferred communication style in the framework of culture are discussed below. Culture and cognitive style are discussed in Section 2.1.2.

*Values and Attitudes* Hofstede (1991) believed that patterns of thinking, feeling, and acting are mental programs, or, as he dubbed them, software of the mind. These mental programs vary as much as the social environments in which they were acquired. The collective programming of the mind is what distinguishes the values and attitudes of one cultural group from another.

In his classic study of cultural variations in workplace values and attitudes, *Culture and Organizations: Software of the Mind*, Hofstede (1991) reports and interprets his findings from administering his Value Survey Module (VSM) to over 116,000 people in 50 countries, mostly white-collar workers at IBM. Hofstede statistically extracted four dimensions along which his subjects in different national cultures systematically differed: (1) power distance, (2) uncertainty avoidance, (3) individualism–collectivism, and (4) masculinity–femininity. Later, he added a fifth dimension, long-term orientation. These are described briefly below:

- **Power Distance** Power distance is defined as the extent to which the less powerful members of institutions and organizations within a society accept that power is distributed unequally. Cultures in which high power distance is the norm tend to have highly demarcated levels of hierarchy. People from the lower rungs of the hierarchy have considerable difficulty in crossing the "boundaries." Malaysia, Philippines, India,

and Arab countries were examples of high-power-distance cultures. Austria, Israel, Ireland, New Zealand, and the Scandinavian countries all ranked very low on Hofstede's power distance scale. The United States ranked in the neutral range on this dimension.

- **Uncertainty Avoidance** Uncertainty avoidance is the extent to which members of organizations in a society are threatened by uncertainty, ambiguity, and unstructured situations. High-uncertainty-avoidance cultures tend to have a greater need for formal rules and have less tolerance for people or groups with deviant ideas or behaviors. Latin American countries, Greece, Portugal, Turkey and Belgium in Europe, and Japan and South Korea in Asia scored high on uncertainty avoidance in Hofstede's research. The Scandinavian countries, Ireland and Great Britain in Europe, Singapore, Hong Kong, and Malaysia in Asia showed the lowest uncertainty avoidance. The United States scored in the middle of the countries sampled.

- **Individualism–Collectivism** Individualism describes a society in which the ties between individuals are loose. Everyone is expected to look after him or herself. Collectivism exists in societies in which people are integrated into strong cohesive groups, which serve to protect them throughout their life. The group receives loyalty in return. The United States, Great Britain, Australia, and Canada had the highest individualism scores in Hofstede's research. The most collectivist countries were in Latin America. Certain Asian countries such as Pakistan, Indonesia, South Korea, and Taiwan also were strongly collectivist.

- **Masculinity–Femininity** Masculinity is found in a society in which social gender roles are very distinct. Men are expected to be assertive, tough, and oriented around material success. Women are supposed to be modest, nurturing, and concerned with the quality of life. In feminine cultures, the gender roles overlap. It is OK for both men and women to show traits of nurturing and concern for quality of life. Japan had the highest score for masculinity in Hofstede's study, followed by Austria, Switzerland, Italy and Germany in Europe, and numerous Caribbean countries. The United States ranked on the masculinity side of the middle range. The Scandinavian countries and The Netherlands were the most feminine countries studied by Hofstede.

- **Long-Term versus Short-Term Orientation** Long-term orientation is found in a society that is oriented toward future rewards. Perseverance and thrift are valued. Cultures with short-term orientation promote the virtues of the past and present. These include respect for tradition, preserving "face," and fulfilling social obligations. China, Hong Kong, Taiwan, Japan, and South Korea all are good examples for long-term oriented cultures. In Hofstede's survey, Pakistan,

Nigeria, Philippines, Canada, Great Britain, and the United States were among the most short-term-oriented cultures.

Since Hofstede's results were originally published, they have been widely used in international management to explain business-related behaviors in different countries (e.g., Trompenaars, 1993). They have been used to understand human-induced causes of airline crashes (Krishnan et al., 1999; Helmreich and Merritt, 1998) and medical errors (Helmreich and Merritt, 1998). Krishnan et al. (1999) proposed a range of cockpit user interface enhancements that would potentially counter certain culturally linked attitudinal tendencies that can lead to errors in the cockpit. Principles relating Hofstede's cultural dimensions and the affect evoked by the visual composition of website content have been proposed by Gould (2001). Finally, much has been written about the implications of these value and attitude dimensions for user interface design (Marcus, 2001).

There is no question that Hofstede's cultural framework can illuminate certain aspects of user interface design, particularly those of an affective or social nature. But as these relationships are postulated and applied to design, one needs to be sensitive to certain issues in applying it to new contexts. First, Hofstede believed that his framework tapped into some of the most fundamental and deeply engrained aspects of national culture. However, he predicted that several of his cultural dimensions, namely uncertainty avoidance and individualism–collectivism, might be influenced by social, economic, and political change. He envisioned that populations might slowly change over time along these attitudinal dimensions. Results from an application of Hofstede's VSM in China (Plocher et al., 2001) provide some evidence for this. Their engineering student subjects at Tsinghua University scored about as expected on four of the Hofstede dimensions. However, their score on the scale of individualism–collectivism was strongly in the opposite direction (strong individualism) to that predicted by the literature. The authors speculated that this particular result may reflect the effects of the rather dramatic social and political upheaval in China over the past 50 years caused by the Chinese revolution and Communist era and, more recently, Western social and economic influences. Young professionals in China may be eschewing the traditional and Communist-era collectivist Chinese values and attitudes for a more Western-like individualist orientation.

In a major research project aimed at understanding national, organizational, and professional influences on the team-related behaviors of commercial airline pilots, Helmreich and Merritt (1998) found the Hofstede (1991) framework to be extremely useful. Many of the airline crashes they reviewed had a significant cultural component when interpreted within the framework provided by Hofstede. However, Helmreich and Merritt sometimes observed behavior that appeared to contradict their expectations based on national culture alone. They concluded that national culture is only one influence on pilot behavior. In some situations the value and attitudinal dimensions of national culture give way to overriding influences imposed by the culture of the organization (e.g., the airline) or by the piloting profession itself. Hofstede's framework provides valuable insights into cross-cultural behaviors but must be applied with an understanding of other influential factors at work in the organization and profession.

Finally, as Bosland (1985) has pointed out, we must remember that Hofstede's cultural dimensions are characteristics of societies that share the same culture. The variables reflect the dominant values of the majority of people living in a given culture. But they do not necessarily reflect the values of every individual person in that culture. Individuals within one culture will have a greater likelihood of holding certain values and attitudes about work than their counterparts in another culture and thus will have a greater likelihood of reacting to events in certain ways. However, within both cultures there will be significant variation, even to the extent of some overlap.

### Preferred Communication Style: Use of Context

Context refers to the amount of information packed into a specific instance of communication (Hall, 1976, 1990). It is one basis for describing communication styles. "High-context" communication style uses terse messages, short on background details. It assumes that the receiver of the message is familiar with the subject matter. "Low-context" communication style uses more lengthy or elaborate messages, contains a lot of background information on the subject of the communication, and assumes that the receiver of the message may not necessarily be familiar with the subject. When high- and low-context people attempt to communicate, misunderstanding or frustration often results. The low-context person wants more detail and background information than the high-context person is willing or able to provide. The high-context person is impatient listening to "information that she already knows" from a low-context person who is only presenting his usual complete and thorough message.

Hall believed that communication style, high or low context, was deeply rooted in culture. While there will be significant variation in style within any one culture, one style will tend to be dominant. Germans, Dutch, English, and Americans tend to prefer low-context communication. French, Italians, Spanish, Latin Americans, and Japanese prefer high-context communication. In their popular book, *Understanding Cultural Differences*, Hall and Hall (1990) describe and interpret the stereotypical business-related behaviors of French, Germans, and Americans in terms of communication style and discuss the kinds of conflicts and misunderstandings that can occur when high context meets low context.

Communication style also affects how people interact with information systems, particularly nonlinear, hypertext systems such as the web (Rau, 2001). High-context people browse information faster and require fewer links to find information than low-context users. However, high-context users also have a greater tendency to become disoriented and lose their sense of location and direction in hypertext. Low-context users

are slower to browse information and link more pages but are less inclined to get lost.

### 2.1.2 Cognition and Human Information Processing

People interact with the world around them by sensing and perceiving the stimuli presented to them, making sense of the information perceived, deciding if a response is needed, then executing the response. It is stated in many cognitive theories how humans perceive (Card et al., 1983), store, and retrieve information from short- and long-term memory (Shiffrin and Atkinson, 1969; Norman and Rumelhart, 1970; Shiffrin and Schneider, 1977; Wickens and Hollands, 2000), manipulate that information to make decisions and solve problems (Newell and Simon, 1972), and carry out responses. The stages in human information processing are represented in a general qualitative model by Wickens and Hollands (2000).

The physiological mechanism of human information processing is universal to all people and can be viewed as culturally independent. The organization and structure of the information at each stage are affected by the experience of each individual. One important factor that governs such experience is culture. As noted in the previous section, culture not only affects the values and attitudes held by people but also can affect people's cognition interacting with the world around them.

***Perception*** Meaningful interaction with the world requires pattern recognition. Reading, understanding speech, and distinguishing the familiar from the unfamiliar all require the recognition of patterns. Our brains organize and give meaning to the constant input of sensory messages through an active process of selecting, ordering, synthesizing, and interpreting.

Everyday objects, symbols, and gestures provide design inspiration and are commonly used in user interfaces, but they can be perceived differently in various parts of the world (Table 1). For example, while a U.S. rural mailbox has been widely used to represent the concept of an email account, it cannot be assumed that people from various cultures will perceive and recognize it as a mailbox. A common Japanese street mailbox looks like a U.S. trashcan (Fernandes, 1995). The use of hand gestures in symbol or icon design also is problematic. The same hand gesture can be perceived differently, sometimes the opposite, by people with different cultural backgrounds. "Thumb up" is well known as "fine" or "good going" in North America and much of Europe, but it is perceived as insulting in Australia (Axtell, 1991). The "thumb up" gesture is also used in counting. For example, in Germany, a person uses the upright thumb to signal "one," and in Japan, the upright thumb is used to signal "five."

***Cognitive Style*** From the 1980s, many researchers have studied the fundamental differences in cognitive behavior between people of different national cultures. The classic paper by Liu (1986) was the first attempt at describing a Chinese cognitive style and the experiential factors that shape Chinese cognitive style during development: the family order, the Chinese educational system, and the nature of the Chinese language. Hall (1984, 1989, 1990) and Hall and Hall (1990) wrote extensively about time cognition, how it was expressed in many different behaviors of daily life, and how it varied across national cultures. Later, Nisbett (2003) and his colleagues in Asia (Nisbett et al., 2001) developed a significant body of experimental evidence characterizing fundamental differences between Easterners and Westerners in reasoning style.

Perhaps the most clearly understood and documented differences in ways of thinking are the differences between Americans and Chinese. The cognitive style of Americans is inferential–categorical (functional), which means that they have a tendency to classify stimuli on the basis of functions or inferences made about the stimuli that are grouped together accordingly (Chiu, 1972). In contrast, Chinese people have a relational–contextual or thematic cognitive style. They tend to classify stimuli on the basis of interrelationships and thematic relationships (Chiu, 1972). The American way of thinking tends to be analytic, abstract, and imaginative. The Chinese way of thinking tends to be synthetic and concrete.

Nisbett theorizes that these cognitive differences result from the fundamental philosophical differences—Aristotelean versus Confucian—which permeate almost every aspect of these societies from childrearing to education to social structure. According to Nisbett, Westerners tend to be analytical–logical in reasoning style. Easterners tend to be holistic–dialectical. He describes these stereotypical differences in reasoning style in the ways (Nisbett, 2003) shown in Table 2.

It is illustrative to consider what can happen when a Western analytical thinker attempts to solve problems or work globally with an Eastern holistic thinker. Two individuals from different cultures might:

- Look at the same information or picture and draw different conclusions.
- Pay different attention to minor evidence and be more or less conclusive about their predictions of actions.

**Table 1 Guideline for Metaphors**

| Category | Guidelines | Supporting Research/ Best Practice |
| --- | --- | --- |
| Symbols and icons | Everyday objects, symbols, and gestures may have different meaning in different cultures. They should be used with an awareness of their meaning in the target culture. | Fernandes, 1995 |

**Table 2  Western and Eastern Reasoning Styles**

| Western Analytical–Logical Thought | Eastern Holistic–Dialectical Thought |
| --- | --- |
| Focus on objects, attributes, categories | Focus on the field surrounding objects; sensitive to covariation in the field; little relevance seen in categories |
| Apply rules based on the categories to predict and explain the objects' behavior | Little use of universal rules; behavior of an object is explained by situational forces and factors in the surrounding field |
| Find category learning easy | Find category learning difficult |
| Organize things functionally (focus on what an object ''does'') | To the extent they use categories, they prefer to organize things thematically (e.g., use the context or environment as a basis for identifying similarities) |
| Use formal logic for reasoning, making categories, and applying and justifying rules | Not much use of formal logic; prefer dialectical reasoning such as synthesis, transcendence, and convergence |
| Eager to resolve contradictions (logic); when logic and experiential knowledge are in conflict, adhere to formal logical rules | Less eager to resolve contradiction; prefer a logic that accepts contradiction |
| Interpret individual's behavior as a result of their disposition or personality | Interpret peoples' behavior as the result of situational pressures |

- Look at opposition leaders and predict different actions based on their preference for dispositional-versus- situational explanations.
- Come to different conclusions about the properties of an object when presented with identical evidence from two other seemingly familiar objects.
- Disagree or miscommunicate over how to group things.

**Organizing and Searching Information**  One basic mental process is that how people group things together into categories. Based on the similarities, people categorize objects perceived to have some certain characteristics. People often decide whether something belongs in a certain group by comparing it to the representative member of that category. Some categories seem to be universal across cultures. For example, facial expressions that signal basic emotions—happiness, sadness, anger, fear, surprise, and disgust—are widely agreed-upon across cultures. Likewise, there is widespread agreement across cultures about which colors are primary and which are secondary. The way people select and remember colors appear to be largely independent of both culture and language.

People categorize items of information, objects, and functions according to perceived similarities and differences. If the items have two or more attributes, then there is a basis for variation in how they are grouped together. Some people will emphasize a certain attribute and sort items into categories accordingly. Others will focus on another attribute and sort accordingly. Chiu (1972) found that Chinese prefer to categorize on the basis of interdependence and relationship, whereas Americans prefer to analyze the components of stimuli and to infer common features. The difference between the analytical and relational thinking styles is mainly based on how subjectivity is treated. The analytical style

separates subjective experience from the inductive process that leads to an objective reality, whereas the relational style of thinking rests heavily on experience and fails to separate the experiencing person from objective facts, figures, or concepts (Stewart and Bennett, 1991).

Choong (1996) conducted research which showed that different cultures often focus on different attributes of the same items or objects. For example, Americans tend to focus on functional attributes, whereas Chinese tend to focus on thematic attributes. As a result, Chinese and Americans tend to group items in fundamentally different ways. The experimental results provide insights for cross-cultural design. The results showed that Chinese and American users of an online department store performed better if the contents of the store were organized in a manner that was consistent with their natural way of organizing objects—functional for Americans and thematic for Chinese. In other words, Americans would prefer to see products in a department store organized by function: cleaning supplies, linens, and furniture. Chinese prefer to organize products by themes, in the case of a department store, the different rooms of a house: kitchen, bathroom, and bedroom. Rau et al. (2004) conducted a cross-cultural study to compare the impact of knowledge representation (abstract and concrete) and interface structure (functional and thematic) on Chinese and American performance. Their study provided additional evidence in line with results from previous studies that the Chinese employ a different thinking style from Americans. It also agreed with the previous study (Choong, 1996) that thematic interface structure was advantageous to Chinese users, especially when error rate is an important factor in task performance.

When performing searching tasks, the structure of information is very important. Zhao (2002) studied the effect of information structure on performance of information-acquiring tasks by people of monochronicity or polychronicity time behaviors (see next section). She first classified users as monochronic or polychronic

by means of a standard survey instrument. Then, she measured their speed of performance on information search and retrieval tasks using two different information structures, hierarchical and network. The hierarchical information structure was basically a tree. It placed information in categories and subcategories, with the vertical structure going up to six levels deep. Crossing over from one vertical branch of the tree to another was possible only at the topmost level. The network information structure allowed more flexible searching based on relationships between items of information, regardless of their place in a hierarchy of categories. The results showed that performance of the monochronic users performed significantly better using a hierarchical information structure. In contrast, polychronic users were significantly faster using a network information structure.

Finally, placing objects into categories may simply be less important for people from Eastern cultures. In a picture-sorting task Nawaz et al. (2007) found that Chinese made less use of specific categories and more use of the category "other" than Danish users.

So what does this mean for user interface design? Categories form the basis for the information architectures underlying user interfaces of software systems. The content and organization of menus in graphical user interfaces, links in websites, and file directories in most software applications are based on categories. Databases are built on domain ontologies that rigorously apply categories to organize the concepts and objects in a domain. Different cultures will have different ways of categorizing concepts and objects. Furthermore, the concepts, per se, may differ among cultures. Some concepts used in one culture may not even exist in another culture. What happens when a user from one culture, say, the United States, attempts to use a software application that is based on categories (e.g., an information architecture) developed in another culture, such as India? If the information architecture has not been carefully localized to American culture or carefully internationalized to remove the biases of the Indian culture, this will result in products with poor usability. The user will be frustrated as he or she attempts to locate functions in menus or information organized in an unfamiliar way by an original designer or engineer thinking with a different style.

In the design process, preferred categories and information structure can be determined by card sorting. Two methods can be used for card sorting, one is by hand with paper note cards, the other is by a computer program. Anderson et al. (2004) describe a method for computing "edit distance" from card sort data as a method for revealing the underlying categories and concepts. Their software tools for conducting edit distance analysis are available on the Web at http://www.cs.washington.edu/research/edtech/CardSorts/.

**Spatial Cognition**   Spatial cognition is the process through which individuals gain knowledge of objects and events linked to space (Gauvain, 1993; Mishra, 1997). Frake (1980) analyzed the use of absolute directions and contingent directions in Southeast Asia

and California. The conclusion shows that cultures influence the use of directions. The research of Spencer and Darvizeh (1983) concludes that the Iranian children gave more detailed object information on the way but less directional information compared with British children.

Ji et al. (2000) conducted rod-and-frame tests to detect the field dependence between Chinese and Americans. They showed that Chinese participants made more mistakes on the rod-and-frame test, reported stronger association between events, and were more responsive to differences in covariation, whereas American participants made few mistakes on the rod-and-frame test, indicating that they were less field dependent than their Chinese counterparts.

**Time Cognition**   Though time should be technically and objectively the same for everyone, Hall's (1984) classic ethnographic observations showed that different cultures have different attitudes toward time. He classified two ways in which people understand time: monochronic and polychronic time systems. The different attitudes toward time are reflected in many aspects of peoples' lives, including how they adhere to schedules, approach the tasks of their job, and cope with competing task demands (Bluedorn et al., 1999; Haase et al., 1979; Kaufman-Scarborough and Lindquist, 1999; Lindquist et al., 2001).

Monochronic time is dominant in Germany, the United Kingdom, the Netherlands, Finland, the United States, and Australia (Hall, 1984; Hall and Hall, 1990). Cultures with a monochronic time orientation treat time in a linear manner. Time is divided into segments that can be easily scheduled and "spent." Monochronic people prefer to follow clear rules and procedures. They prefer to work on one task at a time and are frustrated when other competing tasks disrupt that focus. Monochronic people have a narrower view of the overall situation or activity and may miss significant events related to the waiting tasks (e.g., alarms). Clear procedures are important to monochronic users. They are less inclined to invent procedures in new situations or where standard procedures are not available. Monochronic computer users search for information in hypertext in a deliberate and linear manner, making more links than polychronic users to find the same information (Rau, 2001). Hence, they are slower at searching hyperspaces than polychronic users.

Polychronic time is dominant in Italy, France, Spain, Brazil, and India (Hall, 1984; Hall and Hall, 1990). In contrast to monochronic people, polychronic people perceive time in a less rigid, more flexible way. Adhering to rules, procedures, and schedules are not that important to them. Polychronic users are more inclined to switch back and forth between tasks and applications (Zhang et al., 2004). They have a broader view of the overall situation or ongoing activity, but they are prone to task-switching errors. When they try to resume a task, they may "forget where they left off," resuming at the wrong place in the procedure or process. Standard procedures are less important to polychronic users, and they are more inclined to invent procedures to deal with new situations.

Time orientation, either monochronic or polychronic, is deeply rooted in a culture. Within any one culture, one style tends to be dominant, but there will be significant variation in time orientation and related behaviors. In another word, it is possible that time cognition can be influenced by factors other than national culture. Zhao et al. (2002) found that the natural or preferred time orientation of Chinese industrial workers (monochronic) was quite different from what they displayed on the job (polychronic). During a debriefing of the study, participants revealed that there were many factors at work in the Chinese industrial workplace and in society that simply made polychronic behavior more adaptive.

**Problem Solving and Decision Making**   Problem solving is the process by which people attempt to discover ways of achieving goals that do not seem readily attainable. Cross-cultural research on decision making finds that people of many different cultural groups may use different types of decision-making strategies. Americans may favor considering many possibilities, evaluating each possibility as a hypothesis, and then choosing the best one based on the available information. Other cultures high in uncertainty avoidance, like Chinese, may have a greater tendency to make judgments based on representativeness.

In the research of Yi and Park (2003), more than 800 college students from five countries, Korea, Japan, China, the United States, and Canada, joined the experiments. The conclusion found that cultural differences result in different types of decision making. Compared to Americans and Canadians, Korean students showed higher levels of cooperative decision making. However, Japanese students exhibited the lowest levels of cooperative decision making.

Geary et al. (1992) compared the performance of Chinese and American children on the calculation of simple addition. The experiment concludes that Chinese children solved three times more correctly than the Americans did. They also did it with greater speed, because Chinese children calculate by the strategies of direct retrieval and decomposition, whereas Americans depended mainly on counting.

**Language**   Every human society, however primitive in other terms, has a language. The ability to use language is perhaps the most profound indicator of the power of human cognition (Miller, 1981). Without language, our ability to remember, to reason, and to solve problems would be severely limited, since so much of human information processing and thinking occur at the abstract level of language symbols.

Language is key to meaningful communication among people. The communication will be most effective when the first languages, or "mother tongues," of the peoples of the world are used. There are between 3000 and 4000 spoken languages, with numbers ranging from many millions of speakers down to a few dozen or even fewer. There are hundreds of different written languages represented by scripts in use around the world.

Although a writing system is generally viewed as a system for representing a spoken language, it should

be noted that written language is not always a direct transcription of spoken language (Sampson, 1990). For example, in the Arabic-speaking world, vocabulary, grammar, and phonology between spoken and written varieties of Arabic are different. While it is possible to transcribe Arabic speech directly into Arabic script, the transcription will strike Arabic speakers as bizarre and unnatural.

The means to transcribe spoken words are different for different writing systems. For example, Chinese transcribes whole morphemes, the units of meaning. Each single syllabic character represents a unit of meaning(s) which can be a word itself or sometimes two or more characters form a word. The alphabetic systems such as English transcribe phonemes, the units of sound. There is distinction in accessibility of meanings between morphemic and alphabetical symbols. The meaning of Chinese characters is more manifest perceptually than the meaning of words in alphabetical systems (Hoosain, 1986).

There is evidence showing that language can influence thinking. Hoffman et al. (1986) asked bilingual English–Chinese speakers to read descriptions of individuals and to provide free interpretations of the individuals described. The descriptions were of characters exemplifying personality schemas with economical labels either in English or Chinese. Bilingual subjects thinking in Chinese used Chinese stereotypes in their free interpretations, whereas those thinking in English used English stereotypes. This indicates that languages can affect people's impressions and memory of other individuals.

Logan (1987) claims that the phonetic alphabet is more than a writing system; it is also a system for organizing information. The alphabet has contributed to the development of codified law, monotheism, abstract science, deductive logic, and individualism, each a unique contribution of Western thought. East Asian languages are highly "contextual." Words or phonemes typically have multiple meanings, so they require the context of sentences to be understood. Western languages force a preoccupation with focal objects as opposed to context. For Westerners, it is the self who does the acting; for Easterners, action is something that is undertaken in concert with others or that is the consequence of the self operating in a field of forces (Nisbett, 2003).

## 2.2   Physical Ergonomics and Anthropometry

### 2.2.1   Body Dimension

The study of body sizes and other associated characteristics is generally referred to as anthropometry (Lehto and Buck, 2008). The anthropometry typically refers to the measurements of body size, shape, strength, mobility and flexibility, and working capacity (Pheasant and Haslegrave, 2006). Anthropometric measurements are essential when designing devices, equipments, and systems for users. Humans vary in body dimensions, shapes, and other characteristics; thus ergonomics design requires an understanding of the variability of human beings.

Anthropometry data can be classified by the sample of the subjects: military soldiers or civilians. Anthropometric information about military soldiers has a long history and is rather complete (Kroemer et al., 1990). The military has always had a particular interest in the body dimensions of soldiers in order to provide fitting uniforms, armor, and other equipments. The recently published *Human Integration Design Handbook* (HIDH), NASA/SP-2010-3407 [National Aeronautics and Space Administration (NASA), 2010], provides a good overview of the anthropometry as well as provides guidance for human factors design, especially for human space flight programs and projects. However, military data should be used with caution when applied to a civilian population because of the selection biases of the young and healthy sample. In the recent 10 years, many institutes, libraries, and commercial companies across the world conducted large surveys on collecting anthropometric data of the civilians in different nationalities. The anthropometric data of civilians from different nationalities can be used for designing products for people from different nationalities.

### Traditional One-Dimensional Anthropometry and Three-Dimensional Body Scanning

For years, the measurement of body dimensions used traditional tools to generate one-dimensional measurement data. The tools includes calipers, measuring tapes, anthropometers, weight scales, sliding compasses, head spanners, and other similar instruments. In the recent decade, the emerging and fast development of three-dimensional scanning technology greatly changed the way of anthropometric studies. The three-dimensional body-scanning technology has many advantages over the traditional measurement system. It is capable of capturing hundreds of thousands of points in a few seconds. Moreover, it provides details about the surface shape and three-dimensional locations of measurements relative to each other. The digitalized measurements can be easily transferred to computer-aided design and manufacturing tools. In all, the noncontact, instant, and accurate three-dimensional measurement has made anthropometry studies more convenient.

In recent years, several national-level three-dimensional anthropometric surveys have been conducted. They provided online databases that either can be freely used or have to be purchased for uses. The Civilian American and European Surface Anthropometry Resource (CAESAR) collected 2400 U.S. and Canadian and 2000 European civilians aged 18–65. They provide a three-dimensional as well as one-dimensional database on 40 anthropometric measurements. The U.K. National Sizing Survey (Size UK) collected data from 11,000 subjects aged 16–90+ from U.K. populations. It used three-dimensional whole-body scanners to automatically extract 130 body measurements for standing and seated poses. Size USA conducted a comprehensive sizing survey of the U.S. population by three-dimensional body-scanning technology. It collected data of nearly 11,000 subjects from 12 locations across the United States. Size China collected

the head and face sizes of Chinese population aged 18–70 in six different locations in mainland China. It is the first three-dimensional database of Chinese head and face sizes which can be used for international manufacturers and designers. In Japan, the Research Institute of Human Engineering for Quality Life conducted many three-dimensional anthropometric measurements on the Japanese population. Different from the database listed above, the World Engineering Anthropometry Resource (WEAR) is an international organization which organizes a group of interested experts involved in the application of anthropometry data for design purposes. They collected different anthropometric data across the world as well as the methods in a wide variety of innovative applications. Its aim is to develop data models and software tools of an online worldwide information system for utilizing the latest anthropometric databases in engineering environments.

Besides the databases listed above which can provide up-to-date three-dimensional measurements, there are many anthropometric databases which provide one-dimensional measurements for researchers and designers. For example, PeopleSize 2008 provides 289 body measurement dimensions of American, Australian, Belgian, British, Chinese, French, German, Japanese, and Swedish populations. DINED provides an extensive database for the Dutch population aged 2–80+. The DINBelg 2005 provides body measurements for the Belgian population. The AnthroKids provides children's anthropometric data of the North Americans.

### Application of Anthropometry to Cross-Cultural Designs

Ergonomics and anthropometry are very important in the creation of usable products. Anthropometry data drive the guidelines for the design of a product (NASA, 2010). First, knowing the target user population determines which database defines the anthropometric data to be used. Second, once the target population is defined, designers must decide on the range of the personnel in that population who will be operating and maintaining the product. People from different nationalities vary in their body dimensions; thus designers should carefully consider the culture difference on different body dimensions for people in different nationalities. Lin et al. (2004) compared the anthropometric characteristics among four populations in East Asia. They found significant morphological differences among these peoples on the following four aspects. First, the Mainland Chinese body shape has a narrower body with midrange limbs. Second, the Japanese body shape is wider with shorter limbs. Third, the Korean body shape is midrange among the four peoples, but the upper limbs are longer. Fourth, the Taiwanese body shape has wide shoulders and narrow hips with large hands and long legs.

The revolutionary developments of three-dimensional body scanning bring tremendous benefits to product design. Niu et al. (2009) extensively summarized the benefits of three-dimensional anthropometry in product design as well as in crash test, e-commerce, forensic sciences, and videos and animations.

### 2.2.2 Movement/Reach Zone

Classical anthropometric data provide information on static or structural dimensions of the human body in standard postures. However, these data cannot describe functional performance capabilities, such as reach capabilities and movements. When performing a task, humans do not maintain standard and static postures. Furthermore, human movement varies from whole-body movement (e.g., locomotion or translation) to partial-body movement (e.g., controlling a joystick with the right arm) to a specific joint or segment movement (e.g., pushing a button with a finger while holding the arm steady) (NASA, 2010). Thus, the static postures cannot provide the advantages of dynamic posture that are involved in the design.

***Movement of Human Body*** An ergonomic designer must be familiar with how the human body moves, especially when designing workspaces (Lehto and Buck, 2008). In ergonomic design, the movements often of interests are the movements around a joint, for instance, shoulder movement, wrist movement, hip movement, and ankle movement. Table 3 provides a summary of the resources of human body movement data across the world.

The human body movement data can help designers to determine the proper placement and allowable movement of controls, tools, and equipment (NASA, 2010). Body movement data can be combined with the static body dimensions to calculate the movement ranges and reach zones in the workplace. Table 4 provides a guideline of body dimension and movement.

***Reach Zone*** The reach constraint includes the ability to grasp and operate controls. The area within which manual tasks can be performed easily is defined by the workspace (or reach) envelope (Pheasant and Haslegrave, 2006). In ergonomic design, two aspects of reach should be carefully designed: zones of convenient reach (ZCR) and the normal working area. The zone of convenient reach is the appropriate zone or space in which an object may be reached conveniently by an individual. The normal working area is described as a comfortable sweeping movement of the upper limb, about the shoulder, with the elbow flexed to $90°$ or a little less (Pheasant and Haslegrave, 2006). The data of ZCR for a full grip and the coordinates of the normal working area can be found in the book of Pheasant and Haslegrave (2006). Besides, NASA (2010) provides the data for grasp reach limits with right hands for Americans.

### 2.2.3 Biomechanics

Biomechanics explain characteristics of the human body as a biological system in mechanical terms (Kroemer et al., 1990).

NASA (2010) collected biomechanics data of body surface area, body segment volume, body segment mass properties, body segment center-of-mass location, and body segment movement of inertia of female and male crewmembers. In addition, it also provides strength data for the unsuited, unpressurized suited, and pressurized suited condition. The Chinese standards institute published a series of ergonomic standards (National Technical Committee of Ergonomics Standard, 2009). The GB/T 17245-2004 is the standard of body segment movement of inertia.

## 2.3 Methodology

### 2.3.1 Cross-Cultural User Research

Culture is a complex and multidimensional concept. People from different cultures are different in their perception, cognition and thinking styles, language, color coding and affect, and so on. Thus, a better understanding of different cultural traits in the design process is imperative in cross-cultural design. This is particularly true in the Asian Pacific area, especially in China, since in the future the Chinese will comprise one of the largest user populations.

Chinese users include people in Mainland China, Taiwan, Hong Kong, Singapore, and other areas with Chinese heritage. Chinese people speak Mandarin and other dialects, and even within the Chinese population,

**Table 3  Human Body Movement Database**

| No. | Author | Nationality | Sample | Data |
|---|---|---|---|---|
| 1 | Barter (1957) | U.S. | Military | 19 joint movements |
| 2 | Lehto and Buck (2008) | U.S. | Civilian | 21 joint movements |
| 3 | NASA (2010) | U.S. | Military | 25 joint movements |
| 4 | Hu et al., (2006) | Chinese | Civilian, elderly | 18 joint movements |
| 5 | DINED | Dutch | Civilian | 11 joint movements |

**Table 4  Guidelines for Physical Ergonomics and Anthropometry**

| Category | Guidelines | Supporting Research/ Best Practice |
|---|---|---|
| Body dimension and movement | When designing devices, equipments, and systems, target user's body measurements and movement measurements should be checked and designed to accord with the critical measurements in the anthropometric database for different nations. | Refer to the Appendix. |

there are diverse Chinese subcultural groups. For example, Chinese users in Mainland China use a simplified Chinese writing system and Chinese users in Taiwan use a traditional Chinese writing system. In general, Chinese users in Taiwan have had more opportunities to learn American culture than Chinese users in Mainland China within the past 50 years. Thus, it is expected that the differences between users in Taiwan and users in the United States would be smaller than the differences between users in Mainland China and users in the United States (Rau et al., 2004).

Bonds (1986) presented an extensive overview of the psychology of the Chinese people. In his book, he discussed Chinese patterns of socialization, perceptual processes, cognition style, personality traits, psychopathology, social psychology, and organizational behavior. The book provides insights of the culture differences between the Chinese and people from other cultures.

In addition, Yang (2001a) collected her previous published papers on how to study the Chinese, the indigenous approach. She systematically summarized the ways to conduct studies in China and how to localize studies in China. The indigenous approach Yang adopted is based on local materials and observations, a set of commonly shared meaning systems with which the people under investigation make sense of their lives and their experiences and give out and derive meanings while interacting with each other (Yang, 1991, 2000a, 2000b, 2001b). This also helps indigenous researchers understand and interpret the behaviors manifested by the people under study. The indigenization movement flourished from a general dissatisfaction among psychologists and other social scientists over employment of the Western cross-cultural approach for understanding non-Western peoples (Li et al., 1985).

When conducting cross-cultural studies in China, special issues concerning cross-cultural comparison should be carefully considered to ensure the reliability and validity of the study.

First, cautions should be taken in the explanation of culture differences by different countries. Researchers normally use country as a proxy for culture, for example, they select participants from China and the United States to represent Eastern and Western cultures. According to the results from Schaffer and Riordan (2003), 79% of the cross-cultural organization studies use country as a proxy for culture. That would be very dangerous, since, for example, even within the Asia Pacific Chinese population multilingual and cultural issues exist. Chu et al. (2005) compared the decision process of two closely related nations in East Asia. These results demonstrate the danger of generalizing decision theories across national boundaries, even when the nations are seemingly closely related. The results also indicate that the differences in decision processes among nations cannot be easily characterized as East versus West. Therefore, a better approach for researchers is to start from the theoretical framework to select and collect data, rather than simply characterize the cultural differences by countries.

Second, cautions should be taken in the equivocation of concepts in different cultures. Many studies have been extensively conducted in the United States, and many concepts in academia come from U.S. standards. However, some of the concepts may not carry the same meanings in the Chinese culture. For example, the concepts of personality traits, self-construal, achievement, and so on, may be different. Farh et al. (2004) have compared the forms of organizational citizenship behavior (OCB) that have appeared in the Chinese and Western literature. They found 10 dimensions of OCB in China, with at least one dimension not evident at all in the Western literature. Thus, careful analysis should be performed when examining the equivalence of concepts in different cultures.

Third, caution should be taken on the equivocation of measurements in different cultures. This is especially important in the use of questionnaires. Back-translation is essential for researchers conducting cross-cultural studies with subjects whose native languages are different from the researchers' mother tongue. Brislin's (1970) classic paper investigated the factors that affect translation quality and how equivalence between the source and target versions can be evaluated.

### 2.3.2 International Usability Evaluation

***The Moderator and the Test Subject*** Usability testing involves human social interaction between a test moderator and a test subject. Social and cultural norms affect this interaction just as they affect other interpersonal interactions. There is a growing literature on how Easterners and Westerners react in usability tests. A number of best practices now can be described that help to mitigate cultural bias in usability tests resulting from social interaction effects.

First, it is a good practice to use moderators, evaluators, or interviewers from the same culture as the test subjects. Vatrupu and Pérez-Quiñones (2006) studied how test participants from different cultures behaved in a structured interview setting in which the participant's task was to comment on a website. Indian participants found more usability problems and made more suggestions with an Indian interviewer than with an Anglo-American interviewer, but the comments they made to the Anglo-American interviewer tended to be more positive than negative. With an Anglo-American conducting the interview, Indians also were reluctant to discuss culture-related problems with the website and kept their comments quite general. The participants were more detailed and candid with the Indian interviewer. Yammiyavar et al. (2008) found that when subjects were paired with evaluators from the same culture, they used more head and hand gestures to communicate than if the evaluator was from a different culture, providing a richer source of nonverbal data to analyze. Sun and Shi (2007) studied how using one's primary versus secondary language (English versus Chinese in this case) in a think-aloud test affected the process of the test. Chinese evaluators speaking Chinese to Chinese test users gave more help to users and a more complete introduction to the product being tested and encouraged

users more frequently than Chinese evaluators speaking English during the test.

A second good practice is to avoid pairing evaluators and test users who differ in their perceived status or authority. Particularly in cultures with high power distance, such as China and Malaysia, the behavior of both the test user and the test evaluator is affected by perceived differences in status or authority. Participants in high-power-distance cultures do not challenge or question the evaluator because of the perception of the evaluator as a person of authority (Burmeister, 2000). Yeo (1998) illustrated this with an example of a usability test conducted in Singapore in which a participant broke down and cried from frustration. A posttest interview revealed that the participant's behavior was in part due to the Eastern culture, in which it is not acceptable to criticize the designer openly, because it may cause the designer to lose face. Evers (2002) evaluated cultural differences in the understanding of a virtual campus website across four culturally different user groups (England, North America, the Netherlands and Japan) by using the same methods for each group. The results indicated that Japanese participants who were secondary school students felt uncomfortable speaking out loud about their thoughts and seemed to feel insecure because they could not confer with others to reach a common opinion.

But the effect of culture can go both ways. In a pilot study of think-aloud tests, Sun and Shi (2007) found that evaluator's behavior is also affected by differences in level of perceived authority. When the evaluator's academic title or rank was higher than that of the users, the evaluator tended to more frequently ask the user what he or she was thinking during the test. The evaluator also tended not to provide the user with more detailed instructions during the test.

The third guideline is to train evaluators to combat the "conversational indirectness" of Asian users. Subjects from Asian cultures will tend to seek a compromise and be indirect when evaluating user interfaces. For example, Herman (1996) studied cultural effects on the reliability of objective and subjective usability evaluation. The results of objective and subjective evaluation correlated poorly in Herman's study. The Asian participants were less vocal, very polite, and not inclined to express negative comments in front of observers, so that the results of subjective evaluation tended toward the positive despite clear indication of poor user performance. Herman's solution was to invite test participants to work in pairs to evaluate the interface and make the usability test more of a peer discussion session.

Shi (2008) conducted observations of usability tests in China, India, and Denmark and, like Herman, also noted that Chinese users often kept silent and did not speak out actively, particularly in formative evaluations. Two studies (Shi, 2008; Clemmensen et al., 2009) explained this observation in terms of Nisbett's (2003) cultural theory of Eastern and Western cognition. Accordingly, Chinese people tend to have a holistic process for thinking as opposed to the more analytic style of Westerners. Holistic thought is not as readily verbalized as analytic thought. So, they theorized that

Chinese users in a think-aloud test situation are thinking about the user interface in holistic terms and simply have a more difficult time putting those holistic thoughts into spoken words.

Shi recommended that evaluators receive special training to conduct testing with Chinese users that is based on think-aloud methods. Evaluators should be trained to use reminders and questions, "digging deeper probes," to get the users to talk. For example, evaluators in Shi's study reported that if they knew users were looking for some object or feature on the screen, they would ask, "what are you looking for?" Then, immediately, the user would tell them about what they were looking for. This method of asking related questions to encourage speaking aloud was found to be more natural than just asking people to "keep talking." Evaluators also should be trained to be patient with Eastern participants as they pause and think in between verbalizations (Shi, 2009). All of the above suggestions require a significant amount of training and skill on the part of the evaluator. If such resources are not available to support the test, then perhaps the best approach is to consider an alternative to the think-aloud method for conducting the evaluation.

Shi (2009) found no significant differences in the set of usability problems reported by Chinese and Danish evaluators. However, their ratings of the severity of the usability problems did differ significantly. Chinese evaluators rated problems less severely than Danish evaluators and often rated problems in the middle of the five-point severity scale. Shi (2009) suggests that if problem severity rating is part of the test, then a four-point, rather than five-point, scale should be used to prevent middle-of-the-scale ratings.

***Participant and Evaluator Recruiting***   Recruiting participants with similar background in different places at the same time has made international usability evaluation very difficult. The experimenters have options to carry out international usability evaluation such as going to the foreign country, running the test remotely, hiring a local usability consultant, or asking help from staff in a local branch office (Nielsen, 1990, 2003). Many researchers (Choong and Salvendy, 1998, 1999; Dong and Salvendy, 1999; Fang and Rau 2003; Fukuoka et al., 1999; Prabhu and Harel, 1999; Evers, 2002) chose to recruit participants in two or more countries by going to the countries. The Web has made conducting international usability evaluation a new option. Lee and Harada (1999) conducted an evaluation by recruiting participants on the Web. However, for countries with no experimenters present, they found it difficult to recruit participants on the Web.

Clemmensen et al. (2007) address the problem of "hidden user groups," groups of people who represent significantly different target user segments within the same culture. They suggest that test planners attempt to balance out potential hidden user groups within user segments. For example, users who are accustomed to foreigners and adapt quickly to international test conditions should be balanced by users who are not accustomed to foreigners. Traditional and culturally

sensitive users such as those one might find in rural areas need to be balanced in the pool of test participants by more modern, urbanized people who are less influenced by a country's local culture. To avoid missing critical usability problems during the test, they recommend that evaluators be chosen carefully from evaluator groups suitable to the members of the identified hidden user groups.

### Language

**Verbal Language-Using Interpreters and Translators** Language is a significant factor in international usability evaluation. Without translated content, the target audience may be limited to Web users with a certain education and social background. Nielsen (2003) indicated that the first two concerns for international usability testing are displaying in the user's native language, character set, and notations and translating the user interface documentation into the user's native language. Rau and Liang (2003b) conducted a card-sorting test with Chinese users in Taiwan for an international website. Even though all the information items were translated into Chinese, a dictionary and instructions also were made available to participants. Also, anything offensive in the user's cultural background should be avoided. The testing materials and procedure should accommodate the way business is conducted and the way people communicate in various countries (Nielsen 2003). Interpreters are necessary if the evaluators are not able to speak the user's native language well. The evaluators need to avoid any important information lost in translation so that videotaping and audio recording are useful for further analysis. Nielsen (2003) suggests meeting interpreters beforehand and reminding them that they should not help users during the test.

**Nonverbal Language** Non verbal language displayed in the form of hand and head gestures is commonly used by test participants as an occasional substitute for verbal language or to elaborate on or supplement it. Hand and head gestures also often communicate the comfort level of the participant and his or her readiness to communicate with the evaluator. Observing and analyzing gestures during a test can provide a rich source of data that adds context to the verbal data being recorded.

Yammiyavar et al. (2008) questioned whether users from different cultures exhibited similar or different patterns of nonverbal communication, including the type, frequency, and usage of gestures. They analyzed the occurrence of gestures in video recordings of think-aloud tests that used subjects from Denmark, India, and China. Gestures were grouped into four types (Ekman and Friesen 1969): emblems, illustrators, adapters, and regulators. Emblems replace words with gesture-based signs like nods of head for "yes" or a V sign for victory. They tend to be culture-specific. Illustrators include such actions as banging on the table or sketching shapes in the air. They help the subject to verbalize their thoughts. Yammiyavar et al. (2008) discovered that these can be quite important markers for usability problems because illustrator gestures frequently precede the verbalization of a usability problem to the evaluator. Adapter gestures are actions of the body that convey feeling of pressure or discomfort, for example, cracking one's knuckles, tapping one's feet, stroking hair or chin while in deep contemplation, and "squirming" in one's seat. They indicate the subject's comfort level with the test situation. Finally, subjects control the flow of conversation with the evaluator by using regulator behaviors such as nodding the head up and down to indicate agreement.

Yammiyavar et al. (2008) found that the frequency of using gestures during the test was not significantly different across the three cultural groups. This contradicts the popular belief that Indians, for instance, use more gestures to communicate than other cultural groups. Also, regulator gestures appeared to be used similarly across the cultures studied, but with some tendency for Chinese to use them the least. In contrast, there were significant differences across the cultural groups in specific emblem gestures used to replace words and in adapter behaviors. Certain illustrators appeared to be culture specific as well. The researchers concluded that there is a need to benchmark gestures used in these different cultures and their meanings and then provide those to usability test evaluators to guide observations and understand what they observe.

The facial expression associated with surprise often is used as a marker by evaluators to indicate that a usability problem has been detected by the user. Clemmensen et al. (2009) have questioned the validity of this practice in cross-cultural usability tests. From Nisbett's (2003) theory of cultural cognition, they hypothesize that Easterners will experience less surprise than Westerners when presented with inconsistencies in user interfaces. With their logical, analytic orientation, Westerners tend to focus on fewer causes of observed events, while Easterners, with their holistic orientation, tend to consider more causal factors as well as the context of the event. As Clemmensen et al. (2009) point out, this makes it easier for them to identify a rationale for why the event occurred in the way it did, resulting in less surprise.

***Instructions, Tasks, and Scenarios*** Instructions to the test participant can vary significantly in how much contextual information they provide. At the one extreme, instructions are strictly focused on the task to be performed with the application being tested. No explanation is given about the purpose of the application, when you might use it, or why. At the other extreme, the explanation of the task is embedded in a rich context provided by a real-life scenario. Clemmensen et al. (2009) suggest that Westerners, with their tendency to focus on the central elements presented, such as the details of the task, will be able to obtain that in either presentation of instructions. Easterners, however, will find that the stark instructions are insufficient. With their holistic style of thinking, they will prefer to have the task explanation embedded in the context of a real-life scenario. The recommendation from Clemmensen et al. (2009) is that if cross-cultural testing is to be conducted; then the test planner should

consider adapting the instructions to the cultures of the participants. Planners might want to have different versions of the test protocol prepared that include different types and amounts of background information.

A classic example of adapting test scenarios to the target culture is from Chavan (2005), who engaged Indian test participants by embedding the test tasks in "Bollywood" scenarios. The method capitalizes on the popularity in India of watching Bollywood movies and the fun of openly critiquing them. The test participant is asked to imagine a dramatic scenario similar to a movie script, perform the desired tasks using the application in that context, and critique it.

***Questioning Universal Constructs of Usability*** Underlying much of what is done in user-centered design is the notion that people worldwide understand the fundamental attributes of "good usability" in the same way—effectiveness, ease of use, visual appearance, efficiency, satisfaction, fun, and nonfrustration [International Organization for Standardization (ISO) 9241, 1998; Frandsen-Thorlacius et al., 2009]. As the preceding sections of this chapter have shown, we may culturally adapt specific local instantiations of a user interface to the preferences of local cultures. But it is assumed that the local instantiations are all done with the goal of enhancing the universal construct of usability as defined by these basic attributes.

Recent research questions this assumption of universal constructs of usability (Frandsen-Thorlacious et al., 2009; Hertzum et al., 2007). Frandsen-Thorlacious et al. sampled 412 users from China and Denmark to determine how they understood and prioritized attributes of usability. Chinese users placed greater value on visual appearance, satisfaction, and fun than Danish users. Danish users valued effectiveness, efficiency, and lack of frustration more highly than Chinese users. Clearly, dimensions of usability were weighted differently for these two cultural groups in the study. Hertzum et al. (2007) used repertory grid interviews, a method for identifying how users give meaning to their experience, to explore how personal constructs of usability differed between people from three different cultures, Denmark, China, and India. They found that some of the constructs verbalized by study participants were consistent with common notions of usability such as ease of use and were important at least to some degree to participants from all three cultures. But other constructs differed from commonly used attributes of usability, for example, the attribute of "security." The most important usability attributes for Chinese subjects involved issues of security, task types, training, and system issues. In contrast, Danish and Indian participants focused on more traditional aspects of usability such as "'easy to use," "intuitive," and "'liked." None of the Chinese subjects verbalized these as primary constructs of usability.

These two studies are just a start in understanding what aspects of usability people value in different cultures. However, they raise questions about how user interfaces are currently designed and how they are tested across cultures. If the testing assumes that everyone values the same attributes of usability, then usability test participants from all cultures would be expected to find the same number of usability problems related to the same usability attributes. The two studies reviewed here raise the possibility that test participants in one culture may not identify problems related to a particular attribute of usability simply because that attribute is not as important to them as it is to a participant from another culture. The absence of verbalized problems in a particular area of the user interface design could be misinterpreted as an indication that no problems exist.

From a business perspective, one would have to question why a company would invest scarce development dollars to perfect aspects of a user interface that are not particularly important to a targeted user group. These findings should make companies aware that basing usability requirements for a global product on just one or two cultural groups runs the risk of minimizing attributes that turn out to be important to a second, third, fourth, $n$th, cultural group of users somewhere in their market space, perhaps even to the majority of potential product users. Perhaps a means to identify and weigh what attributes of usability are important to test subjects must become a standard part of global usability testing methodology. And the global locations for usability testing should reflect the segments of the intended global market.

### 2.3.3 Internationalization and Localization

Internationalization is the process of designing a product or system so that it is generic enough to accept many variations and cultural contexts and can be adapted to various languages and regions without engineering changes. Localization, on the other hand, is the process of adapting a product or system so that it can be used by people of a particular cultural context, locale, and area. Many companies perform localization by adapting a product created specifically for its domestic market. A product designed for its creator's domestic market is often embedded with the cultural markings of the creator's cultural context (Hoft, 1996). Any localization after the product has been developed will require recoding and possibly reengineering to accommodate the cultural context of a target locale. Employing internationalization provides the framework and structure in which localization takes place more easily and more efficiently (Luong et al., 1995). Internationalization is the preparatory stage of product development where the embedded culture and language are extracted and generalized (Taylor, 1992). For any company who intends to extend its products in the global market, the approach of product globalization is recommended. Globalization consists of a minimum of two steps: starting with internationalization of a base, culture-free product followed by localization of the base product for each target locale.

Product globalization poses many challenges to product developers beyond text extraction and translation as commonly assumed by people as the only task for global market.

***Cultural Biases*** Designers should avoid using cultural-specific references as they are prone to misunderstanding and misinterpretation cross-culturally.

Visual and verbal puns should not be used since they cannot be translated well. For example, a very popular icon with a light bulb is used to represent "bright idea" in the United States. However, to the rest of the world, it might merely mean a light bulb so that the concept will be easily lost during the localization process.

Designers should stay alert with any potential political, religious, culture-specific values and taboos or other symbolism which may affect the meanings perceived in a particular country.

Also, providing culturally sensitive aesthetics can play a major role in the product acceptance of the target locales as people have internalized perceptions of what looks local and what looks foreign. Culturally sensitive aesthetics should be implemented during the localization process, such as the use of familiar objects, colors, typography, architecture or landmark, and influential geometry.

## 3  USER INTERACTION PARADIGMS AND TECHNOLOGIES

### 3.1  Graphical User Interface

As discussed in previous sessions, how people interact with objects around them is determined by their past experiences with these objects (or similar objects) and their expectations of how things should work when they use the objects. A graphical user interface (GUI) is the graphical and visual representation of, and interaction with, programs, data, and objects on the display of a system. An important feature of GUIs is the ability for users to directly manipulate elements and information on the screen display. A GUI system is portrayed as an extension of the real world. Thus, the users' experiences and expectations play a critical role in the usability of GUI systems. The same is true when it comes to design for users with different cultural backgrounds in that the cultural experiences and expectations of the target audience need to be considered throughout the development lifecycle.

A GUI system usually includes four basic elements: window, icon, menu, and pointing device (WIMP). The WIMP interaction was first developed at Xerox PARC in the 1970s with the desktop metaphor on a personal computer. The WIMP interaction and its variations and extensions are widely used in computing technologies today, including desktop applications, Internet browsing, Web applications, and mobile computing.

#### 3.1.1  Information Organization and Representation

As discussed earlier, differences in cognitive styles exist among people of different cultural backgrounds. Those differences have implications on how information should be represented and organized on GUIs to accommodate different thinking styles due to cultural differences.

Oftentimes on a GUI the information is organized into hierarchical structures. For example, menus display listings of choices or alternatives that users have at appropriate points while using the system or create a set of listings that guide a user from a series of general descriptors through increasingly specific categories on following listings until the lowest level listing is reached. Design guidelines and best practices have been developed for designing usable menus (e.g., ISO, 1997; Galitz, 1997; Weinschenk et al., 1997). Those guidelines call for logical and meaningful organization of menu options. The menu options should be arranged into conventional or natural groups known to users and follow logical order. The groupings should be logical, distinctive, meaningful, and mutually exclusive.

Designing menu structures for people with different cognitive styles cross-culturally needs to accommodate the differences in the tendency of information organization and represent the information accordingly. As Choong (1996) points out when representing information of a system on a GUI, Chinese users will benefit from a thematically organized information structure, whereas American users will benefit from a functionally organized structure. Nawaz et al. (2007) report similar results of cultural differences on the ways the Chinese and Danish group objects, functions, and concepts into categories. In the card-sorting tasks, the Danish subjects prefer to highlight category name by its physical attributes, whereas the Chinese subjects highlight the category by identifying the relation between different entities. The Chinese subjects also utilized more thematic categories than the Danish subjects in the study by Nawaz et al. (2007). Kim et al. (2007) report similar findings with Korean and Dutch users interacting with menu structure designed for mobile phone interfaces. The relational grouping participants (Koreans) were more likely to select and prefer the thematically grouped menu, whereas taxonomic grouping participants (Dutch) had the tendency to select and prefer the functionally grouped menu.

All user interfaces use some forms of metaphors to provide visual and conceptual representations of major user interaction objects and their associated actions. A well-chosen metaphor can be helpful when all or part of the interface includes functions or features that are new to the target users. Metaphors are used to help users connect what they do not know with what they have known. It is imperative that the metaphors are simple, easily understood, and quickly learned. Metaphors should allow target users to easily relate to their real-world experiences. When designing cross-cultural user interfaces, the use of metaphors becomes a challenge as the real world changes from culture to culture. Many companies choose to localize metaphors in their user interfaces by only redesigning the objects or translating the text in a certain metaphor. However, "translating" a metaphor is not sufficient; the entire metaphor will need to be reevaluated and replaced to make the interface mapped to the target users' cultural experience (Evers, 1998). For example, for designing multicultural applications, Salgado et al. (2009) propose five

conceptual metaphors to accommodate users' attitudes and cultural variables: located "at home" metaphor, telescope observer metaphor, close observer metaphor, foreigner "with subtitles" metaphor, and foreigner "without subtitles" metaphor. Salgado et al. expect that those five conceptual metaphors will help designers think about their own cultural perspective and also help the designers to think about how such perspectives can be experienced in different ways.

### 3.1.2  Graphics, Symbols, and Icons

Icons are an essential part in any system with a GUI, that is, WIMP interaction. Icons are small pictorial symbols used on GUIs to represent certain capabilities of the system and to be animated for bringing forth these capabilities for use by the users. The benefits of using icons include to represent visual and spatial concepts, to save screen space, for immediate recognition, for better recall, to reduce user's reading time, and to help products go global (Horton, 1994).

A number of researchers have written about designing icons for specific cultures or for international use. Shen et al. (2007) report on Chinese Web design with cultural icons. Cultural issues in designing international biometric symbols are described in Rau and Liu (2010) and Choong et al. (2010). Pappachan and Ziefle (2008) discuss cultural influences on comprehensibility of icons. Kim and Lee (2005) report on cultural differences in icon recognition on mobile phones. From these we can conclude that a good icon for cross-cultural use has the following properties:

- Mimics both the physical appearance and the function or action of the object it represents
- Clearly represents the state of the object if it is an object that can assume more than one state
- Uses only widely recognized conventions for color and shape
- Is not directional and can be used without rotation
- Is not culture bound
  - No embedded text characters
  - No culture-specific metaphors

### 3.1.3  Presentation, Navigation, and Layout

There are existing guidelines (e.g., ISO, 1997) for designers to follow on how to develop usable presentation and layout of GUI components as well as the navigation among those components. The key is that the design has to match the user's work flow, experience, and expectation. In addition to following GUI design guidelines, there are considerations that need to be addressed in cross-cultural GUI design. For example, different cultures employ different format conventions and measurement systems that will affect the presentation of such information on the GUIs. The arrangement of information on the screen affects how efficiently and comfortably people can scan, read, and find information.

The information should be laid out on the screen in a natural orientation for the target users. For example, in cultures where traditional Chinese form is used, such as Hong Kong and Taiwan, people tend to scan the screen "across the columns" following an "N" pattern. In countries where simplified Chinese is used, for example, China and Singapore, information is mostly printed in rows and read from left to right and top to bottom, that is, a "Z" pattern.

***Language Issues***  The textual information on a GUI imposes significant challenges in cross-cultural design. First, if English is the language of the base product, it is important to start with unambiguous language that minimizes the use of technological jargon, avoids abbreviations, and uses plain, simple English. Second, the same language used in different locales can have dialect differences, including spelling, word usage, grammar, and pronunciation. Third, literal translation of terms often fails to accurately translate the meaning or concept underlying a word.

Linguistic differences should be taken into account, such as text directionality, linguistic boundaries, text wrappings, justifications, and punctuations. When presenting an ordered list or information, a single sequence or ordering of textual information should not be assumed, even if the languages share the same alphabetical system. Collating ideographic characters, such as Chinese Hanzi and Japanese Kanji, is more complex than sorting Latin characters. There are four different collating methods that GUI designers should be aware of and take into consideration: radicals, number of strokes, phonetic sequence, and frequency of use (Rau et al., 2010). Cross-cultural design also will need to take into account physical language variations such as directionality, hyphenation, stressing, fonts, sizes, orientation, layouts, spaces, wrapping, and justification. In European languages, such as German, words tend to be long strings of characters. In contrast, Asian words tend to be shorter, but the characters are much more complex (such as Chinese) and will require more pixels to render clearly on a display screen. Cross-cultural GUI designers should be aware of the possible text expansion required both horizontally and vertically.

Conventional guidelines assume that people read left to right and thus use a left-to-right orientation for labeling text boxes, presenting text, scrolling text within a text box, and presenting a series of control buttons. However, designers should take note that some languages are read from right to left, for example, Arabic or traditional Chinese writing. When designing for such languages, the display features should accommodate the text direction as well as navigation flow.

There are some key linguistic differences to be considered when translating a user interface from one language to another. For user interfaces, adequate screen space needs to be allocated for possible text expansion due to translation. For example, German words are usually longer than their counterpart English words. Composite messages should be avoided, such as warnings with a word or words dynamically determined. The

composite messages will not be translated well since sentence structures could vary dramatically across languages. The rules for word wrapping can also be very different from one language to another.

Sometimes, it is essential to support multiple languages simultaneously. The product designers should avoid using national flags to toggle among languages or using words in one language for selecting among languages. Using national flags for language selection may be offensive to some users since there can be more than one country using the same language. For example, English is used in the United States, United Kingdom, Canada, and many other countries. Using words in one language as language-selecting options is also inappropriate since the users will have to know that language to make the selection. For example, in an English user interface, a Chinese user will have a problem in picking out the word *Chinese* from a list consisting of language options written in English if the Chinese user does not understand English.

Scripts are a collection of characters and glyphs that represent a written version of a spoken language. In many cases, a single script may serve to write tens or even hundreds of languages, for example, the Latin scripts. In other cases, only one language uses a particular script, for example, Hangul, which is used only for the Korean language. The writing systems for some languages may also use more than one script; for example, Japanese makes use of the Han (or Kanji), Hiragana, and Katakana scripts.

Written language can be bidirectional or unidirectional. Most languages are unidirectional. For example, English is written from left to right in a unidirectional fashion. Chinese is another example of unidirectional scripts, but the direction can be left to right, right to left, or top to bottom. A bidirectional script, such as Arabic, can be written from right to left and left to right (in certain situations, such as numbers) in the same context.

Some languages, such as Chinese, demand special consideration of alternative methods of text input. Niu et al. (2010) of Nokia developed a new method called Stroke++ for Chinese character input on mobile phones with the goal of making keypad typing more accessible to novice mobile phone users. Chinese mobile phone input methods use either standardized phonetic notations, such as Pinyin and Zhuyin, or structural information about characters, such as Wubihua or Cangjie. Pinyin requires knowledge of Latinized Chinese Pinyin, which presents difficult barriers to elderly people and nonusers of Pinyin. Wubihua, or Stroke, another popular input method, accepts five distinguishing strokes only when the input sequence is the same as the standard writing order of the character. These methods are difficult to learn because the standard varies between mobile phone manufacturers and there is significant variation in writing habits between different people. The new Stroke++ method exploits the fact that the 600 most frequently used characters in Chinese (out of 20,000 total) can cover 92.9%.of the user's needs in short messages, thus greatly reducing the required set of radicals. However, rather than defining rules to restrict the sequence of entry, this method allows users to input radicals in arbitrary order to form a character, the desired character being selected from a pull-down list of possibilities organized according to frequency of use. In addition, the keypad layout is designed according to Chinese characters' square shape. The radicals are grouped to make the keypad meaningful. For example, the radicals 金 (Metal), 木 (Wood), 水 (Water), 火 (Fire), and 土 (Earth) representing the five elements in traditional Chinese culture are arranged in a single line to help users to remember. 女 and 人 are put together in the middle to make the Chinese word for "woman." Table 5 provides guidelines for use of language.

### Format Conventions and Measurement Systems

Formats are an artifact and are specific to different locales. Numeric values can be represented in different ways. Separators are not always used; even if they are used, different locales use different symbols for separators and with different formats. For the same number, 1,234.56 is used in the United States and

**Table 5 Guidelines for Use of Language**

| Category | Guidelines | Supporting Research/ Best Practice |
|---|---|---|
| Language | Account for physical language variations such as directionality, hyphenation, stressing, fonts, sizes, orientation, layouts, spaces, wrapping, and justification. | Rau et al., 2010 |
| Language | Consider differences in sort order between different languages. | Rau et al., 2010 |
| Language | Allow adequate screen space for possible text expansion due to translation. Languages such as German can be especially problematic. | Rau et al., 2010 |
| Language | Avoid composite messages such as warnings with a word or words dynamically determined. | Rau et al., 2010 |
| Language | Avoid using words from one language as language-selecting options, for example, the word ''China'' to select Chinese language option. | Rau et al., 2010 |
| Language | Consider, for example, the directionality of languages in the design of text boxes. | Rau et al., 2010 |
| Language | Carefully consider the need for alternative text input methods for languages such as Chinese. | Rau et al., 2010 |

**Table 6  Guidelines for Format Conventions and Measurements**

| Category | Guidelines | Supporting Research/ Best Practice |
|---|---|---|
| Format conventions and measurement systems | Be aware of the way in which numbers are represented in the target culture, including treatment of separators and glyph shapes. | Rau et al., 2010 |
| Format conventions and measurement systems | Consider the format conventions for currency, calendars, date and time, addresses, and telephone numbers for the target culture. | Rau et al., 2010 |
| Format conventions and measurement systems | Consider the measurement conventions of the target culture, e.g. weights, dimensions, temperatures, and paper sizes. | Rau et al., 2010 |

Canada; "1.234,56 is used in Germany, Holland, and Italy; and 1.234,56 is used in France and Sweden. The numeric glyph shapes can be different as well.

Other format conventions that need to be taken into account include currency, calendars, date and time formats, names and addresses, and telephone numbers. Research and practice in cross-cultural user interface design have focused on the internationalization and localization of display codes, such features as formats, colors, icons, and graphics (Liang and Plocher, 2003).

Designers need to be aware of different measurement conventions for different locales, for example, dimensions, weights, temperatures, and paper sizes. Adequate accommodations need to be provided so that the product uses the appropriate measurements for the target locales. Table 6 provides guidelines for format conventions and measurements.

### 3.1.4  Color Coding and Affect

***Color Associations and Common Safety Words***
There is no difference between cultures in the actual perception of colors since there are common physiological bases for color vision (Bonds, 1986). Further, in technical applications, there is considerable agreement between Asians and Americans about how common safety words are associated with colors (Courtney, 1986; Luximon et al., 1998; Liang et al., 2000; Kaiser, 2002):

- Danger—red
- Go—green
- Hot—red
- Stop—red
- Safe—green
- Caution—yellow

However, a few color associations have been found to be similar in some cultures (Courtney, 1986; Liang et al. 2000; Kaiser, 2002) but different in others:

- Cold—white: Chinese, Japanese; blue: American
- On—green: Chinese; red: American

The point is that most colors have at least some ambiguity associated with their meaning and should be avoided for signaling important or safety-critical concepts unless they are combined with other coding such as text or icon or both. Designers should be aware of ambiguities and use color coding with caution.

***Color and Affect***  What people feel about colors is more subject to cultural variation. Colors and the combination of colors have different meanings in different cultures. Many researchers have conducted studies on color and its impact for different product design.

Early studies on color codability demonstrated that people in different societies did not have the same array of colors to partition the color spectrum (Whorf, 1956). Berlin and Kay (1969) argued that, if the mechanism underlying color perception is universal, there should be agreement on colors among those who speak different languages from different cultural environments in spite of variations in color vocabulary. They studied 20 languages and discovered meaningful regularities in the use of basic color terms which are names of color categories consisting of only one morpheme. They also noted an evolutionary progression in color terms in the sense that culturally simpler societies tended to have fewer basic color terms than culturally complex societies, for example, large-scale, industrial countries. MacLaury's (1991) work also demonstrates the effect of cultural factors on color coding. A comprehensive study of color naming has been presented by Russell et al. (1997). Davies and Corbett (1997) studied speakers of English, Russian, and Setswana languages, which differ in their number of basic color terms and in how the blue-green region is categorized.

Prabhu and Harel (1999) studied users' needs and preferences for digital imaging products in Japan and China. They found that Japanese men preferred single color fonts and simple fonts without emphasis on all the three lines of help, whereas Japanese women and Chinese men and women preferred multiple colors and highlighted or emphasized fonts. Also, Japanese preferred pastel colors for both the welcome screens and the interaction screens. Though Chinese men preferred Chinese colors, preference for women was mixed between Chinese and Japanese pastel colors.

Minocha et al. (2002) conducted informal observations and analysis for the choice of colors on some e-finance sites in India and Taiwan. Three e-finance

**Table 7  Guidelines for Using Color**

| Category | Guidelines | Supporting Research/ Best Practice |
|---|---|---|
| Color | Avoid using color to signal important or safety-critical concepts unless they are combined with other coding such as text or icon or both. Use the most commonly accepted color–safety word associations. | Courtney, 1986; Luximon et al., 1998; Liang et al., 2000; Kaiser, 2002 |
| Color | If you must use color in your design, be aware of the affective associations common in the target culture(s). | Spartan 1999; Osgood et al., 1975 |

websites in India were studied for Indian cultures, values, and customs: (1) ICICI Bank, the second largest commercial bank in India that has pioneered Internet banking, (2) Allahabad Bank, an interesting and prominent example of a dual-language (Hindi and English) site, and (3) the State Bank of India, India's largest commercial bank. The authors analyzed the website of ICICI Bank as the representative Indian choice of colors. For Indian users, use of red is associated with vitality, energy, prosperity, and health. Red is considered stimulating and shows ambition and initiative. In religious ceremonies and marriages, the guests dress in red-colored clothes. Besides, use of saffron is considered auspicious among the Hindus, Sikhs, Jains, and Buddhists. The combination of red and saffron can be considered to signify prosperity and growth for current and prospective customers.

While color vision is universal, what people feel about color is definitely more subject to culture differences. In a cross-cultural study of color affective meaning, Osgood and his colleagues (1975) compared differential measurements of about 600 concepts in over 20 communities all over the world. Color names were included in the experiment into subjective culture. Results were expressed in terms of three dimensions. *Evaluation* indicates whether subjects felt that concepts were something like good or bad, nice or awful, and so on; *potency* indicates big or small, powerful or powerless, and so on; and *activity* indicates fast or slow, young or old, and so on. Across the language communities tested, color terms produced some universals for affect. These included that the feature brightness universally showed high evaluation, the colors black and grey showed low evaluation and activity, and red showed high potency as well as high activity.

Color has very different meanings around the world and therefore it is a significant element for all forms of international interaction. Holzschlag (1999) and Morton (2003) point out that the color purple represents death in Catholic parts of Europe and that Euro Disney should have done its homework well before choosing purple as a signature color. There are numerous excellent resources that deal with color usage. Jill Morton's colormatters.com website has extensive information on color studies and online books on color. Spartan (1999) gives many examples related to different cultures in their association with colors. For example, in the United States, Americans tend to relate blue to authority whereas in Canada red relates to authority

as well as inspirer of nationalism. In the British Isles, black is commonly used for authority figures. As in China, red and gold are used for honor and authority.

Although the strict meanings of colors may be fading across culture, it is still a good practice to use colors carefully. If the same color is to be used across locales, it would be a good idea not to use primary colors in design because they may still carry negative or positive meanings in some cultures. Table 7 provides guidelines for using color.

## 3.2  Web/Hypermedia

Although designing for the Web and hypermedia is relatively new compared to conventional GUI design, the key components for a good user interface design can be applied to Web and hypermedia design as well.

Information structure and searching are the two important aspects for Web design. The information structure of the Web is concerned with how to organize the information and how to navigate Web users. Searching is concerned with what searching mechanisms should be provided to users and how to present searching results.

### 3.2.1  Information Structure (Navigation and Hyperlinks)

As noted in previous section, people around the world hold different thinking styles. The differences in thinking could affect their performance interacting with computers. Researchers have highlighted that different cultures often focus on different attributes of the same items or objects (Choong, 1996; Choong and Salvendy, 1999). How items or functions should be grouped in pages and how links on a website or buttons and menus on the user interface should be labeled are highly affected by culture. The information of the website should be structured in association with the target user's cultural traits.

Marcus and Gould (2000) contend that navigation will be impacted by cultures. Users from cultures that feel anxiety about uncertain or unknown matters would prefer navigation schemes intended to prevent being lost.

Luna et al. (2002) suggested that websites should be structured to conform to the target cultures, for example, the site could have a hierarchical or a search-based structure, depending on whether the target visitors

belong to a high-context culture (e.g., Japan) in which hierarchical structures might be preferred or a low-context culture (e.g., Germany) in which search-based structures might be preferred. Furthermore, Luna et al. (2002) pointed out that it is important to provide users with a culturally congruent site by offering links to pages that address the respective values, symbols, heroes, and rituals of a particular culture. The culturally appropriate navigation patterns will lead to less confusing and more satisfactory user experience.

Rau and Liang (2003a, 2003b) pointed to well-designed navigational supports to combat the tendency toward disorientation of users in high-context cultures. Rau and Liang (2003a) used a survey designed by Plocher et al. (2001) to classify Web users as either high or low context on Hall's communication style dimension. They postulated that communication style would affect how people interact with information systems, particularly nonlinear, hypertext systems such as the Web. In their experiments, they found that high-context people browsed information faster and required fewer links to find information than did low-context users. However, high-context users also had a greater tendency to become disoriented and lost their sense of location and direction in hypertext. Low-context users were slower to browse information and linked more pages but were less inclined to get lost. In another study, Rau and Liang (2003b) investigated the effects of communication style on user performance in browsing a Web-based service. The results showed that participants with high-context communication style were more disoriented during browsing than were those with low-context communication style.

Cyr and Trevor-Smith (2004) conducted an empirical comparison of German, Japanese, and U.S. website characteristics. They found preference for different navigation and search capabilities across different cultures. The Japanese sites were twice as likely to use symbolic navigation tools as were the German or the American sites. Preferences for vertical and horizontal menus are statistically significant. German and Japanese sites used a "return to home" button twice as much as the U.S. sites. As to the type of hyperlinks used, the results found that the number of external links and the functionality of links differ across cultures. External links are used in almost all Japanese sites, compared to only two-thirds of U.S. and German sites. The Japanese use symbols for links significantly more than do German and U.S. sites. They concluded that providing navigation appropriate to cultures is important in order to avoid the "disorientation" when users make navigational errors when searching for information.

Lo and Gong (2005) studied the cultural impacts on the format and layout design of e-commerce websites. They examined 50 leading e-commerce websites in the United States and 50 leading e-commerce websites in China. They found that the U.S. sites showed a clear trend of using blue color for hyperlinks. But in the Chinese sites three colors, black, blue, and red hyperlinks, are equally likely to be used. Although there is some evidence on the emergence of a global e-commerce culture or some common standards on the color of hyperlinks, there is also clear evidence of the need for e-commerce website designers to consider the impact of local culture, for example, the higher frequency of using red color in Chinese websites.

Lo and Gong (2005) also examined the navigation model of the e-commerce websites in the United States and China. In terms of direction of navigation, five navigation models are possible: left oriented, right oriented, top oriented, bottom oriented, and center oriented. In terms of the appearance of navigation buttons, it can be text based or it can be GUI based. The results showed that the Chinese sites favor the top-oriented navigation model, while the United States is roughly equal in left-, top- and center-oriented navigation models. It is not surprising that Chinese sites favor the top-oriented navigation model, because traditionally Chinese writings are read from top to bottom and right to left. As to the appearance of navigation buttons, it was found that U.S. sites favor GUI-based navigation buttons, while Chinese sites favor text-based navigation buttons. They further discussed the reasons that Chinese sites employ the top-oriented and text-based navigation model on the cultural aspects. Yahoo is one of the early entries to China's e-commerce market, and it was quite successful. The Chinese Yahoo site employs the top-oriented and text-based navigation model. This mode was also used by the other two leading e-commerce sites in China: sina.com and sohu.com. China has a high-power-distance, collectivism, and high-uncertainty-avoidance society, the success of the above three websites are clearly recognized, and thus subsequent entries into the Chinese e-commerce market tend to imitate their approach using the same navigation model. In the United States, the situation is quite different. Because the United States is rated higher in individualism, a greater variety of navigation styles and site design approaches were found at U.S. sites.

Kralisch et al. (2005) studied the impact of culture on website navigation behavior. In their study, they were concerned with the impact of cultural dimensions (Hofstede's long-term orientation and uncertainty avoidance and Hall's mono-/polychronicity) on user behavior. They collected behavioral data by sorting through records of navigation steps in the Web server log of a frequently used international multilingual website. The results demonstrated the impact of culture on website navigation behavior. Members of short-term oriented cultures spent less time on visited pages than members of long-term oriented cultures. In addition, more information is collected by members of high-uncertainty-avoidance countries than members of low-uncertainty-avoidance countries. Finally, monochromic cultures showed more linear navigation patterns than polychromic cultures and vice versa. They recommended that for monochromic users information should be placed in linear order and links should emphasize hierarchical structure. Table 8 provides guidelines for information architecture.

**Table 8 Guidelines for Information Architecture**

| Category | Guidelines | Supporting Research/ Best Practice |
| --- | --- | --- |
| Information architecture (navigation and hyperlinks) | When designing hypermedia, be aware of the structure of information and the navigation in association with the target user's cultural traits. | Choong (1996), Choong and Salvendy (1999), Marcus and Gould (2000), Luna et al. (2002), Cyr and Trevor-Smith (2004) |
| Information architecture (navigation and hyperlinks) | The navigation of a website should be designed to meet users' expectation by clearly indicating where they are, where they have been, and what they can access to and how they can proceed. | Luna et al. (2002) |
| Information architecture (navigation and hyperlinks) | Provide extra navigational aids for Japanese, Arabic, and Mediterranean users or users in high-context communication style. | Rau and Liang (2003a, 2003b) |
| Information architecture (navigation and hyperlinks) | The color of the hyperlinks should be designed considering the impact of local culture. | Lo and Gong (2005) |
| Information architecture (navigation and hyperlinks) | Provide top-oriented navigation model for Chinese sites. Provide left-, top-, and center-oriented navigation models for American sites. | Lo and Gong (2005) |
| Information architecture (navigation and hyperlinks) | Provide text-based navigation buttons for Chinese sites. Provide GUI-based navigation buttons for American sites. | Lo and Gong (2005) |
| Information architecture (navigation and hyperlinks) | Information should be placed in linear order, and links should emphasize hierarchical structure for monochromic users. | Kralisch et al. (2005) |

### 3.2.2 Searching

Searching is a key element in Web design. Researchers have indicated the significance of searching mechanisms for Web design. Morkes and Nielsen (1997) suggested that designers should provide search mechanisms and structure information to facilitate focused navigation on all websites. They found that 79% of participants scanned text and only 16% read word for word. Users with different cultural background may have different needs for searching mechanisms. Most websites have two types of search mechanisms built in: Web directories and search engines.

Fang and Rau (2003) examined the effects of cultural differences between Chinese and Americans on the perceived usability and search performance of Web portal sites. They found that Chinese participants tended to use keyword search to start a task. If they failed after one or more trials, they would then try to browse the categories to complete the task. On the other hand, American participants tended to browse categories at the beginning of a task. If they failed, they might use keyword search to supplement category search.

Besides color and graphics, discussed earlier, which are affective surface characteristics, the searching outcomes and user satisfaction associated with them also influence the user's affective experience. Fang and Rau (2003) found that Chinese participants were less satisfied with their searching performance than their American counterparts, even though no significant difference was found on their browsing performance for most of the searching tasks. The differences in consequence attribution of American and Chinese users may explain their differences in satisfaction. The Chinese tend to attribute consequence of events more internally than the Americans. The Chinese participants might think that if they had tried harder or had paid more attention than they did in the test, they would have done better. Therefore, providing possible outcomes and results of operations as much as possible is recommended for Asian users.

Kralisch and Berendt (2004) studied the searching behavior on websites for different cultures. They found that cultural dimensions, in particular amount of information needed and the perception of time and space, have an impact on the users' search behavior. The differences in search behavior are likely to be caused by the inherent thinking patterns determined by different cultural backgrounds. Website providers offering information to an international audience should take these results into consideration when designing search options and information access on their websites. The results of their study showed a clear stronger preference for search engines among the high-uncertainty-avoidance (UA) group and stronger preference for content-organized links among the low-UA group. The higher use of search engines among the high-UA group is consistent with the higher amount of information needed by these users. Table 9 provides guidelines for searching.

**Table 9 Guidelines for Searching**

| Category | Guidelines | Supporting Research/ Best Practice |
|---|---|---|
| Searching | Inherent thinking patterns determined by different cultural background influence the search behaviors. | Kralisch and Berendt (2004) |
| Searching | Be aware if user's satisfaction of searching is influenced by cultures. | Fang and Rau (2003) |

## 3.3  Mobile Computing

### 3.3.1  Usage Behaviors

Some recent studies investigated how mobile phone usage differs between cultures and how to design mobile services accordingly. Two studies designed mobile entertainment services in this way. The culture in Latin America focuses on interacting with other people. People there "like to contact with people, listen to music on loudspeakers and avoid being alone" (Otero et al., 2010). Therefore, mobile phones should focus on social connection when isolated (Otero et al., 2010). The culture difference also exits in different regions of China. Northern China is more traditionally interdependent while Eastern China is more independent. This results in different acceptance behaviors toward mobile entertainment service. Rural people in Northern China are most influenced by social influence while those in Eastern China are most influenced by self-efficacy (Liu et al., 2010). Similarly, a difference in mobile browsing is found. Kaikkonen (2008) found differences in mobile phone browsing between people in Asia and other continents. Users in Asia were less technical and preferred a mobile tailored Web. In contrast, users in North America and Europe are more technical and preferred full Web content.

Nokia researchers Nettamo et al. (2006) studied how users in New York and Hong Kong acquired, retrieved, consumed, and shared music on mobile devices. The Internet and one's friends played important roles in discovering and acquiring music in both cultures and music was shared in both cultures by means of emails and instant messaging. However, in New York, emphasis was placed on mobile music as a means of emphasizing individualism, while in Hong Kong the emphasis was on music as a means of bonding with friends. Also, the New York participants retrieved a greater variety of sources and carried a larger selection of music with them compared to the Hong Kong participants. They tended to listen to their mobile music player when commuting to create a private space, whereas in Hong Kong mobile music was valued primarily for entertainment value. Owning a device of a specific brand, such as the Apple iPod, was more important in New York than in Hong Kong. In the latter location, the style and industrial design of the player in general were perceived as more significant than any given brand.

According to an online survey with a large sample size of 3518 respondents from Korea, Hong Kong, and Taiwan (Kim et al., 2004), four cultural factors, that is,

uncertainty avoidance, individualism, contextuality, and time perception, have significant influences on users' postadoption perceptions of mobile Internet services. The service may offer functions with limited options and free trials if the user has a strong inclination to uncertainty avoidance (Lee et al., 2007).

### 3.3.2  Acceptance of Mobile Phones and Services

Many studies have analyzed culture traits of one specific culture, but few have compared different cultures. One important study investigated purchase intentions. Mobile phones are an important item with which to express oneself. Self-expression is different among different cultures. For example, Filipinos with high income or those at a lower social status tend to be flashy in their lifestyle, especially in dressing up. The flamboyance influences their choice of personal items. In contrast, Singaporeans dress up simply and do not outwardly demonstrate an extravagant way of living. Seva and Helander (2009) investigated how the cultural differences between Filipinos and Singaporeans influence their intention to purchase mobile phones. In the field experiment at mobile phone stores, participants were asked to choose one positively attractive and one negatively attractive mobile phone. Then they filled in questionnaires to indicate their affect and rate purchase intention. The results indicated that culture differences do influence emotional responses. Aesthetic attributes influence prepurchase attraction of Filipinos to a particular phone design, while functional attributes (e.g. display area, weight thickness) influence the Singaporeans.

Many studies have focused on cross-cultural design of icons in mobile phones. This is very important to launch mobile phones in the international market. The metaphors and contextual information for the objects to be represented as icons should consider the local culture (Krisnawati and Restyandito, 2008). Cultural differences between Americans and Koreans influence mobile phone icon styles. Among abstract, semiabstract, and concrete icons, Koreans performed significantly better with concrete icons while Americans showed the opposite tendencies (Kim and Lee, 2005). One example in this area is the Apple iPhone. To tailor to users in China, India, and the United States, Oren et al. (2009) investigated how easily users in each country found the icons and redesigned the current Apple iPhone icons based on their investigation. Since there are solutions for different cultures, is it possible to design for all? Pappachan and Ziefle (2008) answered affirmatively.

**Table 10 Guidelines for Mobile Computing**

| Category | Guidelines | Supporting Research/ Best Practice |
|---|---|---|
| Mobile computing | Be aware of the difference of culture on using mobile services. | Kim et al. (2004); Nettamo et al. (2006); Seva and Helander (2009); Otero et al. (2010); Liu et al. (2010) |

Regardless of cultural backgrounds, people can interpret icons better if an icon is more concrete and contains more information. They investigated mobile phone icon design differences between India and Germany. The results identified two critical issues for cross-cultural design of icons, domain knowledge and cultural-specific concept. Table 10 provides guidelines for mobile computing.

## 4 CONCLUSION

Prior to the Year 1995, relatively little attention in human factors design was devoted to consideration of cross-cultural issues. The research and design guidelines that were available at that time focused on surface issues such as the use of colors, symbols, language, and numbers. Fernandes's 1995 book, *International User Interfaces*, remains the classic design reference for many of these issues. Prabhu and Harel's (1999) extensive study of cross-cultural design issues in Japan and China also was a benchmark in this line of human factors research.

By the Year 1999, concomitant with the explosive growth in the use of mobile phones, the Internet, and online services, research in cross-cultural design greatly expanded. That research took two directions. One direction, typified by Marcus (2001), extrapolated user interface design guidelines from the classic work of Hofstede (1991) on cultural differences in attitudes and values and the work of Hall (1976, 1984, 1989, 1990) on time orientation and communication. A significant amount of research has followed Marcus's lead in that direction. A second and relatively less explored direction of cross-cultural design research has focused on cultural differences in cognition. Choong and Salvendy's (1999) research on cultural differences in user interface information structure launched this line of research. Nisbett's basic research (Nisbett et al., 2001) on the broad differences in cognitive style between Confucian and Aristotelian cultures has made us aware of the significant role these differences might play in user interface design for users in the East and the West. Unfortunately, very little research on these cognitive issues in cross-cultural user interface design has followed. It remains a relatively unexplored, yet extremely vital area for cross-cultural design research in the future.

Finally, many new types of technologies, products, and services are still in their infancy but are just beginning to emerge on the global market. Among the most important ones are those associated with global megatrends such as energy management and health care

accessibility. All of these new initiatives will succeed only if they are designed with the needs, preferences, anthropometry, and information-processing styles of their intended users in different cultures.

For example, online health information systems have great potential to reach users in areas of the world that are underserved with local medical services. However, across different cultures, the concepts of disease, symptomology, treatment, and even the human body itself vary widely. An online health information system based on Western concepts of medicine will most certainly fail if launched in China. It is doubtful that all or even many of the concepts of Western medicine could be expressed in the Chinese language. Even more importantly, Chinese users would have difficulty searching and understanding an online system based on unfamiliar concepts of medicine. Carefully aligning the underlying information model and structure in the online health information system with the concepts of the target culture will result in a much more highly usable and successful health information service.

Another example, smart home energy management, also has great promise to help solve another global problem, the difficulty in matching energy supplies with energy demands. The human factors problem begging to be solved here is that of consumer motivation. Wide variations exist in how consumers, both within and between cultures, perceive and define comfort and convenience. Further, there is great variation in consumer tolerance for trading off comfort and convenience against energy cost savings. Research shows that monetary rewards for more careful energy management increase compliance with conservation programs only to a certain level but then reache an asymptote (Wilson and Dowlatabadi, 2007). To motivate homeowners to higher energy conservation levels, other incentives must be used (Peterson et al., 2009; Faruqui et al., 2009; McMakin et al., 2002). The nature of these incentives will depend on the values, attitudes, and lifestyle of the target culture. For example, one might speculate that in a highly collectivist culture such as China effective incentives might be found in appeals to the common good of neighborhood, community, and country. In a highly individualist culture such as the United States, other incentives would likely have to be designed. In all cases, one cannot assume that incentives that are effective in one culture will be effective in another. The success of home energy management on a global scale absolutely depends on understanding the target cultures, their values, attitudes, lifestyles, and needs, and designing an incentive system that is compatible with them.

**APPENDIX: SUMMARY OF ANTHROPOMETRIC DATABASE FOR DIFFERENT NATIONS (ONLINE RESOURCES)**

| Database | Institute/Author | Nationality | Age range | Dimension | Data | Methodology | Link | Product |
|---|---|---|---|---|---|---|---|---|
| CAESAR® | Government and industry | North American; European | 18–65, civilian | 1D 3D | 40 body dimensions | From April 1998 to early 2000, the project collected data on 2400 U.S. and Canadian and 2000 European civilians. | http://store.sae .org/caesar/ | CAESAR™ |
| Size UK | Collaboration of the U.K. government, 17 major U.K. retailers, leading academics, and technology companies | United Kingdom population | 16–90⇓, civilian | 3D | 130 body measurements | The U.K. National Sizing Survey collected data on 11,000 subjects from three geographic regions. | http://www .size.org/ | Sizemic |
| Size USA | [TC]2 | U.S. | Civilian | 3D | — | Size USA collected data of nearly 11,000 individuals in 12 locations across the U.S. | http://www .sizeusa.com/ | SizeUSA |
| ERGODATA | Laboratory of Applied Anthropology, France | North American, European, Asian, Central American, South American, African, Oceania | Civilian | 1D 3D in the development process | Body measurements; biostereometric data | The Individual data file (DI) including 64 inquiries, with 60,000 subjects coming from 20 different countries. The Aggregate data file (DA) including 160 articles, which represent more than 1 million subjects, with 60 represented countries. Tridimensional data file (3D) which is in the development process. | http://www .biomedicale .univ-paris5 .fr/LAA/ eindex.htm | Free |

| Database | Institute/Author | Nationality | Age range | Dimension | Data | Methodology | Link | Product |
|---|---|---|---|---|---|---|---|---|
| Size China | Hong Kong Polytechnic University | Chinese | 18–70, civilian | 3D | Head and face size | From 2006 to 2007, the project collected data on 2000 male and female volunteers in six diverse locations in China. | http://www.sizechina.com/html/index.html; http://www.certiform.org/products.html | Certiform™ |
| Size Japan | Research Institute of Human Engineering for Quality Life | Japanese | Civilian | 3D | Body measures | — | http://www.hql.jp/ | — |
| WEAR | Global partners | Depends on different studies | Depends on different studies | Depends on different studies | Depends on different studies | WEAR is a group of interested experts involved in the application of anthropometry data for design purposes. | http://wear.io.tudelft.nl/ | Free |
| PeopleSize 2008 | Open Ergonomics Ltd. | American, Australian, Belgian, British, Chinese, French, German, Japanese, Swedish | Infant, child, and adult | 1D | Body measures | 289 individual body measurement dimensions | http://www.openerg.com/psz/index.html | PeopleSize |

| Database | Institute/Author | Nationality | Age range | Dimension | Data | Methodology | Link | Product |
|---|---|---|---|---|---|---|---|---|
| DINED | Delft University of Technology | Dutch; International: North American; Latin American (Indians); Latin American (rest); North Europe; Central Europe; Eastern Europe; Southeast Europe; France; Spain and Portugal; North Africa; West Africa; Southeast Africa; Middle East; North India; South India; North India; South China; South East Asia; Australia (European); Japan) | 2–80⇓ Dutch civilian | 1D | Body measures, force exercise, and joint excursion | Long-term project collecting data from different nations. Latest version for Dutch adults is dined 2004, for Dutch elderly is geron 1998, Dutch children is kima 1993. The version for international is 1989. | www.dined.nl | DINED, free |
| AnthroKids | Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST) and the CPSC | North American | Children | 1D | Body measurements | Two studies performed in the years 1975 and 1977 | http://ovrt.nist .gov/projects/ anthrokids/ | Free |
| DINBelg 2005 | Several Belgium schools | Belgian | 2–80, civilian | 1D | Body measurements | The aim of DINBelg 2005 is to gather up-to-date anthropometric dimensions of the Belgian population. | http://www .dinbelg.be/ anthropometry .htm | Free |

## REFERENCES

Anderson, R., Anderson, R., and Deibel, K. (2004, November), "Analyzing Concept Groupings of Introductory Computer Programming Students," Technical Report, University of Washington, Seattle, WA.

Axtell, R. E. (1991), *Gestures: The DO's and TABOOs of Body Language around the World*, Wiley, New York.

Barter, J. T. (1957), "Estimation of the Mass of Body Segments," WADC Technical Report, pp. 57–260. Wright Air Development Center, Wright-Patterson Air Force Base, OH.

Berlin, B., and Kay, P. (1969), *Basic Color Terms*, University of California Press, Berkeley and Los Angeles.

Bluedorn, A. C., Kalliath, T. J., Strube, M. J., and Martin, G. D. (1999), "Polychronicity and the Inventory of Polychronic Values (IPV): The Development of an Instrument to Measure a Fundamental Dimension of Organizational Culture," *Journal of Managerial Psychology*, Vol. 14, Nos. 3–4, pp. 205–230.

Bonds, M. H. (1986), *The Psychology of the Chinese People*, Oxford University Press, New York.

Bosland, N. (1985), *An Evaluation of Replication Studies Using the Value Survey Model*, Institute for Research on Cultural Cooperation, Tilburg.

Brislin, R. W. (1970), "Back-Translation for Cross-Cultural Research," *Journal of Cross Cultural Psychology*, Vol. 1, No. 3, pp. 185–216.

Burmeister, O. K. (2000), "Usability Testing: Revisiting Informed Consent Procedures for Testing Internet Sites," paper presented at the 2nd Australian Institute of Computer Ethics Conference, Canberra, Australia, December 2000.

Card, S. K., Moran, T. P., and Newell, A. L. (1983), *The Psychology of Human Computer Interaction*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Chavan, A. L. (2005), "Another Culture, Another Method," in *Proceedings of the 11th International Conference on Human Computer Interaction*, Lawrence Erlbaum Associates, Mahwah, NJ.

Chiu, L. H. (1972), "A Cross-Cultural Comparison of Cognitive Styles in Chinese and American Children," *International Journal of Psychology*, Vol. 7, No. 4, pp. 235–242.

Choong, Y. Y. (1996, August), "Design of Computer Interfaces for the Chinese Population," Ph.D. Dissertation, Purdue University, West Lafayette, IN.

Choong, Y. Y., and Salvendy, G. (1998), "Design of Icons for Use by Chinese in Mainland China," *Interacting with Computers*, Vol. 9, pp. 417–430.

Choong, Y. Y., and Salvendy, G. (1999), "Implications for Design of Computer Interfaces for Chinese Users in Mainland China," *International Journal of Human Computer Interaction*, Vol. 11, No. 1, pp. 29–46.

Choong, Y., Stanton. B., and Theofanos, M. (2010). Biometric symbol design for the public—case studies in the United States and four Asian countries. In Kaber, D. B., and Boy, G. (Ed.), *Advances in Cognitive Ergonomics*, New York: Taylor & Francis Group, p. 1–12.

Chu, P. C., Spires, E. E., Farn C. K., and Sueyoshi, T. (2005), "Decision Processes and Use of Decision Aids—Comparing Two Closely Related Nations in East Asia," *Journal of Cross-Cultural Psychology*, Vol. 36, No. 3, pp. 304–320.

Clemmensen, T., Shi, Q., Kumar, J., Li, H., Sun, X. H., and Yammiyavar, P. (2007), "Cultural Usability Tests—How Usability Tests Are Not the Same All over the World," in *Usability and Internationalization, Part I, HCII 2007*, N. Aykin, Ed., LNCS 4559, 2007. Springer-Verlag, Berlin and Heidelberg, pp. 281–290.

Clemmensen, T., Hertzum, M., Hornbæk, K., Shi, Q., and Yammiyavar, P. (2009), "Cultural Cognition in Usability Evaluation," *Interacting with Computers*, Vol. 21, No. 3, pp. 212–220.

Courtney, A. J. (1986), "Chinese Population Stereotypes: Color Associations," *Human Factors*, Vol. 28, No. 1, pp. 97–99.

Cyr, D., and Trevor-Smith, H. (2004), "Localization of Web Design: An Empirical Comparison of German, Japanese, and United States Web Site Characteristics," *Journal of the American Society for Information Science and Technology*, Vol. 55, No. 13, pp. 1199–1208.

Davies, I. R., and Corbett, G. G. (1997), "A Cross-Cultural Study of Colour Grouping: Evidence for Weak Linguistic Relativity," *British Journal of Psychology*, Vol. *88*, pp. 493–517.

del Galdo, E., and Nielsen, J. (1996), *International User Interfaces*, Wiley, New York.

Dong, J., and Salvendy, G. (1999), "Designing Menus for the Chinese Population Horizontal or Vertical?" *Behaviour & Information Technology*, Vol. 18, No. 6, pp. 467–471.

Ekman, P., and Friesen, W. V. (1969), "The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding," *Semiotica*, Vol. 1, pp. 49–98

Evers, V. (1998), "Cross-cultural Understanding of Metaphors in Interface Design," in *Proceedings CATAC'98, Cultural Attitudes towards Technology and Communication*, C. Ess and F. Sudweeks, Eds., 1–3 August, Science Museum, London. University of Sydney, Australia

Evers, V. (2002), "Cross-Cultural Applicability of User Evaluation Methods: A Case Study amongst Japanese, North-American, English and Dutch Users," paper presented at the ACM Conference on Human Factors in Computing Systems (CHI) 2002, Minneapolis, MN, April 20–25, 2002.

Fang, X., and Rau, P. L. P. (2003), "Culture Differences in Design of Portal Sites," *Ergonomics*, Vol. 46, Nos. 1–3, pp. 242–254.

Farh, J. L., Zhong, C. B., and Organ, D. W. (2004), "Organizational Citizenship Behavior in the People's Republic of China," *Organization Science*, Vol. 15, pp. 241–253.

Faruqui, A., Sergici, S., and Sharif, A. (2009, May), "The Impact of Informational Feedback on Energy Consumption—A Survey of the Experimental Evidence," available: http://ssrn.com/abstract=1407701.

Fernandes, T. (1995), *Global Interface Design: A Guide to Designing International User Interfaces*, AP Professional, Chestnut Hill, MA.

Frake, C. (1980), "The Ethnographic Study of Cognitive System," in *Language and cultural descriptions*, C. Frake, Ed., Stanford University Press, Stanford, CA, pp. 1–17.

Frandsen-Thorlacius, O., Hornbæk, K., Hertzum, M., and Clemmensen, T. (2009), "Non-Universal Usability?: A Survey of How Usability Is Understood by Chinese and Danish Users," in *Proceedings of the CHI 2009 Conference on Human Factors in Computing Systems*, Boston, MA, April 5–9, pp. 41–50. ACM Press, New York.

Fukuoka, W., Kojima, Y., and Spyridakis, J. H. (1999), "Illustrations in User Manuals: Preference and Effectiveness with Japanese and American Readers," *Technical Communication,* Vol. 46, No. 2, pp. 167–176.

Galitz, W. O. (1997), *The Essential Guide to User Interface Design*, Wiley, New York.

Gauvain, M. (1993), "The Development of Spatial Thinking in Everyday Activity," *Developmental Review*, Vol. 13, pp. 92–121.

Geary, D. C., Fan, L., and Bow-Thomas, C. C. (1992), "Numerical Cognition: Loci of Ability Differences Comparing Children from China and the United States," *Psychological Science*, Vol. 3, pp. 180–185.

Gould, E. W. (2001), "More than Content: Web Graphics, Crosscultural Requirements, and a Visual Grammar," paper presented at the HCI International 2001 Conference, New Orleans, LA, August 2001.

Haase, R. F., Lee, D. Y., and Banks, D. L. (1979), "Cognitive Correlates of Polychronicity," *Perceptual and Motor Skills*, Vol. 49, pp. 271–282.

Hall, E. T. (1976), *Beyond Culture*, Doubleday, New York.

Hall, E. T. (1984), *Dance of Life: The Other Dimension of Time*, Intercultural Press, Yarmouth, ME.

Hall, E. T. (1989), *Beyond Culture*, Anchor, New York.

Hall, E. T. (1990), *The Hidden Dimension*, Anchor, New York.

Hall, E. T., and Hall, M. (1990), *Understanding Cultural Differences: Germans, French and Americans*, Intercultural Press, Yarmouth, ME.

Helmreich, R. L., and Merritt A. C. (1998), *Culture at Work in Aviation and Medicine*, Ashgate, Aldershot, United Kingdom.

Herman, L. (1996), "Toward Effective Usability Evaluation in Asia: Cross Cultural Differences", In *Proceedings of the 6th Australian Conference on Computer-Human Interaction*, pp. 135–136). Washington, DC: IEEE Computer Society.

Hertzum, M., Clemmensen, T., Hornbæk, K., Kumar, J., Shi, Q., and Yammiyavar, P. (2007), "Usability Constructs: A Cross-Cultural Study of How Users and Developers Experience Their Use of Information Systems," *Lecture Notes in Computer Science*, Vol. 4559, pp. 317–327.

Hoffman, C., Lau, I., and Johnson, D. R. (1986), "The Linguistic Relativity of Person Cognition: An English-Chinese Comparison," *Journal of Personality & Social Psychology*, Vol. 51, pp. 1097–1105.

Hofstede, G. (1991), *Cultures and Organizations: Software of the Mind. Intercultural Cooperation and Its Importance for Survival*, McGraw-Hill International, London.

Hoft, N. (1996), "Developing a Cultural Model," in *International User Interfaces*, E. del Galdo and J. Nielsen, Eds., Wiley Computer Publishing, New York, pp. 41–71.

Holzschlag, M. E. (1999), "Satisfying Customers with Color, Shape and Type," *Web Techniques*, Vol. 4, No. 11, pp. 24–29.

Hoosain, R. (1986), "Perceptual Processes of the Chinese," in *The Psychology of the Chinese People*, M. H. Bond, Ed., Oxford University Press, New York, pp. 38–72.

Horton, W. (1994), *The Icon Book: Visual Symbols for Computer Systems and Documentation*, Wiley, New York.

Hu, H., Li, Z., Yan, J., Wang, X., Xiao, H., Duan, J., and Li, Z. (2006), "Measurements of Voluntary Joint Range of Motion of the Chinese Elderly Living in Beijing Area by a Photographic Method," *International Journal of Industrial Ergonomics*, Vol. 36, pp. 861–867.

International Organization for Standardization (ISO) (1998), *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)—Part 12: Presentation of Information*, ISO 9241-12: 1998 (E), ISO, Geneva.

International Organization for Standardization (ISO) (1997), *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)—Part 14: Menu Dialogues*, ISO 9241-14: 1997 (E), ISO, Geneva.

Ji, L., Peng, K., and Nisbett, R. E. (2000), "Culture, Control, and Perception of Relationships in the Environment," *Journal of Personality and Social Psychology*, Vol. 78, pp. 943–955.

Kaikkonen, A. (2008), "Full or Tailored Mobile Web—Where and How Do People Browse on Their Mobiles?" paper presented at the International Conference on Mobile Technology, Applications, and Systems: Proceedings of the International Conference on Mobile Technology, Applications, and Systems, Yilan, Taiwan.

Kaiser, J. (2002), "Web Design and Color Symbolism," available: http://webdesign.about.com/mbody.htm.

Kaufman-Scarborough, C., and Lindquist, J. D. (1999), "Time Management and Polychronicity: Comparisons, Contrasts, and Insights for the Workplace," *Journal of Managerial Psychology*, Vol. 14, Nos. 3–4, pp. 288–312.

Kim, J., Lee, I., Lee Y., Choi, B., Hong, S., Tam, K., Naruse, K., and Maeda, Y. (2004), "Exploring E-Business Implications of the Mobile Internet: A Cross-National Survey of Hong Kong, Japan and Korea," *International Journal of Mobile Communication*, Vol. 2, No.1, pp. 1–21.

Kim, J. H., and Lee, K. P. (2005), "Cultural Difference and Mobile Phone Interface Design: Icon Recognition According to Level of Abstraction," paper presented atthe 7th International Conference on Human Computer Interaction with Mobile Devices and Services, Salzburg, Austria, September 19–22, 2005.

Kim, J. H., Lee, K. P., and You, I. K. (2007), "Correlation between Cognitive Style and Structure and Flow in Mobile Phone Interface: Comparing Performance and Preference of Korean and Dutch Users," in *Usability and Internationalization*, N. Aykin, Ed., Part 1, HCII 2007, LNCS 4559, Springer-Verlag, Heidelberg, pp. 531–540.

Kralisch, A., and Berendt, B. (2004), "Cultural Determinants of Search Behaviour on Websites," in *Proceedings of the IWIPS 2004 Conference on Culture, Trust and Design Innovation*, Vancouver, Canada, 8–10 July, 2004, pp. 61–74), Product & Systems Internationalisation, Inc., Vancouver, Canada.

Kralisch, A., Berendt, B., and Eisend, M. (2005), "Impact of Culture of Website Navigation Behaviour," in *Proceedings of the 11th International Conference on Human-Computer Interaction*, Lawrence Erlbaum Associates, Mahwah, NJ.

Krishnan, K., Plocher, T. A., and Garg, C. (1999), "Cross-Cultural Issues in Commercial Aviation," White Paper, Honeywell Technology Center.

Krisnawati, L. D., and Restyandito (2008), "Localized User Interface for Improving Cell Phone Users' Device Competency," *International Journal of Information Technology and Web Engineering*, Vol. 3, No. 1, pp. 38–52.

Kroemer, K. H. W., Kroemer, H. J., and Kroemer-Elbert, K. E. (1990), *Engineering Physiology, Bases of Human Factors/Ergonomics*, Van Nostrand Reinhold, New York.

Lee, K. P., and Harada, A. (1999, September), "A Study of the Cultural Effects on User-Interface Design: With the Emphasis on the Cross-Cultural Usability Testing through World Wide Web," *Proceedings of the 4th Asian Design Conference* Nagaoka, Japan.

Lee, I., Choi, B., Kim, J., and Hong, S. (2007), "Culture-Technology Fit: Effects of Cultural Characteristics on the Post-Adoption Beliefs of Mobile Internet Users,"

*International Journal of Electronic Commerce*, Vol. 11, No. 4, pp. 11–51.

Lehto, M., and Buck, J. (2008), *Introduction to Human Factors and Ergonomics for Engineers*, Lawrence Erlbaum, New York.

Li, Y. Y., Yang, K. S., and Wen, C. Y. (1985), *Modernization and Chinese Characteristics*, Kuei-Kuan Publishing, Taipei, China.

Liang, S-F. M., and Plocher, T. (2003), "Towards a Cross-Cultural Human-Computer Interaction," paper presented at the International Ergonomics Association XVth Triennial Congress, Seoul, South Korea, August 2003.

Liang, S-F. M., Plocher, T. A., Lau, P. W. C., Chia, Y. T. B., Rafi, N., and Tan, T. H. R. (2000), "Cross-Cultural Stereotypes of Colors and Graphics in Honeywell Process Control Workstations," in *Proceedings of the 4th Asia Pacific Conference on Computer Human Interaction (APCHI 2000) and 6th South East Asian Ergonomics Society Conference (ASEAN Ergonomics 2000)*, Lim, K. Y., Ed, 27 November–December, Singapore, Elsevier, pp. 120–124.

Lin, Y. C., Wang, M. J. J., and Wang, E. M. (2004), "The Comparisons of Anthropometric Characteristics among Four Peoples in East Asia," *Applied Ergonomics*, Vol. 35, No. 2, pp. 173–178.

Lindquist, J. D., Knieling, J., and Kaufman-Scarborough, C. (2001), "Polychronicity and Consumer Behavior Outcomes among Japanese and U.S. Students: A Study of Response to Culture in a U.S. University Setting," in *Proceedings of the Tenth Biennial World Marketing Congress*, H. Spotts and H. L. Meadow, Eds, Cardiff, Wales, Great Britain, Academy of Marketing Science, FL.

Liu, I. M. (1986), "Chinese Cognition," in *The Psychology of the Chinese People*, M. H. Bonds, Ed., Oxford University Press, New York, pp. 73–106.

Liu, J., Liu, Y., Rau, P. L. P., Li, H., Wang, X., and Li, D. (2010), "How Socio-Economic Structure Influences Rural Users' Acceptance of Mobile Entertainment," in *Proceedings of the 28th ACM Conference on Human Factors in Computing Systems (CHI 2010)*, Atlanta, Georgia, April 12–April 15, 2010 E. D. Mynatt, D. Schoner, G. Fitzpatrick, S. E. Hudson, W. K. Edwards, and T. Rodden, Eds., ACM New York.

Lo, B. W. N., and Gong, P. (2005), "Cultural Impact on the Design of e-Commerce Websites: Part I—Site Format and Layout," *Issues in Information Systems*, Vol. 6, No. 2, pp. 182–188.

Logan, R. K. (1987), *The Alphabet Effect: The Impact of the Phonetic Alphabet on the Development of Western Civilization*, St. Martin's Press, New York.

Luna, D., Peracchio, L. A., and deJuan, M. D. (2002), "Cross-Cultural and Cognitive Aspects of Web Site Navigation," *Journal of the Academy of Marketing Science*, Vol. 30, No. 4, pp. 397–410.

Luong, T. V., Lok, J. S. H., Taylor, D. J., and Driscoll, K. (1995), *Internationalization: Developing Software for Global Markets*, Wiley, New York.

Luximon, A., Lau, W. C., and Goonetilleke, R. S. (1998), "Safety Signal Words and Color Codes: The Perception of Implied Hazard by Chinese People," paper presented at the 6th Pan-Pacific Conference on Occupational Ergonomics, Beijing, China.

MacLaury, R. E. (1991), "Social and Cognitive Motivations of Change: Measuring Variability in Color Semantics," *Language*, Vol. 67, No. 1, pp. 34–62.

Marcus, A. (2001), "Cross-Cultural User-Interface Design", in *Proceeding of the Human-Computer Interface International (HCII) Conference*, Vol. 2, M. J. Smith and G. Salvendy, Eds., Lawrence Erlbaum Association, New Orleans, pp. 502–505.

Marcus, A., and Gould, E. M. (2000, July/August), "Cross-currents: Cultural Dimensions and Global Web User-Interface Design," *Interactions*, Vol. 7, No. 4, pp. 32–46.

McMakin, A. H., Malone, E. L., and Lundgren, R. E. (2002, November), "Motivating Residents to Conserve Energy without Financial Incentives," *Environment and Behavior Journal*, Vol. 34-6, pp. 848–863.

Miller, G. (1981), *Language and Speech*, Freeman, San Francisco.

Minocha, S., French, T., and Smith, A. (2002), "eFinance Localisation: An Informal Analysis of Specific eCulture Attractors in Selected Indian and Taiwanese Sites," paper presented at the 4th International Workshop on Internationalization of Products and Systems, Austin, TX, 11–13 July.

Mishra, R. C. (1997), "*Cognition and Cognitive Development*," in *Handbook of Cross-Cultural Psychology*, Vol. 2, *Basic Processes and Human Development*, J. W. Berry, P. R. Dasen, and T. S. Saraswathi, Eds., Allyn & Bacon, Boston, pp. 143–176.

Morkes, J., and Nielsen, J. (1997), Concise, Scannable and Objective: How to Write for the Web, available: http://www.useit.com/papers/webwriting/writing.html, accessed October, 1 2010.

Morton, J. (2003), "Color Matters," available: http://www.colormatters.com, accessed October 1, 2010.

National Aeronautics and Space Administration (NASA) (2010), *Human Integration Design Handbook*, NASA, Washington, DC.

National Technical Committee of Ergonomics Standard (2009), *Ergonomics Standards: Anthropometry and Biomechanics* (in Chinese), China Standards Publishing House, Beijing.

Nawaz, R., Plocher, T., Clemmensen, T., Qu, W., and Sun, X. (2007), "Cultural Differences in the Structure of Categories in Denmark and China," working paper 03-2007, Department of Informatics, Copenhagen Business School, Copenhagen.

Nettamo, E., Nirhamo, M., and Hakkila, J. (2006), "A Cross-Cultural Study of Mobile Music—Retrieval, Management and Consumption", paper presented atthe Australian Conference on Computer-Human Interaction (OZCHI 2006), Sydney, Australia, November 20–24, 2006.

Newell, A., and Simon, H. (1972), *Human Problem Solving*, Prentice Hall, Englewood Cliffs, NJ.

Nielsen, J. (1990), "Usability Testing of International Interfaces," in *Advances in Human Factors/Ergonomics*, Vol. 13, *Designing User Interfaces for International Use*, J. Nielson, Ed., Elsevier Science, New York, pp. 39–44.

Nielsen, J. (2003), "International Usability Testing," available: http://www.useit.com/papers/international_usetest.html.

Nisbett, R. E. (2003), *The Geography of Thought: How Asians and Westerners Think Differently and Why*, Free Press, New York, NY.

Nisbett, R. E., Peng, K., Choi, I., and Norenzayan, A. (2001), "Culture and Systems of Thought: Holistic vs. Analytic Cognition," *Psychological Review*, Vol. 108, pp. 291–310.

Niu, J., Zhang, Y., and Wu, J. (2009), "Challenges in the Application of Three-Dimensional Anthropometry,"

paper presented at the 4th IEEE Conference on Industrial Electronics and Applications, May 25–27, 2009, pp. 864–868.

Niu, J., Zhu, L., Yan, Q., Liu, Y., and Wang, K. (2010), "Stroke++: A Hybrid Chinese Input Method for Touch Screen Mobile Phones," paper presented at the *International Conference on Human-Computer Interaction with Mobile Devices and Services* (MobileHCI'10), Lisbon, Portugal, September 7–10, 2010, ACM.

Norman, D. A., and Rumelhart, D. E. (1970), "A System for Perception and Memory," in D. A. Norman, Ed., *Models of Human Memory*, Academic, New York, pp. 19–64.

Oren, M. A., Seth, U., Huang, F., and Kang, S. (2009), "Cross-Cultural Design and Evaluation of the Apple iPhone," in *Proceedings of the 3rd International Conference on Internationalization, Design and Global Development* in the 13th International Conference on Human-Computer Interaction, Aykin, N. Ed., San Diego, CA, pp. 79–88.

Osgood, C. E., May, W. H., and Miron, M. S. (1975), *Cross-Cultural Universals of Affective Meaning*, University of Illinois Press, Urbana, IL.

Otero, Y., Herrera, M., Maehr, W., and Castillo, M. I. (2010), "Mobile Internet User Experience in Latin America," available: http://research.nokia.com/files/Otero-Latin America.pdf.

Pappachan, P., and Ziefle, M. (2008), "Cultural Influences on the Comprehensibility of Icons in Mobile-Computer Interaction," *Behaviour & Information Technology*, Vol. 27, No. 4, pp. 331–337.

Peterson, D., Steele, J., and Wikerson, J. (2009), "WattBot: A Residential Electricity Monitoring and Feedback System," paper presented at the *ACM Conference on Human Factors in Computing Systems (CHI 2009),* Boston, MA, April 4–9, 2009, ACM.

Pheasant, S., and Haslegrave, C. M. (2006), *Bodyspace: Anthropometry, Ergonomics, and the Design of Work*, 3rd ed., Taylor & Francis, Boca Raton, FL.

Plocher, T. A., Zhao, C., Liang, S. M., Sun, X., and Zhang, K. (2001), "Understanding the Chinese User: Attitudes Toward Automation, Work, and Life," in *Proceedings of the 9th International Conference on Human-Computer Interaction*, S. Constantine, Ed., New Orleans, LA, August, 2001, Lawrence Erlbaum Associates, Mahwah, NJ.

Prabhu, G., and Harel, D. (1999), "GUI Design Preference Validation for Japan and China—A Case for KANSEI Engineering," paper presented at the 8th International Conference on Human-Computer Interaction (HCI99), Munich, Germany.

Rau, P.-L. P. (2001), "Cross-Cultural User Interface Research and Design with Emphasis on Asia: Chinese Users in Taiwan," final report to Honeywell Singapore Laboratory, October 2001.

Rau, P.-L. P., and Liang, S-F. M. (2003a), "A Study of Cultural Effects on Designing User Interface for a Web-Based Service," *International Journal of Services Technology and Management*, Vol. 4, Mos. 4–6, pp. 480–493.

Rau, P.-L. P., and Liang, S-F. M. (2003b), "Internationalization and Localization: Evaluating and Testing a Web Site for Asian Users," *Ergonomics*, Vol. 46, Nos. 1–3, pp. 255–270.

Rau, P.-L. P., Choong, Y-Y., and Salvendy, G. (2004), "A Cross Culture Study of Knowledge Representation and Interface Structure in Human Computer Interface," *International Journal of Industrial Ergonomics*, Vol. 34, No. 2, pp. 117–129.

Rau, P.-L. P., and Liu, J. (2010). Design and Evaluate Biometric Device Symbols for Chinese. *Proceedings of the 3rd International Conference on Applied Human Factors and Ergonomics (AHFEI 2010)*, Miami, Florida, USA, July 17–20, 2010.

Rau, P-L. P., Plocher, T., and Choong, Y-Y. (2010), "Cross-Cultural Web Design," in *Handbook of Human Factors in Web Design*, 2nd ed., R. W. Proctor and K-P. L. Vu, Ed., CRC Press, Mahwah, NJ.

Russell, P. A., Deregowski, J. B., and Kinnear, P. R. (1997), "Perception and Aesthetics," in *Handbook of Cross-Cultural Psychology*, 2nd ed., Vol. 2, J. W. Berry, P. R. Dasen, and T. S. Saraswathi, Eds., Allyn & Bacon, Boston, MA, pp. 107–142.

Salgado, L. C. C., Souza, C. S., and Leitao, C. F. (2009), "Conceptual Metaphors for Designing Multi-Cultural Applications," in *Proceedings of the 2009 Latin American Web Congress (la-web 2009)*, E. Chavez, Eds., Computer Society Press, Merida, mexico, pp. 105–111.

Sampson, G. (1990), *Writing Systems: A Linguistic Introduction*, Stanford University Press, Stanford, CA.

Schaffer, B. S., and Riordan, C. M. (2003), " A Review of Cross-Cultural Methodologies for Organizational Research: A Best- Practices Approach," *Organizational Research Methods*, Vol. 6, No. 2, pp. 169–215.

Seva, R. R., and Helander, M. G. (2009), "The Influence of Cellular Phone Attributes on Users' Affective Experiences: A Cultural Comparison," *International Journal of Industrial Ergonomics,* Vol. 39, No. 2, pp. 341–346.

Shen, S. T., Prior, S. D., Chen, K. M., and You, M. L. (2007), "Chinese Web Browser Design Utilising Cultural Icons," in *Usability and Internationalization, Part II, HCII 2007*, LNCS 4560, pp. 249–258, N. Aykin, Ed., Springer-Verlag, Berlin and Heidelberg.

Shi, Q. (2008), "A Field Study of the Relationship and Communication between Chinese Evaluators and Users in Thinking Aloud Usability Tests," paper presented at the Nordic Conference on Human-Computer Interaction (NordiCHI), Lund, Sweden, October 18–22, 2008.

Shi, Q. (2009), "An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective," Ph.D. Thesis, LIMAC PhD School, Programme in Informatics, Copenhagen Business School (CBS), Copenhagen, Denmark. November 2009.

Shiffrin, R. M., and Atkinson, R. C. (1969), "Storage and Retrieval Processes in Long-Term Memory," *Psychological Review*, Vol. 76, pp. 179–193.

Shiffrin, R. M., and Schneider, W. (1977), "Controlled and Automatic Human Information Processing : II. Perceptual Learning, Automatic Attending, and a General Theory," *Psychological Review*, Vol. 84, pp. 127–190.

Spartan (1999), "Multimedia Presentation on the Colors of Culture," available: http://www.mastep.sjsu.edu/Alquist/workshop2/color_and_culture_files/frame.htm, accessed September 30, 2010.

Spencer, C. P., and Darvizeh, Z. (1983), "Young Children's Place Descriptions, Maps and Route Finding: A Comparison of Nursery School Children in Iran and Britain," *International Journal of Early Childhood,* Vol. 15, No. 1, pp. 26–31.

Stewart, E. C., and Bennett, M. J. (1991), *American Cultural Patterns: A Cross-Cultural Perspective*, Intercultural, Yarmouth, ME.

Sun, X., and Shi, Q. (2007), "Language Issues in Cross Cultural Usability Testing: APilot Study in China", *Lecture Notes in Computer Science*, Vol. 4560, pp. 274–284.

Taylor, D. (1992), *Global Software: Developing Applications for the International Market*, Springer-Verlag, New York.

Trompenaars, F. (1993), *Riding the Waves of Culture*, Nicholas Brealy, London.

Vatrupu, R., and Pérez-Quiñones, M. A. (2006), "Culture and Usability Evaluation: The Effects of Culture in Structured Interviews," *Journal of Usability Studies*, Vol. 1, No. 4, pp. 156–170.

Weinschenk, S., Jamar, P., and Yeo, S. C. (1997), *GUI Design Essentials*, Wiley, New York.

Wickens, C. D., and Hollands, J. G. (2000), *Engineering Psychology and Human Performance*, Prentice Hall, Englewood Cliffs, NJ.

Wilson, C., and Dowlatabadi, H. (2007), "Models of Decision-making and Residential Energy Use," *Annual Review of Environmental Resources*, Vol. 32, pp. 169–203.

Whorf, B. L. (1956), *Language, Thought, and Reality*, MIT Press, Boston.

Yammiyavar, P., Clemmensen, T., and Kumar, J. (2008), "Influence of Cultural Background on Non-Verbal Communication in a Usability Testing Situation," Special Issue on Cultural Aspects of Interaction Design, *International Journal of Design*, Vol. 2, No. 2, pp. 31–40.

Yang, C. F. (1991), "'ziji' of Chinese: Theory and Research Direction (in Chinese)", in *Chinese People and Chinese Psychology—Society and Personality*, C. F. Yang, and S. R. Gao, Eds., Yuan-Liou Publishing, Taipei, pp. 93–145.

Yang, C. F. (2000a), "Toward a New Conceptualization of Guanxi and Renqing," *Indigenous Psychological Studies in Chinese Societies*, Vol. 12, pp. 105–179.

Yang, C. F. (2000b), "In the Wrong Places? Or with the Wrong People?: Commentary on 'In Search of the Chinese in All

the Wrong Places!'" *Journal of Psychology in Chinese Societies*, Vol. 1, pp. 153–158.

Yang, C. F. (2001a), *How to Study the Chinese: A Collection of Papers on the Indigenous Approach*, Yuan-Liou Publishing, Taipei.

Yang, C. F. (2001b), "The Chinese Conception of the Self: Toward a Person-Making Perspective," paper presented at Scientific Advances in Indigenous Psychologies: Philosophical, Cultural and Empirical Contributions, Taipei, Taiwan, October 27–30, 2001.

Yeo, A. (1998, April), "Cultural Effects in Usability Assessment, Doctoral Consortium," in *Proceedings of the Conference on CHI 98 Summary: Human Factors in Computing Systems*, C.-M. Karat, A. Lund, J. Coutaz, and J. Karat, Eds., Los Angeles, CA, ACM, pp. 71–75.

Yi, J., and Park, S. (2003), "Cross-Cultural Differences in Decision-Making Styles: A Study of College Students in Five Countries," *Social Behavior & Personality: An International Journal*, Vol. 31, No. 1, p. 35.

Zhang, Y., Goonetilleke, R. S., Plocher, T., and Liang, S. -F. M. (2004), "Time Orientation and Human Performance," in *Proceedings of the 7th International Conference on Work With Computing Systems*, Kuching, Sarawak, Malaysia, pp. 248–252.

Zhao, C. (2002, May), "Effect of Information Structure on Performance of Information Acquiring: A Study Exploring Different Time Behavior: Monochronicity/Polychronicity," Ph.D. Dissertation, Institute of Psychology, Chinese Academy of Sciences, Beijing, China.

Zhao, C., Plocher, T., Xu, Y., Zhou, R., Liu, X., Liang, S-F. M., and Zhang, K. (2002), "Understanding the Polychronicity of Chinese," in *Proceedings of the APCHI2002, 5th Asia-Pacific Conference on Computer Human Interaction*, Beijing, China, Vol. 1, pp. 189–195.

# CHAPTER 7

# DECISION-MAKING MODELS, DECISION SUPPORT, AND PROBLEM SOLVING

**Mark R. Lehto**
**Purdue University**
**West Lafayette, Indiana**

**Fiona Fui-Hoon Nah**
**University of Nebraska–Lincoln**
**Lincoln, Nebraska**

**Ji Soo Yi**
**Purdue University**
**West Lafayette, Indiana**

## 1 INTRODUCTION

This chapter focuses on the broad topic of human decision making. Decision making is often viewed as a stage of human information processing because people must gather, organize, and combine information from various sources to make decisions. However, as decisions grow more complex, information processing actually becomes part of decision making, and methods of decision support that help decision makers process information become of growing importance. Decision making also overlaps with problem solving. The point where decision making becomes problem solving is fuzzy, but many decisions require problem solving, and the opposite is true as well. Cognitive models of problem solving are consequently relevant for describing many aspects of human decision making. They become especially relevant for describing steps taken in the early stages of a decision where choices are formulated and alternatives are identified.

A complete treatment of human decision making is well beyond the scope of a single chapter.[*] The topic has its roots in economics and is currently a

---

[*] No single book covers all the topics addressed here. More detailed sources of information are referenced throughout the chapter. Sources such as von Neumann and Morgenstern (1947), Savage (1954), Luce and Raiffa (1957), Shafer (1976), and Friedman (1990) are useful texts for readers desiring an introduction to normative decision theory. Raiffa (1968), Keeney and Raiffa (1976), Saaty (1990), Buck (1989), and Clemen (1996) are applied texts on decision analysis. Kahneman et al. (1982), von Winterfeldt and Edwards (1986), Payne et al. (1993), Svenson and Maule (1993), Heath et al. (1994), Yates (1992), Koehler and Harvey (2004), and Camerer et al. (2004), among numerous others, are texts addressing elements of behavioral decision theory. Klein et al. (1993) and Klein (1998, 2004) provide introductions to naturalistic decision making.

focus of operations research and management science, psychology, sociology, and cognitive engineering. These fields have produced numerous models and a substantial body of research on human decision making. At least three objectives have motivated this work: to develop normative prescriptions that can guide decision makers, to describe how people make decisions and compare the results to normative prescriptions, and to determine how to help people apply their "natural" decision-making methods more successfully. The goals of this chapter are to synthesize the various elements of this work into a single picture and provide some depth of coverage in particularly important areas. The integrative model presented in Section 1.3 focuses on the first goal. The remaining sections address the second goal.

## 1.1 Role and Utility of the Chapter

This chapter is intended to provide an overall perspective on human decision making to human factors practitioners, developers of decision tools (e.g., expert systems), product designers, and others who are interested in how people make decisions and how decision making might be improved. Consequently, we present a broad set of prescriptive and descriptive approaches. Numerous applications are presented and strengths and weaknesses of particular approaches are noted. Emphasis is also placed on providing useful references containing additional information on topics that the reader may find to be of special interest.

Section 2 addresses various decision-making models, which are grouped into normative decision models (Section 2.1), behavioral decision models (Section 2.2), and naturalistic decision models (Section 2.3). Normative decision models are based on principles of rational choice and they prescribe how decisions should be made. In contrast, behavioral and natural decision models focus more on describing human decision making. Several descriptive models of human judgment, preference, and choice are discussed and compared to normative models in Section 2.2. Naturalistic decision models should be of interest to practitioners interested in the process to which many real-world decisions are made, the quality of these decisions, and why people use particular methods to make decisions. The discussion provides insight into how people perform diagnostic tasks, make decisions involving risks, and develop expertise.

Section 3 introduces the topic of group decision making. The discussion addresses conflict resolution both within and between groups, group performance and biases, and methods of group decision making.

The next section addresses the topic of decision support and problem solving. The discussion begins by addressing the topic of decision analysis (Section 4.1), which refers to the application of normative decision theory to improve decisions. The discussion considers the advantages of the various approaches, how they can be applied, and what problems might arise during their application. Sections 4.2 and 4.3 cover methods of assisting or supporting the decision making of individuals, groups, and organizations. Section 4.4 discusses the implication of problem-solving research for decision making.

## 1.2 Elements of Decision Making

Decision making requires that the decision maker make a choice between two or more alternatives (note that doing nothing can be viewed as making a choice). The alternative selected results in real or imaginary consequences to the decision maker. Judgment is a closely related process in which a person rates or assigns values to attributes of the alternatives considered. For example, a person might judge both the safety and attractiveness of a car being considered for purchase. Obtaining an attractive car is a desirable consequence of the decision, while obtaining an unsafe car is an undesirable consequence. A rational decision maker seeks desirable consequences and attempts to avoid undesirable consequences.

The nature of decision making can vary greatly, depending on the decision context. Certain decisions, such as deciding where and what to eat for lunch, are routine and repeated often. Other choices, such as purchasing a house, choosing a spouse, or selecting a form of medical treatment for a serious disease, occur rarely, may involve much deliberation, and take place over a longer period. Decisions may also be required under severe time pressure and involve potentially catastrophic consequences, such as when a fire chief decides whether to send firefighters into a burning building. Previous choices may constrain or otherwise influence subsequent choices (e.g., a decision to enter the graduate school might constrain a future employment-related decision to particular job types and locations). The outcomes of choices may be uncertain and in certain instances are determined by the actions of potentially adverse parties, such as competing manufacturers of a similar product. Decisions may be made by a single person or by a group. Within a group, there may be conflicting opinions and differing degrees of power between individuals or factions. Decision makers may also vary greatly in their knowledge and degree of aversion to risk.

Conflict occurs when a single decision maker is not sure which choice should be selected or when there is lack of consensus within a group regarding the choice. Both for groups and single decision makers, conflict occurs, at the most fundamental level, because of uncertainty or conflicting objectives. Uncertainty can take many forms and is one of the primary reasons that decisions can be difficult. In ill-structured decisions, decision makers may not have identified the current condition, alternatives to choose between, or their consequences. Decision makers also may be unsure what their aspirations or objectives are or how to choose between alternatives. At least four reasons for conflict may exist after a decision has been structured. First, when alternatives have both undesirable and desirable consequences, decision makers may experience conflict due to conflicting objectives. For example, a decision maker considering the purchase of an air bag–equipped car may experience conflict because an air bag increases the cost but improves safety. Second, decision makers may be unsure of their reaction to a consequence. For example, people considering whether to enter a raffle where the prize is a sailboat may be unsure how much they want a sailboat. Third, decision makers may not know whether

a consequence will be sure to happen. Even worse, they may be unsure what the probability of the consequences is or may not have enough time to evaluate the situation carefully. They may also be uncertain about the reliability of their information. For example, it may be difficult to determine the truth of a salesperson's claim regarding the probability that a product will break down immediately after the warranty expires.

To resolve conflicts, decision makers must deal appropriately with uncertainty, conflicting objectives, or a lack of consensus. Conflict resolution therefore becomes a primary focus of decision theory. In the following section we present an integrative model of decision making that relates conflict resolution to the elements of decision making discussed above. This model considers specifically how decision making changes when different sources of conflict are present. It also matches methods of conflict resolution to particular sources of conflict and decision rules.

### 1.3 Integrative Model of Decision Making

Human decision making can be viewed as a stage of information processing that falls between perception and response execution (Welford, 1976). The integrative model of human decision making, presented in Figure 1, shows how the elements of decision making discussed above fit into this perspective. From this view, decision making is the process followed when a response to a perceived stimulus is chosen. The process followed depends on what decision strategy is applied and can vary greatly between decision contexts.[*] Decision strategies in Figure 1 correspond to different paths between situation assessment and executing an action. The particular decision strategy followed depends on both the decision context and on whether or not the decision maker experiences conflict.[†]

At least four, sometimes overlapping, categories of decision making can be distinguished. *Group decision making* occurs when multiple decision makers interact and is represented at the highest level of the model as a source of conflict that might be resolved through debate, bargaining, or voting. For example, members of a university faculty committee might debate and bargain before voting for the best candidate for a job opening.

*Dynamic decision making* occurs in a changing environment, in which the results of earlier decisions affect future decisions. The decisions made in such settings often make use of feedback and are multistage in nature. For example, a decision to take a medical test almost always requires a subsequent decision regarding what to do after receiving the test results. Dynamic decision

making is represented at the lowest level of the model by the presence of two feedback loops, which show how the action taken and its effects can feed forward to the assessment of a new decision or feed back to the reassessment of the current decision.

*Routine decision making* occurs when decision makers use knowledge and past experience to decide quickly what to do and is especially prevalent in dynamic decision-making contexts. Routine decision making is represented in Figure 1 as a single pattern-matching step or associative leap between situation assessment and executing an action. For example, a driver, after perceiving a stop sign, decides to stop. Similarly, the user of a word-processing system, after perceiving a misspelled word, decides to activate the spell checker. Since routine decisions are often made in dynamic task environments, routine decision making is discussed in this chapter as a subtopic of dynamic decision making.

*Conflict-driven decision making* occurs when various forms of conflict must be resolved before an alternative action can be chosen and often involves a complicated path between situation assessment and executing an action.[‡] Before executing an action, the decision maker experiences conflict, somehow resolves it, and then either recognizes the best action (conflict resolution might transform the decision to a routine one) or applies a decision rule. Applying the decision rule leads ideally to a choice which is then executed. Attempting to apply the decision rule may, however, cause additional conflicts, leading to more conflict resolution. For example, decision makers may realize that they need more information to apply a particular decision rule. In response, they might decide to use a different decision rule that requires less information. Along these lines, when choosing a home, a decision maker might decide to use a satisficing decision rule after seeing that hundreds of homes are listed in the classified ads of the local newspaper.

Potential sources of conflict, methods of conflict resolution, and the results of conflict resolution are listed at the top of Figure 1. Each source of conflict maps to a particular method of conflict resolution, which then provides a result necessary to apply a decision rule, as illustrated schematically in the figure.[§]

Accordingly, conflict occurs at the most fundamental level when the current condition, alternative actions, or their consequences have not been identified. At the next most fundamental level, conflict occurs when the decision maker is unsure how to compare the alternatives. In other words, the decision maker has not yet selected a decision rule. Given that the decision maker

---

[*] The notion that the best decision strategy varies between decision contexts is a fundamental assumption of the theory of contingent decision making (Payne et al., 1993), cognitive continuum theory (Hammond, 1980), and other approaches discussed later in the chapter.

[†] Conflict has been recognized as an important determinant of what people do in risky decision-making contexts (Janis and Mann, 1977). Janis and Mann focus on the stressful nature of conflict and on how affective reactions in stressful situations can affect decision strategies.

---

[‡] The distinction between routine and conflict-driven decision making made here is similar to Rasmussen's (1983) distinction between (1) routine skill or rule-based levels of control and (2) nonroutine knowledge-based levels of control in information-processing tasks.

[§] Note that multiple sources of conflict are possible for a given decision context. An attempt to resolve one source of conflict may also make the decision maker aware of other conflicts that must first be resolved. For example, decision makers may realize they need to know what the alternatives are before they can determine their aspiration levels.
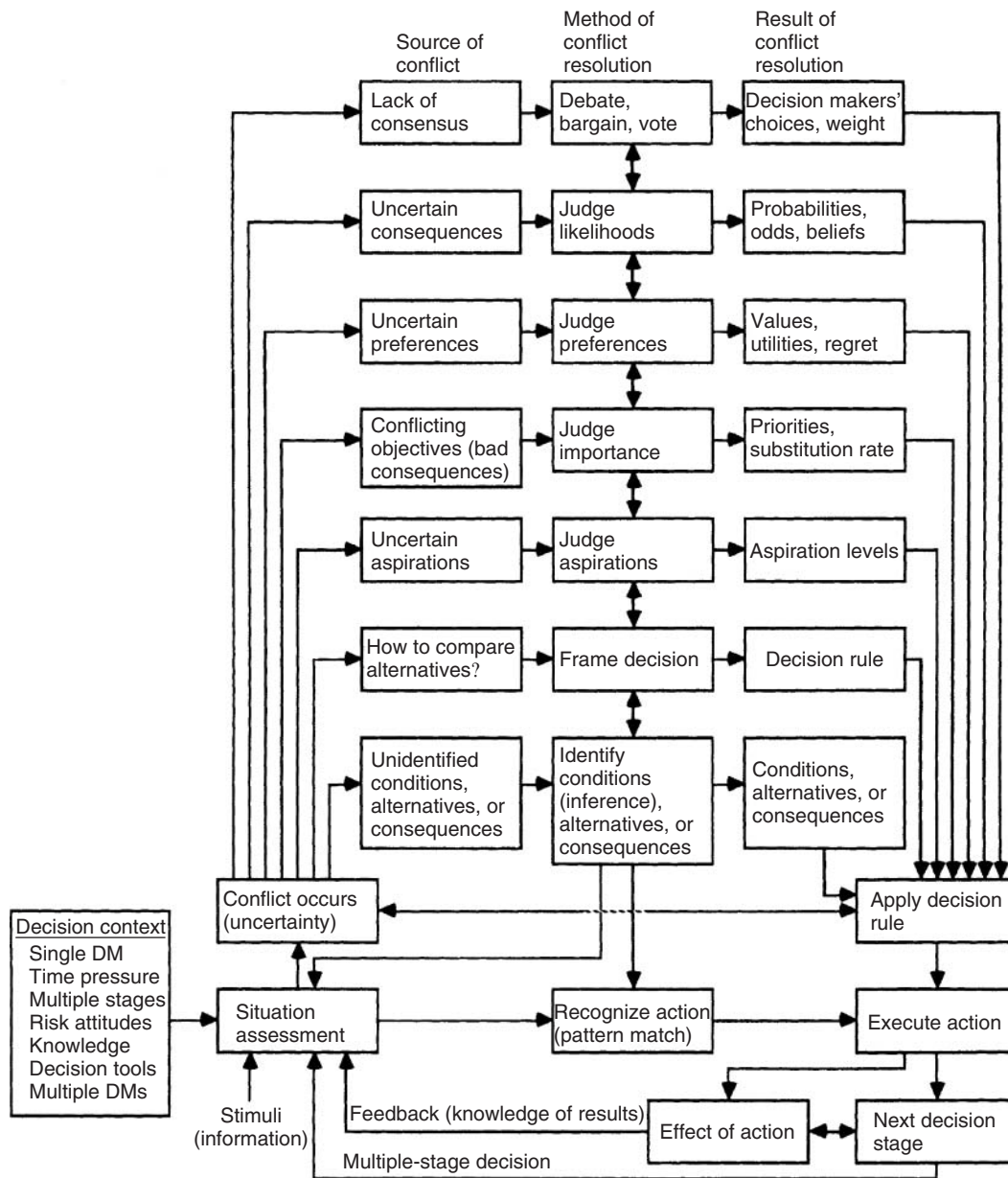
**Figure 1** Integrative model of human decision making (DM, decision maker).

has a decision rule, conflict can still occur if the needed inputs are not available. These sources of conflict and associated methods of conflict resolution are addressed briefly below in relation to the remainder of this chapter.

Identifying the current condition, alternative actions, and their consequences is an important part of decision making. This topic is emphasized in both naturalistic decision theory (Klein et al., 1993) and decision analysis[*] (Raiffa, 1968; Clemen, 1996). Decision trees, influence diagrams, and other tools for structuring decisions are covered in Section 4.1. In Section 2.2.1,

---

[*] Clemen (1996) includes a chapter on creativity and decision structuring. Some practitioners claim that structuring the decision is the greatest contribution of the decision analysis process.

we describe several descriptive models of human inference and discuss their limitations. Section 4.1 covers discussion group decision-making methods that may be useful at this decision-making stage.

When decision makers are unsure how to compare alternatives, they must consider what information is available and then frame a decision appropriately. The way the decision is framed then determines (1) which decision rules are appropriate, (2) what information is needed to make the decision using the rules given, and (3) the choices selected. As discussed in Section 2.2.3, there are reasons to believe that people apply different decision-making strategies in different decision contexts. We discuss the appropriateness of decision rules and how the particular rule used can affect choices. When the specific inputs needed by a decision rule are not available, the resulting conflict might be resolved by judging aspirations, importance, preference, or likelihood. It might also be resolved by choosing a different decision rule or strategy. As noted in Section 2.3, there is a prevalent tendency among decision makers in naturalistic settings to minimize analysis and the cognitive effort required. In group situations, conflict due to a lack of consensus among decision makers might be resolved through debate, bargaining, or voting (Section 3).

## 2  DECISION-MAKING MODELS

In this section, decision-making models are categorized into three types: normative (Section 2.1), behavioral (Section 2.2), and naturalistic (Section 2.3). Normative decision models date back to early application of economics and statistics to specify how to make optimal decisions (von Neumann and Morgenstern, 1947; Savage, 1954); thus, they focus heavily on the notion of rationality (Savage, 1954; von Winterfeldt and Edwards 1986). In contrast, behavioral decision models acknowledge the limitations of human decision makers: they are systematically biased and use heuristics to overcome cognitive limitations. In these models, decision makers are not rational but boundedly rational (Simon, 1955). Naturalistic decision models extend this perspective to understand how people actually make decisions and cognitively complex tasks in realistic demanding situations, which cannot be easily replicated in a laboratory setting.

## 2.1  Normative Models

Classical decision theory represents preference and choice problems in terms of four basic elements: (1) a set of potential actions ($A_i$) to choose between, (2) a set of events or world states ($E_j$), (3) a set of consequences ($C_{ij}$) obtained for each combination of action and event, and (4) a set of probabilities ($P_{ij}$) for each combination of action and event. For example, a decision maker might be deciding whether to wear a seat belt when traveling in an automobile. Wearing or not wearing a seat belt corresponds to two actions, $A_1$ and $A_2$. The expected consequence ($C_{ij}$) of either action depends on whether an accident occurs. Having or not having an accident corresponds to two events, $E_1$ and $E_2$. Wearing a seat belt reduces the expected consequences ($C_{11}$) of having an accident ($E_1$). As the probability of having an accident increases, use of a belt should therefore become more attractive.

Normative models are based on basic axioms (or what are felt to be self-evident assumptions) of rational choice. In the following discussion we first present some of the most basic axioms.

### 2.1.1  Axioms of Rational Choice

Numerous axioms have been proposed that are essential either for a particular model of choice or for the method of eliciting numbers used for a particular model (von Winterfeldt and Edwards, 1986). The best known set of axioms (Table 1) establishes the normative principle of *subjective expected utility* (SEU) as a basis for making decisions [see Savage (1954) and Luce and Raiffa (1957) for a more rigorous description of the axioms]. On an individual basis, these axioms are intuitively appealing (Stukey and Zeckhauser, 1978), but as discussed in Section 2.2.2, people's preferences can deviate significantly from the SEU model in ways that conflict with certain axioms. Consequently, there has been a movement toward developing less restrictive standards of normative decision making (Zey, 1992; Frisch and Clemen, 1994).

Frisch and Clemen (1994, p. 49) propose that "a good decision should (a) be based on the relevant consequences of the different options (*consequentialism*), (b) be based on an accurate assessment of the world

## Table 1  Basic Axioms of Subjective Expected Utility Theory

A. *Ordering/quantification of preference*. Preferences of decision makers between alternatives can be quantified and ordered using the relations:

> $>$, where $A > B$ means that $A$ is preferred to $B$

> $=$, where $A = B$ means that $A$ and $B$ are equivalent

> $\geq$, where $A \geq B$ means that $B$ is not preferred to $A$

B. *Transitivity of preference*. If $A_1 \geq A_2$ and $A_2 \geq A_3$, then $A_1 \geq A_3$.

C. *Quantification of judgment*. The relative likelihood of each possible consequence that might result from an alternative action can be specified.

D. *Comparison of alternatives*. If two alternatives yield the same consequences, the alternative yielding the greater chance of the preferred consequence is preferred.

E. *Substitution*. If $A_1 > A_2 > A_3$, the decision maker will be willing to accept a gamble [$p(A_1)$ and $(1 - p)(A_3)$] as a substitute for $A_2$ for some value of $p \geq 0$.

F. *Sure-thing principle*. If $A_1 \geq A_2$, then for all $p$, the gamble [$p(A_1)$ and $(1 - p)(A_3)$] $\geq$ [$p(A_2)$ and $(1 - p)(A_3)$].

and a consideration of all relevant consequences (*thorough structuring*), and (c) make tradeoffs of some form (*compensatory decision rule*)." Consequentialism and the need for thorough structuring are both assumed by all normative decision rules. Most normative rules are also compensatory. However, when people make routine habitual decisions, they often do not consider the consequences of their choices, as discussed in Section 2.3. In addition, because of cognitive limitations and the difficulty of obtaining information, it becomes unrealistic in many settings for the decision maker to consider all the options and possible consequences. To make decisions under such conditions, decision makers may limit the scope of the analysis by applying principles such as satisficing and other noncompensatory decision rules, which will be discussed in Section 2.2.3. They may also apply heuristics, based on their knowledge or experience, leading to performance that can approximate the results of applying compensatory decision rules (Section 2.2).

### 2.1.2 Dominance

Dominance is perhaps the most fundamental normative decision rule. *Dominance* is said to occur between two alternative actions, $A_i$ and $A_j$, when $A_i$ is at least as good as $A_j$ for all events $E$, and for at least one event $E_k$, $A_i$ is preferred to $A_j$. For example, one investment might yield a better return than another regardless of whether the stock market goes up or down. Dominance can also be described for the case where the consequences are multidimensional. This occurs when for all events $E$, the $k$th consequence associated with action $i$ ($C_{ik}$) and action $j$ ($C_{jk}$) satisfies the relation $C_{ik} \geq C_{jk}$ for all $k$ and for at least one consequence $C_{ik} > C_{jk}$. For example, a physician choosing between alternative treatments has an easy decision if one treatment is *both* cheaper and more effective for all patients.

Dominance is obviously a normative decision rule, since a dominated alternative can never be better than the alternative that dominates it. Dominance is also conceptually simple, but it can be difficult to detect when there are many alternatives to consider or many possible consequences. The use of tests for dominance by decision makers in naturalistic settings is discussed further in Section 2.3.

### 2.1.3 Maximizing Expected Value

From elementary probability theory, return is maximized by selecting the alternative with the greatest expected value. The expected value of an action $A_i$ is calculated by weighting the decision maker's preference $V(C_{ik})$ for its consequences $C_{ik}$ over all events $k$ by the probability $P_{ik}$ that the event will occur. The expected value of a given action $A_i$ is therefore

$$\text{EV}[A_i] = \sum_k P_{ik} V(C_{ik}) \qquad (1)$$

Monetary value is a common value function. For example, lives lost, units sold, or air quality might all be converted into monetary values. More generally,

however, value reflects preference, as illustrated by ordinary concepts such as the value of money or the attractiveness of a work setting. Given that the decision maker has large resources and is given repeated opportunities to make the choice, choices made on the basis of expected monetary value are intuitively justifiable. A large company might make nearly all of its decisions on the basis of expected monetary value. Insurance buying and many other rational forms of behavior cannot, however, be justified on the basis of expected monetary value. It has long been recognized that rational decision makers made choices not easily explained by expected monetary value (Bernoulli, 1738). Bernoulli cited the St. Petersburg paradox, in which the prize received in a lottery was $2^n$, $n$ being the number of times a flipped coin turned up heads before a tail was observed. The probability of $n$ flips before the first tail is observed is $0.5^n$. The expected value of this lottery becomes

$$\text{EV}[L] = \sum_k P_{ik} V(C_{ik}) = \sum_{n=0}^{\infty} 0.5^n 2^n \rightarrow \infty \qquad (2)$$

The interesting twist is that the expected value of the lottery above is infinite. Bernoulli's conclusion was that preference cannot be a linear function of monetary value since a rational decision maker would never pay more than a finite amount to play the lottery. Furthermore, the value of the lottery can vary between decision makers. According to utility theory, this variability, described in utility, reflects rational differences in preference between decision makers for uncertain consequences.

### 2.1.4 Subjective Expected Utility Theory

Expected utility theory extended expected value theory to describe better how people make uncertain economic choices (von Neumann and Morgenstern, 1947). In their approach, monetary values are first transformed into utilities using a utility function $u(x)$. The utilities of each outcome are then weighted by their probability of occurrence to obtain an expected utility. The SEU theory added the notion that uncertainty about outcomes could be represented with subjective probabilities (Savage, 1954). It was postulated that these subjective estimates could be combined with evidence using Bayes' rule to infer the probabilities of outcomes.[*] This group of assumptions corresponds to the Bayesian approach to statistics. Following this approach, the SEU of an alternative $A_i$, given subjective probabilities $S_{ik}$ and consequences $C_{ik}$ over events $E_k$, becomes

$$\text{SEU}[A_i] = \sum_k S_{ik} U(C_{ik}) \qquad (3)$$

Note the similarity between formulation (3) for SEU and equation (1) for expected value. The EV and SEU

---

[*] When no evidence is available concerning the likelihood of different events, it was postulated that each consequence should be assumed to be equally likely. The Laplace decision rule makes this assumption and then compares alternatives on the basis of expected value or utility.

are equivalent if the value function equals the utility function. Methods for eliciting value and utility functions differ in nature (Section 4.1). Preferences elicited for uncertain outcomes measure utility.[*] Preferences elicited for certain outcomes measure value. It has, accordingly, often been assumed that value functions differ from utility functions, but there are reasons to treat value and utility functions as equivalent (von Winterfeldt and Edwards, 1986). The latter authors claim that the differences between elicited value and utility functions are small and that "severe limitations constrain those relationships, and only a few possibilities exist, one of which is that they are the same."

When people are presented with choices that have uncertain outcomes, they react in different ways. In some situations, people find gambling to be pleasurable. In others, people will pay money to reduce uncertainty: for example, when people buy insurance. SEU theory distinguishes between risk-neutral, risk-averse, risk-seeking, and mixed forms of behavior. These different types of behavior are described by the shape of the utility function (Figure 2).

A risk-neutral decision maker will find the expected utility of a gamble to be the same as the utility of the gamble's expected value. That is, expected $u(\text{gamble}) = u(\text{gamble's expected value})$. For a risk-averse decision maker, expected $u(\text{gamble}) < u(\text{gamble's expected value})$; for a risk-seeking decision maker, expected $u(\text{gamble}) > u(\text{gamble's expected value})$. On any given point of a utility function, attitudes toward risk are described formally by the coefficient of risk aversion:

$$C_{\text{RA}} = \frac{u''(x)}{u'(x)} \qquad (4)$$

where $u'(x)$ and $u''(x)$ are, respectively, the first and second derivatives of $u(x)$ taken with respect to $x$. Note that when $u(x)$ is a linear function of $x$ [i.e., $u(x) = ax + b$], then $C_{\text{RA}} = 0$. For any point of the utility function, if $C_{\text{RA}} < 0$, the utility function depicts risk-averse behavior, and if $C_{\text{RA}} > 0$, the utility function depicts risk-seeking behavior. The coefficient of risk aversion therefore describes attitudes toward risk at each point of the utility function given that the utility function is continuous. SEU theory consequently provides a powerful tool for describing how people might react to uncertain or risky outcomes. However, some commonly observed preferences between risky alternatives cannot be explained by SEU. Section 2.2.2 focuses on experimental findings showing deviations from the predictions of SEU.

A major contribution of SEU is that it represents differing attitudes toward risk and provides a normative model of decision making under uncertainty. The prescriptions of SEU are also clear and testable. Consequently, SEU has played a major role in fields other than economics, both as a tool for improving human decision making and as a stepping stone for developing models
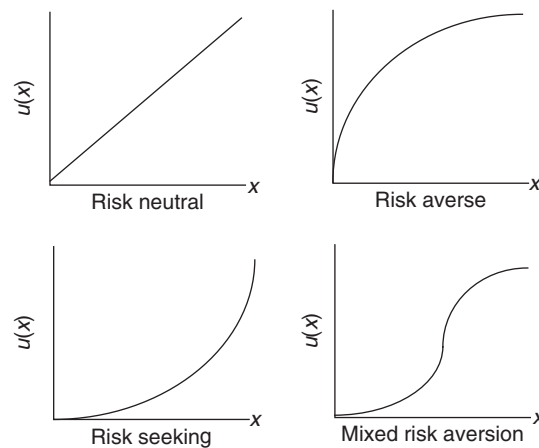


**Figure 2**  Utility functions for differing risk attitudes.

that describe how people make decisions when outcomes are uncertain. As discussed further in Section 2.2, much of this work has been done in psychology.

### 2.1.5  Multiattribute Utility Theory

Multiattribute utility theory (MAUT) (Keeney and Raiffa, 1976) extends SEU to the case where the decision maker has multiple objectives. The approach is equally applicable for describing utility and value functions. Following this approach, the utility (or value) of an alternative $A$, with multiple attributes $x$, is described with the multiattribute utility (or value) function $u(x_1, \ldots, x_n)$, where $u(x_1, \ldots, x_n)$ is some function $f(x_1, \ldots, x_n)$ of the attributes $x$. In the simplest case, MAUT describes the utility of an alternative as an additive function of the single-attribute utility functions $u_n(x_n)$. That is,

$$u(x_1, \ldots, x_n) = \sum_{i=1}^{n} k_i u_i(x_i) \qquad (5)$$

where the constants $k_n$ are used to weight each single-attribute utility function ($u_n$) in terms of its importance.[†] Assuming that an alternative has three attributes, $x$, $y$, and $z$, an additive utility function is $u(x, y, z) = k_x u_x(x) + k_y u_y(y) + k_z u_z(z)$. Along these lines, a community considering building a bridge across a river versus building a tunnel or continuing to use the existing ferry system might consider the attractiveness of each option in terms of the attributes of economic benefits, social benefits, and environmental benefits.

More complex multiattribute utility functions include multiplicative forms and functions that combine utility functions for subsets of two or more attributes (Keeney and Raiffa, 1976). An example of a simple multiplicative function would be $u(x, y) = u_x(x) u_y(y)$. A function

---

[*] Note that classical utility theory assumes that utilities are constant. Utilities may, of course, fluctuate. The random-utility model (Bock and Jones, 1968) allows such fluctuation.

[†] To develop the multiattribute utility function, the single-attribute utility functions ($u_n$) and the importance weights ($k_n$) are determined by assessing preferences between alternatives. Methods of doing so are discussed in Section 4.1.3.

that combines utility functions for subsets would be $u(x, y, z) = k_{xy} u_{xy}(x, y) + k_z u_z(z)$. The latter type of function becomes useful when utility independence is violated.

Utility independence is violated when the utility function for one attribute depends on the value of another attribute. Along these lines, when assessing $u_{xy}(x, y)$, it might be found that $u_x(x)$ depends on the value of $y$. For example, people's reaction to the level of crime in their own neighborhood might depend on the level of crime in a nearby suburb. In the latter case, it is probably better to measure $u_{xy}$ ($x$ is crime in one's own neighborhood and $y$ is crime in a nearby suburb) directly than to estimate it from the single-attribute functions. Assessment of utility and value functions is discussed in Section 4.1.

MAUT has been applied to a wide variety of problems (Clemen, 1996; Keeney and Raiffa, 1976; Saaty, 1990; Wallenius et al., 2008; von Winterfeldt and Edwards, 1986). An advantage of MAUT is that it helps structure complex decisions in a meaningful way. Alternative choices and their attributes often naturally divide into hierarchies. The MAUT approach encourages such divide-and-conquer strategies and, especially in its additive form, provides a straightforward means of recombining weights into a final ranking of alternatives. The MAUT approach is also a compensatory strategy that allows normative trade-offs between attributes in terms of their importance.

## 2.2 Behavioral Decision Models

As a normative ideal, classical decision theory has influenced the study of decision making in a major way. Much of the earlier work in behavioral decision theory compared human behavior to the prescriptions of classical decision theory (Edwards, 1954; Slovic et al., 1977; Einhorn and Hogarth, 1981). Numerous departures were found, including the influential finding that people use heuristics during judgment tasks (Tversky and Kahneman, 1974). On the basis of such research, psychologists have concluded that other approaches are needed to describe the process of human decision making. Descriptive models that relax assumptions of the normative models but retain much of their essence are now being evaluated in the field of judgment and decision theory (Stevenson et al., 1993). One of most exciting developments is that fast and frugal heuristics can perform very well even when compared to sophisticated optimization model (Gigerenzer, 2008).

The following discussion summarizes findings from this broad body of literature. The discussion begins by considering research on statistical estimation and inference. Attention then shifts to the topic of decision making under uncertainty and risk.

### 2.2.1 Statistical Estimation and Inference

The ability of people to perceive, learn, and draw inferences accurately from uncertain sources of information has been a topic of much research. In the following discussion we first consider briefly human abilities and limitations on such tasks. Attention then shifts to several heuristics that people may use to cope with their limitations and how their use can cause certain biases. In the next section, we then consider briefly the role of memory and selective processing of information from a similar perspective. Attention then shifts to mathematical models of human judgment that provide insight into how people judge probabilities, the biases that might occur, and how people learn to perform probability judgment tasks. In the final section, we summarize findings on debiasing human judgments.

***Human Abilities and Limitations*** Research conducted in the early 1960s tested the notion that people behave as "intuitive statisticians" who gather evidence and apply it in accordance with the Bayesian model of inference (Peterson and Beach, 1967). Much of the earlier work focused on how good people are at estimating statistical parameters such as means, variances, and proportions. Other studies have compared human inferences obtained from probabilistic evidence to the prescriptions of Bayes' rule.

A number of interesting results were obtained (Table 2). The research first shows that people can be fairly good at estimating means, variances, or proportions from sample data. Sedlmeier et al. (1998) point out that "there seems to be broad agreement with" (p. 754) the conclusion of Jonides and Jones (1992) that people can give answers that reflect the actual relative frequencies of many kinds of events with great fidelity. However, as discussed by von Winterfeldt and Edwards (1986), like other psychophysical measures, subjective probability estimates are noisy. Their accuracy will depend on how carefully they are elicited and on many other factors. Studies have shown that people are especially likely to have trouble estimating accurately the probability of unlikely events, such as nuclear plant explosions. For example, when people were asked to estimate the risk associated with the use of consumer products (Dorris and Tabrizi, 1978; Rethans, 1980) or various technologies (Lichtenstein et al., 1978), the estimates obtained were often weakly related to accident data. Weather forecasters are one of the few groups of people that have been documented as being able to estimate high and low probabilities accurately (Winkler and Murphy, 1973).

Part of the issue is that when events occur rarely, people will not be able to base their judgments on a representative sample of their own observations. Most of the information they receive about unlikely events will come from secondary sources, such as media reports, rather than from their own experience. This tendency might explain why risk estimates are often related more strongly to factors other than likelihood, such as catastrophic potential or familiarity (Lichtenstein et al., 1978; Slovic 1978, 1987; Lehto et al., 1994). Media reporting focuses on "newsworthy" events, which tend to be more catastrophic and unfamiliar. Consequently, judgments based on media reports might reflect the latter factors instead of likelihood. Weber (1994) provides additional evidence that subjective probabilities are related to factors other than likelihood and argues that people will overestimate the chance of a highly

**Table 2 Sample Findings on the Ability of People to Estimate and Infer Statistical Quantities**

| | Reference |
|---|---|
| **Statistical estimation** | |
| Accurate estimation of sample means | Peterson and Beach (1967) |
| Variance estimates correlated with mean | Lathrop (1967) |
| Variance biases not found | Levin (1975) |
| Variance estimates based on range | Pitz (1980) |
| Accurate estimation of event frequency | Estes (1976), Hasher and Zacks (1984), Jonides and Jones (1992) |
| Accurate estimates of sample proportions between 0.75 and 0.25 | Edwards (1954) |
| Severe overestimates of high probabilities; severe underestimates of low proportions | Fischhoff et al. (1977), Lichtenstein et al. (1982) |
| Reluctance to report extreme events | Du Charme (1970) |
| Weather forecasters provided accurate probabilities | Winkler and Murphy (1973) |
| Poor estimates of expected severity | Dorris and Tabrizi (1978) |
| Correlation of 0.72 between subjective and objective measures of injury frequency | Rethans (1980) |
| Risk estimates lower for self than for others | Weinstein (1980, 1987) |
| Risk estimates related to catastrophic potential, degree of control, familiarity | Lichtenstein et al. (1978) |
| Evaluations of outcomes and probabilities are dependent | Weber (1994) |
| Overweighting the probability of rare events in decisions from descriptions; underweighting decisions from experiences | Hertwig et al. (2004) |
| **Statistical inference** | |
| Conservative aggregation of evidence | Edwards (1968) |
| Nearly optimal aggregation of evidence in naturalistic setting | Lehto et al. (2000) |
| Failure to consider base rates | Tversky and Kahneman (1974) |
| Base rates considered | Birnbaum and Mellers (1983) |
| Overestimation of conjunctive events | Bar-Hillel (1973) |
| Underestimation of disjunctive events | |
| Tendency to seek confirming evidence, tendency to discount disconfirming evidence, tendency to ignore reliability of the evidence | Einhorn and Hogarth (1978), Baron (1985) |
| Subjects considered variability of data when judging probabilities | Kahneman and Tversky (1973) |
| People insensitive to information missing from fault trees | Evans and Pollard (1985) |
| Overconfidence in estimates | Fischhoff et al. (1978) |
| Hindsight bias | Fischhoff et al. (1977) |
| Illusionary correlations | Fischhoff (1982), Christensen-Szalanski and Willham (1991) |
| Gambler's fallacy | Tversky and Kahneman (1974) |
| Misestimation of covariance between items | Arkes (1981) |
| Misinterpretation of regression to the mean | Tversky and Kahneman (1974) |
| Optimism bias | Armor and Taylor (2002) |

positive outcome because of their desire to obtain it. Weber also argues that people will overestimate the chance of a highly undesirable outcome because of their fear of receiving it. Traditional methods of decision analysis separately elicit and then recombine subjective probabilities with utilities, as discussed earlier, and assume that subjective probabilities are independent of consequences. A finding of dependency therefore casts serious doubt on the normative validity of this commonly accepted approach.

When studies of human inference are considered, several other trends become apparent (Table 2). In particular, several significant deviations from the Bayesian model have been found:

1. Decision makers tend to be conservative in that they do not give as much weight to probabilistic evidence as does Bayes' rule (Edwards, 1968).

2. Decision makers do not consider base rates or prior probabilities adequately (Tversky and Kahneman, 1974).

3. Decision makers tend to ignore the reliability of the evidence (Tversky and Kahneman, 1974).

4. Decision makers tend to overestimate the probability of conjunctive events and underestimate the probability of disjunctive events (Bar-Hillel, 1973).

5. Decision makers tend to seek out confirming evidence rather than disconfirming evidence and place more emphasis on confirming evidence when it is available (Einhorn and Hogarth, 1978; Baron, 1985). The order in which the evidence is presented has an influence on human judgments (Hogarth and Einhorn, 1992).

6. Decision makers are overconfident in their predictions (Fischhoff et al. 1977), especially in hindsight (Fischhoff, 1982; Christensen-Szalanski and Willham, 1991).

7. Decision makers show a tendency to infer illusionary causal relations (Tversky and Kahneman, 1973).

A lively literature has developed regarding these deviations and their significance* (Evans, 1989; Caverni et al., 1990; Wickens, 1992; Klein et al., 1993; Doherty, 2003). From one perspective, these deviations demonstrate inadequacies of human reason and are a source of societal problems (Baron, 1998; and many others). From the opposite perspective, it has been held that the foregoing findings are more or less experimental artifacts that do not reflect the true complexity of the world (Cohen, 1993). A compelling argument for the latter point of view is given by Simon (1955, 1983). From this perspective, people do not use Bayes' rule to compute probabilities in their natural environments because it makes unrealistic assumptions about what is known or knowable. Simply put, the limitations of the human mind and time constraints make it nearly impossible for people to use principles such as Bayes' rule to make inferences in their natural environments. To compensate for their limitations, people use simple heuristics or decision rules that are adapted to particular environments. The use of such strategies does not mean that people will not be able to make accurate inferences, as emphasized by both Simon and researchers embracing the ecologica† (i.e., Hammond, 1996; Gigerenzer et al., 1999) and naturalistic (i.e., Klein et al., 1993) models of decision making. In fact, as discussed further in this section, the use of simple heuristics in rich environments can lead to inferences that are in many cases

more accurate than those made using Naïve Bayes, or linear regression (Gigerenzer et al., 1999).

There is an emerging body of literature that shows, on the one hand, that deviations from Bayes' rule can in fact be justified in certain cases from a normative view and, on the other hand, that these deviations may disappear when people are provided with richer information or problems in more natural contexts. For example, drivers performing a simulated passing task combined their own observations of the driving environment with imperfect information provided by a collision-warning system, as predicted by a distributed signal detection theoretic model of optimal team decision making (Lehto et al., 2000). Other researchers have pointed out that:

1. A tendency toward conservatism can be justified when evidence is not conditionally independent (Navon, 1979).

2. Subjects do use base-rate information and consider the reliability of evidence in slightly modified experimental settings (Birnbaum and Mellers, 1983; Koehler, 1996). In particular, providing natural frequencies instead of probabilities to subjects can improve performance greatly (Gigerenzer and Hoffrage, 1995; Krauss et al., 1999).

3. A tendency to seek out confirming evidence can offer practical advantages (Cohen, 1993) and may reflect cognitive failures, due to a lack of understanding of how to falsify hypotheses, rather than an entirely motivational basis (Klayman and Ha, 1987; Evans, 1989).

4. Subjects prefer stating subjective probabilities with vague verbal expressions rather than precise numerical values (Wallsten et al., 1993), demonstrating that they are not necessarily overconfident in their predictions.

5. There is evidence that the hindsight bias can be moderated by familiarity with both the task and the type of outcome information provided (Christensen-Szalanski and Willham, 1991).

Based on such results, numerous researchers have questioned the practical relevance of the large literature showing different types of biases. One reason that this literature may be misleading is that researchers overreport findings of bias (Evans, 1989; Cohen, 1993). A more significant concern is that studies showing bias are almost always conducted in artificial settings where people are provided information about an unfamiliar topic. Furthermore, the information is often given in a form that forces use of Bayes' rule or other form of abstract reasoning to get the correct answer. For example, consider the simple case where a person is asked to predict how likely it is that a woman has breast cancer given a positive mammogram (Martignon and Krauss, 2003). In the typical study looking for bias, the subject might be told to assume (1) the probability that a 40-year-old woman has breast cancer is 1%, (2) the probability of a positive mammogram given that a woman has cancer is 0.9, and (3) the probability of a positive mammogram given that a woman does not

---

* Doherty (2003) groups researchers on human judgment and decision making into two camps. The optimists focus on the success of imperfect human beings in a complex world. The pessimists focus on the deficiencies of human reasoning compared to normative models.

† As noted by Gigerenzer et al. (1999, p. 18), because of environmental challenges, "organisms must be able to make inferences that are fast, frugal, and accurate." Similarly, Hammond (1996) notes that a close correspondence between subjective beliefs and environmental states will provide an adaptive advantage.

have cancer is 0.1. Although the correct answer can be easily calculated using Bayes' rule, it is not at all surprising that people unfamiliar with probability theory have difficulty determining it. In the real world, it seems much more likely that a person would simply keep track of how many women receiving a mammogram actually had breast cancer. The probability that a woman has breast cancer, given a positive mammogram, is then determined by dividing the number of women receiving a mammogram who actually had breast cancer by the number of women receiving a mammogram. The latter calculation gives exactly the same answer as using Bayes' rule and is much easier to do.

The implications of the example above are obvious: First, people can duplicate the predictions of the Bayes rule by keeping track of the right relative frequencies. Second, if the right relative frequencies are known, accurate inferences can be made using very simple decision rules. Third, people will have trouble making accurate inferences if they do not know the right relative frequencies. Recent studies and reevaluations of older studies provide additional perspective. The finding that subjects are much better at integrating information when they are provided data in the form of natural frequencies instead of probabilities (Gigerenzer and Hoffrage, 1995; Krauss et al., 1999) is particularly interesting.

One conclusion that might be drawn from the latter work is that people are Bayesians after all if they are provided adequate information in appropriate representations (Martignon and Krauss, 2003). Other support for the proposition that people are not as bad at inference as it once seemed includes Dawes and Mulford's (1996) review of the literature supporting the overconfidence effect or bias, in which they conclude that the methods used to measure this effect are logically flawed and that the empirical support is inadequate to conclude that it really exists. Part of the issue is that much of the psychological research on the overconfidence effect "overrepresents those situations where cue-based inferences fail" (Juslin and Olsson, 1999). When people rate objects that are selected randomly from a natural environment, overconfidence is reduced. Koehler (1996) provides a similarly compelling reexamination of the base-rate fallacy. He concludes that the literature does not support the conventional wisdom that people routinely ignore base rates. To the contrary, he states that base rates are almost always used and that their degree of use depends on task structure and representation as well as their reliability compared to other sources of information.

Because such conflicting results can be obtained, depending on the setting in which human decision making is observed, researchers embracing the ecological (i.e., Hammond, 1996; Gigerenzer et al., 1999) and naturalistic (Klein et al., 1993; Klein, 1998) models of decision making strongly emphasize the need to conduct ecologically valid research in rich realistic decision environments.

***Heuristics and Biases*** Tversky and Kahneman (1973, 1974) made a key contribution to the field when they showed that many of the above-mentioned discrepancies between human estimates of probability and Bayes' rule could be explained by the use of three heuristics. The three heuristics they proposed were those of representativeness, availability, and anchoring and adjustment.

The *representativeness* heuristic holds that the probability of an item A belonging to some category B is judged by considering how representative A is of B. For example, a person is typically judged more likely to be a librarian than a farmer when described as "a meek and tidy soul who has a desire for order and structure and a passion for detail." Applications of this heuristic will often lead to good probability estimates but can lead to systematic biases. Tversky and Kahneman (1974) give several examples of such biases. In each case, representativeness influenced estimates more than other, more statistically oriented information. In the first study, subjects ignored base-rate information (given by the experimenter) about how likely a person was to be either a lawyer or an engineer. Their judgments seemed to be based entirely on how representative the description seemed to be of either occupation. Tversky and Kahneman (1973) found people overestimated conjunctive probabilities in a similar experiment. Here, after being told that "Linda is 31 years old, single, outspoken, and very bright," most subjects said it was more likely she was both a bank teller and active as a feminist than simply a bank teller. In a third study, most subjects felt that the probability of more than 60% male births on a given day was about the same for both large and small hospitals (Tversky and Kahneman, 1974). Apparently, the subjects felt that large and small hospitals were equally representative of the population.

Other behaviors explained in terms of representativeness by Tversky and Kahneman (1974) included gambler's fallacy, insensitivity to predictability, illusions of validity, and misconceptions of statistical regression to the mean. With regard to gambler's fallacy, they note that people may feel that long sequences of heads or tails when flipping coins are unrepresentative of normal behavior. After a sequence of heads, a tail therefore seems more representative. Insensitivity to predictability refers to a tendency for people to predict future performance without considering the reliability of the information on which they base the prediction. For example, a person might expect an investment to be profitable solely on the basis of a favorable description without considering whether the description has any predictive value. In other words, a good description is believed to be representative of high profits, even if it states nothing about profitability. The illusion of validity occurs when people use highly correlated evidence to make a conclusion. Despite the fact that the evidence is redundant, the presence of many representative pieces of evidence increases confidence greatly. Misconception of regression to the mean occurs when people react to unusual events and then infer a causal linkage when the process returns to normality on its own. For example, a manager might incorrectly conclude that punishment works after seeing that unusually poor performance improves to normal levels following punishment. The same manager might also conclude that rewards do not work

after seeing that unusually good performance drops after receiving a reward.

The *availability* heuristic holds that the probability of an event is determined by how easy it is to remember the event happening. Tversky and Kahneman state that perceived probabilities will therefore depend on familiarity, salience, effectiveness of memory search, and imaginability. The implication is that people will judge events as more likely when the events are familiar, highly salient (such as an airplane crash), or easily imaginable. Events will also be judged more likely if there is a simple way to search memory. For example, it is much easier to search for words in memory by the first letter rather than by the third letter. It is easy to see how each item above affecting the availability of information can influence judgments. Biases should increase when people lack experience or when their experiences are too focused.

The *anchoring-and-adjustment* heuristic holds that people start from an initial estimate and then adjust it to reach a final value. The point chosen initially has a major impact on the final value selected when adjustments are insufficient. Tversky and Kahneman (1974) refer to this source of bias as an anchoring effect. They show how this effect can explain under- and overestimates of disjunctive and conjunctive events. This happens if the subject starts with a probability estimate of a single event. The probability of a single event is, of course, less than that for the disjunctive event and greater than that for the conjunctive event. If adjustment is too small, under- and overestimates occur, respectively, for the disjunctive and conjunctive events. Tversky and Kahneman also discuss how anchoring and adjustment may cause biases in subjective probability distributions.

The notion of heuristics and biases has had a particularly formative influence on decision theory. A substantial recent body of work has emerged that focuses on applying research on heuristics and biases (Kahneman et al., 1982; Heath et al., 1994). Applications include medical judgment and decision making, affirmative action, education, personality assessment, legal decision making, mediation, and policy making. It seems clear that this approach is excellent for describing many general aspects of decision making in the real world. However, research on heuristics and biases has been criticized as being pretheoretical (Slovic et al., 1977) and, as pointed out earlier, has contributed to overselling of the view that people are biased. The latter point is interesting, as Tversky and Kahneman (1973) have claimed all along that using these heuristics can lead to good results. However, nearly all the research conducted in this framework has focused on when they might go wrong.

### Memory Effects and Selective Processing of Information

The heuristics-and-biases framework has been criticized by many researchers for its failure to adequately address more fundamental cognitive processes that might explain biases (Dougherty et al., 2003). This follows because the availability and representativeness heuristics can both be described in terms of more fundamental memory processes. For example, the availability heuristic proposes that the probability of

an event is determined by how easy it is to remember the event happening. Ease of recall, however, depends on many things, such as what is stored in memory, how it is represented, how well it is encoded, and how well a cue item matches the memory representation.

Dougherty et al. (2003) note that three aspects of memory can explain many of the findings on human judgment: (1) how information is stored or represented, (2) how information is retrieved, and (3) experience and domain knowledge. The first aspect pertains to what is actually stored when people experience events. The simplest models assume that people store a record of each instance of an experienced event and, in some cases, additional information such as the frequency of the event (Hasher and Zacks, 1984) or ecological cue validities (Brehmer and Joyce, 1988). More complex models assume that people store an abstract representation or summary of the event (Pennington and Hastie, 1988), in some cases at multiple levels of abstraction (Reyna and Brainerd, 1995). The way information is stored or represented can explain several of the observed findings on human judgment.

First, there is strong evidence that people are often excellent at storing frequency information[*] and the process by which this is done is fairly automatic (Hasher and Zacks, 1984; Gigerenzer et al., 1991). Gigerenzer et al. conclude that with repeated experience people should also be able to store ecological cue validities. The accuracy of these stored representations would, of course, depend on how large and representative the sample of encoded observations is. Such effects can be modeled with simple adding models that might include the effects of forgetting (or memory trace degradation) or other factors, such as selective sampling or the amount of attention devoted to the information at the time it is received. As pointed out by Dougherty et al. (2003) and many others, many of the biases in human judgment follow directly from considering how well the events are encoded in memory. In particular, except for certain sensory qualities which are encoded automatically, encoding quality is assumed to depend on attention. Consequently, some biases should reflect the tendency of highly salient stimuli to capture attention. Another completely different type of bias might reflect the fact that the person was exposed to an unrepresentative sample of events. Lumping these two very different biases together, as is done by the availability heuristic, is obviously debatable.

Other aspects of human memory mentioned by Dougherty et al. (2003) that can explain certain findings on human judgment include the level of abstraction of the stored representation and retrieval methods. One interesting observation is that people often find it preferable to reason with gist-based representations rather than verbatim descriptions of events (Reyna and Brainerd, 1995). When the gist does not contain adequate detail, the reasoning may lead to flawed conclusions. Some of the differences observed between highly skilled experts and novices might correspond to situations where

---

[*] This point directly confirms Tversky and Kahneman's (1973) original assumption that the availability heuristic should often result in good predictions.

experts have stored a large number of relevant instances and their solutions in memory, whereas novices have only gist-based representations. In such situations, novices will be forced to reason using the information provided. Experts, on the other hand, might be able to solve the problem with little or no reasoning, simply by retrieving the solution from memory. The latter situation would correspond to Klein's (1989, 1998) recognition-primed decision making. However, there is also reason to believe that people are more likely to develop abstract gist-type representations of events with experience (Reyna and Brainerd, 1995). This might explain the findings in some studies that people with less knowledge and experience sometimes outperform experts. A particularly interesting demonstration is given by Gigerenzer et al. (1999), who discuss a study where a simple recognition heuristic based on the collective recognition of the names of companies by 180 German laypeople resulted in a phenomenally high yield of 47% and outperformed the Dax 30 market index by 10%. It outperformed several mutual funds managed by professionals by an even greater margin.

Memory models and processes can also be used to explain primacy and recency effects in human judgment (Hogarth and Einhorn, 1992). Such effects seem similar to the well-known serial position effect.[*] Given that human judgment involves retrieval of information from memory, it seems reasonable that judgments would also show primacy and recency effects. Several mathematical models have been developed that show how the order in which evidence is presented to people might affect their judgments. For example, Hogarth and Einhorn (1992) present an anchoring and adjustment model of how people update beliefs that predicts both primacy and recency effects. The latter model holds that the degree of belief in a hypothesis after collecting $k$ pieces of evidence can be described as

$$S_k = S_{k-1} + w_k[s(x_k) - R] \tag{6}$$

where $S_k$ is the degree of belief after collecting $k$ pieces of evidence, $S_{k-1}$ is the anchor or prior belief, $w_k$ is the adjustment weight for the $k$th piece of evidence, $s(x_k)$ is the subjective evaluation of the $k$th piece of evidence, and $R$ is the reference point against which the $k$th piece of evidence is compared. In evaluation tasks, $R = 0$. This corresponds to the case where evidence is either for or against a hypothesis.[†] For estimation tasks, $R \neq 0$. The different values of $R$ result in an additive model for evaluation tasks and an averaging model for estimation tasks. Also, if the quantity $s(x_k) - R$ is evaluated for several pieces of evidence at a time, the model predicts primacy effects. If single pieces of

evidence are evaluated individually in a step-by-step sequence, recency effects become more likely.

Biases in human judgment, which in some but not all cases are memory related, can also be explained by models of how information is processed during task performance. Along these lines, Evans (1989) argues that factors which cause people to process information in a selective manner or attend to irrelevant information are the major cause of biases in human judgment. Evans's model of selective processing of information is consistent with other explanations of biases. Among such explanations, information overload has been cited as a reason for impaired decision making by consumers (Jacoby, 1977). The tendency of highly salient stimuli to capture attention during inference tasks has also been noted by several researchers (Nisbett and Ross, 1980; Payne, 1980). Nisbett and Ross suggest that vividness of information is determined by its emotional content, concreteness and imagability, and temporal and spatial proximity. As noted by Evans and many others, these factors have also been shown to affect the memorability of information. The conclusion is that biases due to salience can occur in at least two different ways: (1) People might focus on salient but irrelevant items while performing the task and (2) people might draw incorrect inferences when the contents of memory are biased due to salience effects during earlier task performance.

### Debiasing or Aiding Human Judgments

The notion that many biases (or deviations from normative models) in statistical estimation and inference can be explained has led researchers to consider the possibility of *debiasing* (a better term might be *improving*) human judgments (Keren, 1990). Part of the issue is that the heuristics people use often work very well. The nature of the heuristics also suggests some obvious generic strategies for improving decision making. One conclusion that follows directly from the earlier discussion is that biases related to the availability and representativeness heuristics might be reduced if people were provided better, more representative samples of information. Other strategies that follow directly from the earlier discussion include making ecologically valid cues more salient, providing both outcome and cognitive feedback, and helping people do analysis. These strategies can be implemented in training programs or guide the development of decision aids.[‡]

Emerging results from the field of naturalistic decision making support the conclusion that decision-making skills can be improved through training (Fallesen and Pounds, 2001; Pliske et al., 2001; Pliske and Klein, 2003). The use of computer-based training to develop task-specific decision-making skills is one very

---

[*] When people are asked to memorize lists of information, they almost always are able to remember items at the beginning or end of a list better than items in the middle of the list (Ebbinghaus, 1913). The improved ability for words at the start of the list is called the *primacy effect*. The improved ability for words at the end of the list is called the *recency effect*.

[†] It is easy to see that equation (6) approximates the log-odds form of Bayes' rule, where evidence for or against the hypothesis is combined additively.

[‡] These strategies seem to be especially applicable to the design of information displays and decision support systems. Chapter 42 addresses the issue of display design. Computer-based decision support is addressed in Section 4.2. These strategies also overlap with decision analysis. As discussed in Section 4.1, decision analysis focuses on the use of analytic methods to improve decision quality.

interesting development (Sniezek et al., 2002). Decision-making games (Pliske et al., 2001) and cognitive simulation (Satish and Streufert, 2002) are other approaches that have been applied successfully to improve decision-making skills. Other research shows that training in statistics reduces biases in judgment (Fong et al., 1986). In the latter study, people were significantly more likely to consider sample size after training.

These results supplement some of the findings discussed earlier, indicating that judgment biases can be moderated by familiarity with the task and the type of outcome information provided. Some of these results discussed earlier included evidence that providing feedback on the accuracy of weather forecasts may help weather forecasters (Winkler and Murphy, 1973), and research showing that cognitive feedback about cues and their relationship to the effects inferred leads to quicker learning than does feedback about outcomes (Balzer et al., 1989). Other studies have shown that simply asking people to write down reasons for and against their estimates of probabilities can improve calibration and reduce overconfidence (Koriat et al., 1980). This, of course, supports the conclusion that judgments will be less likely to be biased if people think carefully about their answers. Other research showed that subjects were less likely to be overconfident if they expressed subjective probabilities verbally instead of numerically (Zimmer, 1983; Wallsten et al., 1993). Conservatism, or the failure to modify probabilities adequately after obtaining evidence, was also reduced in Zimmer's study.

The results above support the conclusion that it might be possible to improve or aid human judgment. On the other hand, many biases, such as optimistic beliefs regarding health risks, have been difficult to modify (Weinstein and Klein, 1995). People show a tendency to seek out information that supports their personal views (Weinstein, 1979) and are quite resistant to information that contradicts strongly held beliefs (McGuire, 1966; Nisbett and Ross, 1980). Evans (1989) concludes that "pre-conceived notions are likely to prejudice the construction and evaluation of arguments." Other evidence shows that experts may have difficulty providing accurate estimates of subjective probabilities even when they receive feedback. For example, many efforts to reduce both overconfidence in probability estimates and the hindsight bias have been unsuccessful (Fischhoff, 1982). One problem is that people may not pay attention to feedback (Fischhoff and MacGregor, 1982). They also may attend only to feedback that supports their hypothesis, leading to poorer performance and at the same time greater confidence (Einhorn and Hogarth, 1978). Several efforts to reduce confirmation biases, the tendency to search for confirming rather than disconfirming evidence, through training have also been unsuccessful (Evans, 1989).[*]

---

[*] Engineers, designers, and other real-world decision makers will find it very debatable whether the confirmation bias is really a bias. Searching for disconfirming evidence obviously makes sense in hypothesis testing. That is, a single negative instance is enough to disprove a logical conjecture. In real-world settings, however, checking for evidence that supports a hypothesis can be very efficient.

The conclusion is that debiasing human judgments is difficult but not impossible. Some perspective can be obtained by considering that most studies showing biases have focused on statistical inference and generally involved people not particularly knowledgeable about statistics, who are not using decision aids such as computers or calculators. It naturally may be expected that people will perform poorly on such tasks, given their lack of training and forced reliance on mental calculations (von Winterfeldt and Edwards, 1986). The finding that people can improve their abilities on such tasks after training in statistics is particularly telling and also encouraging. Another encouraging finding is that biases are occasionally reduced when people process information verbally instead of numerically. This result might be expected given that most people are more comfortable with words than with numbers.

### 2.2.2 Preference and Choice

Much of the research on human preference and choice has focused on comparing observed preferences to the predictions of SEU theory (Goldstein and Hogarth, 1997). Early work examining SEU as a descriptive theory drew generally positive conclusions. However, it soon became apparent that people's preferences for risky or uncertain alternatives often violated basic axioms of SEU theory. The finding that people's preferences change when the outcomes are framed in terms of costs, as opposed to benefits, has been particularly influential. Several other common deviations from SEU have been observed. One potentially serious deviation is that preferences can be influenced by sunk costs or prior commitment to a particular alternative. Preferences change over time and may depend on which alternatives are being compared or even the order in which they are compared. The regret associated with making the "wrong" choice seems to play a major role when people compare alternatives. Accordingly, the satisfaction people derive from obtaining particular outcomes after making a decision is influenced by positive and negative expectations prior to making the decision. Other research on human preference and choice has shown that people choose between and apply different decision strategies depending on the cognitive effort required to apply a decision strategy successfully, the needed level of accuracy, and time pressure. Certain strategies are more likely than others to lead to choices consistent with those prescribed by SEU theory.

Alternative models, such as prospect theory and random-utility theory, were consequently developed to explain human preferences under risk or uncertainty.[†] The following discussion will first summarize some common violations of the axioms underlying SEU theory before moving on to framing effects and preference reversals. Attention will then shift to models of choice and preference. The latter discussion will begin with prospect theory before addressing other models of labile

---

[†] Singleton and Hovden (1987) and Yates (1992) are useful sources for the reader interested in additional details on risk perception, risk acceptability, and risk taking behavior. Section 2.3 is also relevant to this topic.

or conditional preferences. Decision-making strategies, and how people choose between them, are covered in Section 4.3.

***Violation of the Rationality Axioms*** Several studies have shown that people's preferences between uncertain alternatives can be inconsistent with the axioms underlying SEU theory. One fundamental violation of the assumptions is that preferences can be intransitive (Tversky, 1969; Budescu and Weiss, 1987). Also, as mentioned earlier, subjective probabilities may depend on the values of consequences (violating the independence axiom), and as discussed in the next section, the framing of a choice can affect preference. Another violation is given by the *Myers effect* (Myers et al., 1965), where preference reversals between high- (H) and low- (L) variance gambles can occur when the gambles are compared to a certain outcome, depending on whether the certain outcome is positive (H preferred to L) or negative (L preferred to H). The latter effect violates the assumption of independence because the ordering of the two gambles depends on the certain outcome.

Another commonly cited violation of SEU theory is that people show a tendency toward uncertainty avoidance, which can lead to behavior inconsistent with the "sure-thing" axiom. The Ellsberg and Allais paradoxes (Allais, 1953; Ellsberg, 1961) both involve violations of the sure-thing axiom (see Table 1) and seem to be caused by people's desire to avoid uncertainty. The *Allais paradox* is illustrated by the following set of gambles. In the first gamble, a person is asked to choose between gambles $A_1$ and $B_1$, where:

Gamble $A_1$ results in \$1 million for sure. Gamble $B_1$ results in \$2.5 million with a probability of 0.1, \$1 million with a probability of 0.89, and \$0 with a probability of 0.01.

In the second gamble, the person is asked to choose between gambles $A_2$ and $B_2$, where:

$A_2$ results in \$1 million with a probability of 0.11 and \$0 with a probability of 0.89. Gamble $B_2$ results in \$2.5 million with a probability of 0.1 and \$0 with a probability of 0.9.

Most people prefer gamble $A_1$ to $B_1$ and gamble $B_2$ to $A_2$. It is easy to see that this set of preferences violates expected utility theory. First, if $A_1 > B_1$, then $u(A_1) > u(B_1)$, meaning that $u(\$1 \text{ million}) > 0.1u(\$2.5 \text{ million}) + 0.89u(\$1 \text{ million}) + 0.01u(\$0)$. If a utility of 0 is assigned to receiving \$0 and a utility of 1 to receiving \$2.5 million, then $u(\$1 \text{ million}) > 1/11$. However, from the preference $A_2 > B_2$, it follows that $u(\$1 \text{ million}) < \$1/11$. Obviously, no utility function can satisfy this requirement of assigning a value both greater than and less than 1/11 to \$1 million.

Savage (1954) mentioned that the set of gambles above can be reframed in a way that shows that these preferences violate the sure-thing principle. After doing so, Savage found that his initial tendency toward choosing $A_1$ over $B_1$ and $A_2$ over $B_2$ disappeared. As noted by Stevenson et al. (1993), this example is one of the first cases cited of a preference reversal caused by reframing a decision, the topic discussed below.

***Framing of Decisions and Preference Reversals***
A substantial body of research has shown that people's preferences can shift dramatically depending on the way a decision is represented. The best known work on this topic was conducted by Tversky and Kahneman (1981), who showed that preferences between medical intervention strategies changed dramatically depending on whether the outcomes were posed as losses or gains. The following question, worded in terms of benefits, was presented to one set of subjects:

Imagine that the United States is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:

If program A is adopted, 200 people will be saved.

If program B is adopted, there is a 1/3 probability that 600 people will be saved and a 2/3 probability that no people will be saved.

Which of the two programs would you favor?

The results showed that 72% of subjects preferred program A. The second set of subjects was given the same cover story but worded in terms of costs:

If program C is adopted, 400 people will die.

If program D is adopted, there is a 1/3 probability that nobody will die and a 2/3 probability that 600 people will die.

Which of the two programs would you favor?

The results now showed that 78% of subjects preferred program D. Since program D is equivalent to B and program A is equivalent to C, the preferences for the two groups of subjects were strongly reversed. Tversky and Kahneman concluded that this reversal illustrated a common pattern in which choices involving gains are risk averse and choices involving losses are risk seeking. The interesting result was that the way the outcomes were worded caused a shift in preference for identical alternatives. Tversky and Kahneman called this tendency the *reflection effect*. A body of literature has since developed showing that the framing of decisions can have practical effects for both individual decision makers (Kahneman et al., 1982; Heath et al., 1994) and group decisions (Paese et al., 1993). On the other hand, recent research shows that the reflection effects can be reversed by certain outcome wordings (Kuhberger, 1995); more important, Kuhberger provides evidence that the reflection effect observed in the classic experiments can be eliminated by fully describing the outcomes (i.e., referring to the paragraph above, a more complete description would state: "If program C is adopted, 400 people will die *and* 200 *will live*").

Other recent research has explored the theory that perceived risk and perceived attractiveness of risky outcomes are psychologically distinct constructs (Weber et al., 1992). In the latter study, it was concluded that perceived risk and attractiveness are "closely related but distinct phenomena." Related research has shown weak negative correlations between the perceived risk

and value of indulging in alcohol-related behavior for adolescent subjects (Lehto et al., 1994). The latter study also showed that the rated propensity to indulge in alcohol-related behavior was strongly correlated with perceived value ($R = 0.8$) but weakly correlated with perceived risk ($R = -0.15$). Both findings are consistent with the theory that perceived risk and attractiveness are distinct constructs, but the latter finding indicates that perceived attractiveness may be the better predictor of behavior. Lehto et al. conclude that intervention methods attempting to lower preferences for alcohol-related behavior should focus on lowering perceived value rather than on increasing perceived risk.

***Prospect Theory*** Prospect theory (Kahneman and Tversky, 1979) attempts to account for behavior not consistent with the SEU model by including the framing of decisions as a step in the judgment of preference between risky alternatives. Prospect theory assumes that decision makers tend to be risk averse with regard to gains and risk seeking with regard to losses. This leads to a value function that weights losses disproportionately. As such, the model is still equivalent to SEU, assuming a utility function expressing mixed risk aversion and risk seeking. Prospect theory, however, assumes that the decision maker's reference point can change. With shifts in the reference point, the same returns can be viewed as either gains or losses.[*] The latter feature of prospect theory, of course, is an attempt to account for the framing effect discussed above. Prospect theory also deviates significantly from SEU theory in the way in which probabilities are addressed. To describe human preferences more closely, perceived values are weighted by a function $\pi(p)$ instead of the true probability, $p$. Compared to the untransformed form of $p$, $\pi(p)$ overweights very low probabilities and underweights moderate and high probabilities. The function $\pi(p)$ is also generally assumed to be discontinuous and poorly defined for probability values close to 0 or 1.

Prospect theory assumes that the choice process involves an editing phase and an evaluation phase. The editing phase involves reformulation of the options to simplify subsequent evaluation and choice. Much of this editing process is concerned with determining an appropriate reference point in a step called *coding*. Other steps that may occur include the segregation of riskless components of the decision, combining probabilities for events with identical outcomes, simplification by rounding off probabilities and outcome measures, and search for dominance. In the evaluation phase, the perceived values are then weighed by the function $\pi(p)$. The alternative with the greatest weighed value is then selected. Several other modeling approaches that differentially weigh utilities in risky decision making have been proposed (Goldstein and Hogarth, 1997). As in prospect theory, such models often assume that

the subjective probabilities, or decision weights, are a function of outcome sign (i.e., positive, neutral, or negative), rank (i.e., first, second, etc.), or magnitude. Other models focus on display effects (i.e., single-stage vs. multistage arrangements) and distribution effects (i.e., two outcome lotteries vs. multiple-outcome lotteries). Prospect theory and other approaches also address how the value or utility of particular outcomes can change between decision contexts, as discussed below.

More recently, Hertwig et al. (2004) divided decision from description, where full description of probability of risky events is given, and decision from experience, where decision makers learn the probability from experiences. They reported that two different decisions can lead to dramatically different choice behavior. In the case of decisions from description, people make choices as if they overweight the probability of rare events, as described by prospect theory. In contrast, the case of decisions from experience lead choices as if they underweight the probability of rare events. This idea created interesting debate because some (e.g., Fox and Hadar, 2006) attributed the division to simply sampling error.

***Labile Preferences*** There is no doubt that human preferences often change after receiving some outcome. After losing money, an investor may become risk averse. In other cases, an investor may escalate her commitment to an alternative after an initial loss, even if better alternatives are available. From the most general perspective, any biological organism becomes satiated after satisfying a basic need, such as hunger. Preferences also change over time or between decision contexts. For example, a 30-year-old decision maker considering whether to put money into a retirement fund may currently have a very different utility function than at retirement. The latter case is consistent with SEU theory but obviously complicates analysis.

Economists and behavioral researchers have both focused on mathematically modeling choice processes to explain intransitive or inconsistent preference orderings of alternatives (Goldstein and Hogarth, 1997). Game theory provides interesting insight into this issue. From this perspective, preferences of the human decision maker are modeled as the collective decisions obtained by a group of internal agents, or selves, each of which is assumed to have distinct preferences (see Elster, 1986). Intransitive preferences and other violations of rationality on the part of the human decision maker then arise from interactions between competing selves.[†] Along these lines, Ainslie (1975) proposed that impulsive preference switches (often resulting in risky

---

[*] The notion of a reference point against which outcomes are compared has similarities to the notion of making decisions on the basis of regret (Bell, 1982). Regret, however, assumes comparison to the best outcome. The notion of different reference points is also related to the well-known trend that buying and selling prices of assets often differ for a decision maker (Raiffa, 1968).

---

[†] As discussed further in Section 4.3, group decisions, even though they are made by rational members, are subject to numerous violations of rationality. For example, consider the case where the decision maker has three selves that are, respectively, risk averse, risk neutral, and risk seeking. Assume that the decision maker is choosing between alternatives $A$, $B$, and $C$. Suppose that the risk-averse self rates the alternatives in the order $A$, $B$, $C$; the risk-neutral self rates them in the order $B$, $C$, $A$; and the risk-seeking self rates them in the order $C$, $A$, $B$. Also assume that the selves are equally powerful. Then two of the three agents always agree that $A > B$, $B > C$, and $C > A$. This ordering is, of course, nontransitive.

or unhealthy choices) arise as the outcome of a struggle between selves representing conflicting short- and long-term interests, respectively.

Another area of active research has focused on how experiencing outcomes can cause shifts in preference. One robust finding is that people tend to be more satisfied if an outcome exceeds their expectations and less satisfied if it does not (i.e., Feather, 1966; Connolly et al., 1997). Expectations therefore provide a reference point against which outcomes are compared. A related result found in marketing studies is that negative experiences often have a much larger influence on product preferences and future purchasing decisions than positive experiences (Baumeister et al., 2001; Oldenburger et al., 2007).

Other studies have shown that people in a wide variety of settings often consider sunk costs when deciding whether to escalate their commitment to an alternative by investing additional resources (Arkes and Blumer, 1985; Arkes and Hutzel, 2000). From the perspective of prospect theory, sunk costs cause people to frame their choice in terms of losses instead of gains, resulting in risk-taking behavior and consequently, escalating commitment. Other plausible explanations for escalating commitment include a desire to avoid waste or to avoid blame for an initially bad decision to invest in the first place. Interestingly, some recent evidence suggests that people may deescalate commitment in response to sunk costs (Heath, 1995). The latter effect is also contrary to classical economic theory, which holds that decisions should be based solely on marginal costs and benefits. Heath explains such effects in terms of mental accounting. Escalation is held to occur when a mental budget is not set or expenses are difficult to track. Deescalation is held to occur when people exceed their mental budget, even if the marginal benefits exceed the marginal costs.

Other approaches include value and utility as random variables within models of choice to explain intransitive or inconsistent preference orderings of alternatives. The random utility model (Iverson and Luce, 1998) describes the probability $P_{a,A}$ of choosing a given alternative $a$ from a set of options $A$ as

$$P_{a,A} = \text{Prob } (U_a \geq U_b, \text{ for all } b \text{ in } A) \qquad (7)$$

where $U_a$ is the uncertain utility of alternative $a$ and $U_b$ is the uncertain utility of alternative $b$. The most basic random utility models assign a utility to each alternative by sampling a single value from a known distribution. The sampled utility of each alternative then remains constant throughout the choice process. Basic random utility models can predict a variety of preference reversals and intransitive preferences for single- and multiple-attribute comparisons of alternatives (i.e., Tversky, 1972).

Sequential sampling models extend this approach by assuming that preferences can be based on more than one observation. Preferences for particular alternatives are accumulated over time by integrating or otherwise summing the sampled utilities. The utility of an alternative at a particular time is proportional to the latter sum. A choice is made when the summed preferences for a particular alternative exceed some threshold, which itself may vary over time or depend on situational factors (Busemeyer and Townsend, 1993; Wallsten, 1995). It is interesting to observe that sequential sampling models can explain speed–accuracy trade-offs in signal detection tasks (Stone, 1960) as well as shifts in preferences due to time pressure (Busemeyer and Townsend, 1993; Wallsten, 1995) if it is assumed that people adjust their threshold downward under time pressure. That is, under time pressure, people sample less information before making a choice. In the following section we explore further how and why decision strategies might change over time and between decision contexts.

### 2.2.3 Adaptive Decision Behavior

The main idea of adaptive decision behavior, or contingent decision behavior, is that an individual decision maker uses different strategies in different situations (Payne et al., 1993). These strategies also include some short-cuts or heuristics that reduce the complexity of the problem while increasing chances to select suboptimal choices. Various decision strategies have been identified (Bettman et al., 1998; Wright, 1975) and will be described next. In addition, these different strategies have trade-off relationships. In other words, some strategies are less cognitively burdensome but their accuracy is also low. Other strategies are more cognitively burdensome, but their accuracy could be higher. This relationship is discussed below.

***Decision Strategies*** According to the taxonomy of Wright (1975), decision strategies are organized along two dimensions: "data combination processes" and "choice rules." Data combination processes have two levels, compensatory and noncompensatory data integration. In compensatory data integration, a good value of one attribute compensates for a bad value of another attribute. In contrast, noncompensatory data integration could drop a choice with a bad value of an attribute, even if the choice or alternative has perfect values for the other attributes (Edwards and Fasolo, 2001). The other dimension, choice rule, also has two levels: "best" and "cutoff." The best-choice rule chooses the best option through heuristics (e.g., choosing an alternative that has the highest number of good features and the smallest number of bad features), whereas the cutoff rule merely eliminates available options based on a decision maker's threshold (e.g., eliminating alternatives that have bad aspects that do not meet criteria until only a few alternatives remain).

Table 3 outlines these strategies and heuristics along the two dimensions just described. The distinction among levels in the two dimensions will be clarified when each decision strategy is discussed. In Table 3, three heuristics (WADD, EQW, and VOTE) are categorized in both the best and cutoff rules since the calculated score through these heuristics can be used as a cutoff criterion. Payne et al. (1988) also introduced some combination of multiple strategies such as EBA+MCD, EBA+WADD, and EBA+EQW.

In applying a weighted adding strategy (WADD), the subjective importance ratings are associated with

**Table 3  Decision Heuristics**

| Choice Rule | Data Combination Process | |
|---|---|---|
| | Compensatory | Noncompensatory |
| Best | WADD | LEX |
| | EQW | LEXSEMI |
| | VOTE | MAXIMAX |
| | MCD | MINIMAX |
| Cutoff | WADD | SAT |
| | EQW | EBA |
| | VOTE | |

*Source*: Adapted from Wright (1975) and Bettman et al. (1998).
*Abbreviations*: Weighted adding (WADD), lexicographic (LEX), lexicographic semiorder (LEXSEMI), satisficing (SAT), elimination by aspects (EBA), equal weight (EQW), majority of confirming dimensions (MCD), feature voting (VOTE), maximize minimum (MINIMAX), and maximize maximum (MAXIMAX)

each attribute. These subjective ratings, or weights, are multiplied by the corresponding attribute value obtained by a particular alternative. The worth of any particular alternative, then, would be the sum of these products (i.e., assigned weight by attribute value), and the decision maker selects the alternative with the greatest value. Equal weight (EQW) is a special case of weighted adding. Since each attribute is considered equally important in this method, the value of each alternative is simply the sum of all its attribute values.

The strategy of feature voting (VOTE) is based on the number or frequency of occurrences of positive or negative features within an alternative. Here, the decision maker subjectively defines the positive and negative values for each attribute. The worth of the alternative is determined by the number of good votes (positive features) and bad votes (negative features). The decision maker may also disregard the number of bad votes or focus on minimizing the number of bad votes.

Paired comparisons are central to the strategy of majority of confirming dimensions (MCD). In this strategy, decision makers start by evaluating the first two alternatives. They then count the number of times each alternative has the higher score across all the attributes. The alternative with the higher count is then compared with the next alternative, and the process is repeated until all the alternatives have been considered. Finally, the alternative with the most "wins" is selected.

The lexicographic ordering principle (LEX) (Fishburn, 1974) considers the case where alternatives have multiple consequences or attributes. For example, a purchasing decision might be based on both the cost and performance of the product considered. The various consequences are first ordered in terms of their importance. Returning to the example above, performance might be considered more important than cost. The decision maker then compares each alternative sequentially, beginning with the most important consequence. If an alternative is found that is better than the others on the first consequence, it is selected immediately. If no alternative is best on the first dimension, the alternatives are compared for the next most important consequence.

This process continues until an alternative is selected or all the consequences have been considered without making a choice. The latter situation can happen only if the alternatives have the same consequences.

The lexicographic semiorder (LEXSEMI) method is a variation of LEX. In LEXSEMI, the condition of "equally good" is loosened by introducing the concept of "just-noticeable difference." For example, if car A is $20,000 and car B is $20,050 (but has better gas mileage), a car buyer using LEX will choose car A because car A is cheaper than car B. However, if another car buyer using LEXSEMI looks at gas mileage, which may be the second most important attribute, he may choose car B since $50 would become insignificant compared with future savings on gas costs.

The satisficing (SAT) method is a matter of selecting the first alternative that meets all the decision maker's requirements on the attributes. As such, decisions resulting from this technique depend on the order that the alternatives are presented. For example, a person considering the purchase of a car might stop looking once he or she has found an attractive deal, instead of comparing every model on the market. More formally, the comparison of alternatives stops once a choice is found that exceeds a minimum aspiration level $S_{ik}$ for each of its consequences $C_{ik}$ over the possible events $E_k$. Satisficing can be a normative decision rule when (1) the expected benefit of exceeding the aspiration level is small, (2) the cost of evaluating alternatives is high, or (3) the cost of finding new alternatives is high. More often, however, it is viewed as an alternative to maximizing decision rules. From this view, people cope with incomplete or uncertain information and their limited rationality by satisficing in many settings instead of optimizing (Simon, 1955, 1982).

Two similar noncompensatory decision strategies are minimize maximum loss (MINIMAX) and maximize maximum gains (MAXIMAX). A decision maker "applying MINIMAX compares the options on their worst attributes, rejecting one if another worst attributes is less offensive or if another has fewer worst attributes that are equally offensive. That is, the decision maker minimizes the maximum possible loss. MAXIMAX implies a decision maker compares options on their best attribute, choosing one over another if its best attribute is more desirable or if it possesses more best attributes of equal desirability" (Wright, 1975, p. 61).

Elimination by aspects (EBA) (Tversky, 1972) is a closely related sequential strategy of eliminating alternatives that do not meet selected criteria. EBA is similar to the LEX. It differs in that the consequences used to compare the alternatives are selected in random order, where the probability of selecting a consequence dimension is proportional to its importance. Often, the most important aspect (or attribute) is selected for inspection. If other alternatives have bad values for the attributes, they are eliminated from the candidates. For example, if one thinks that gas mileage is important, he can eliminate cars that have bad mileage, say less than 20 miles per gallon, from the pool of candidates.

Even though these strategies describe more accurately how people make decisions than the normative

approaches, they also have some problems. Some non-compensatory methods mistakenly eliminate the optimal alternative(s). For example, if a decision maker uses EBA, and if the optimal alternative has a bad value for a certain attribute, the optimal alternative can be eliminated from the pool of candidates due to this particular bad attribute even though the overall utility may be better than that of the remaining alternatives. Even compensatory strategies pose difficulties for those seeking to make the optimal choice. For example, if one uses MCD, the choice can vary depending on the order of comparisons, as the decision maker may eliminate a good alternative merely because it takes a "loss" in an early comparison and may have won several other comparisons with the remaining alternative.

***Effort-Accuracy Framework*** The theory of contingent decision making describes the trade-off situations in more detail (Payne et al., 1993). Payne and his colleagues measure the effort expended and accuracy resulting from the application of different strategies. They assume that performing each strategy has associated costs in the form of small information-processing tasks referred to as elementary information processes (EIPs). The number of EIPs for a decision strategy is assumed to be the measure of effort in making the decision. The accuracy of a strategy is measured in relative terms; "that is, the quality of the choice expected from a rule [strategy] is measured against the standard of accuracy provided by a normative model like the weighted additive rule [WADD]" (Payne et al., 1993, p. 93). (*Note*: Brackets have been added for terminology consistency.) As shown in Figure 3, use of the WADD method represents the most accurate, albeit most costly (in terms of effort), strategy to perform. In contrast, EBA is the most effortless strategy, but accuracy must often be sacrificed. Note that random choice is not considered here because it is not considered a method of decision making per se.
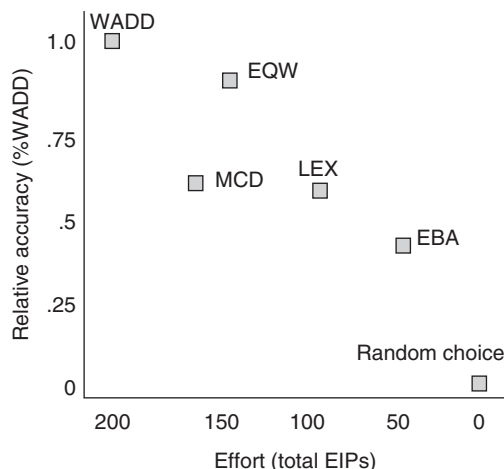


**Figure 3** Trade-off between relative accuracy and effort. (Adapted from Payne et al., 1993.)

The question then is how to help people make decisions more accurately and effortlessly. Todd and Benbasat (1991, 1993) conducted a series of empirical studies showing the effectiveness of computer-based decision aids. They showed that "when a more accurate normative strategy [WADD] is made less effortful to use, it is used" (Todd and Benbasat, 2000, p. 91). (*Note*: Brackets have been added for terminology consistency.) This decision aid helps decision makers perform more accurate decision strategies with less effort. Another interesting result of the same study is that although the decision aid also provided features to support noncompensatory strategies (EBA), the noncompensatory strategies were not significantly promoted. Todd and Benbasat argued that the noncompensatory strategies are only preferred when support for the compensatory strategies is low.

However, it should be noted that supporting a compensatory strategy is cumbersome. According to Edwards and Fasolo (2001), who reviewed representative Web-based decision aids, compensatory strategies have been sparingly employed by Web-based decision aids since interfaces supporting compensatory strategies are complicated and difficult to use. Thus, providing proper decision aids that assist target users requires some additional design effort to minimize the required efforts. Some creative solutions, probably using InfoVis techniques, will be required.

Another lesson from reviewing these results is that information overload plays an important role in selecting proper decision strategies. Under high information overload, noncompensatory strategies such as EBA appear to be more effective since they filter out unnecessary information. After filtering is over, compensatory strategies or relatively normative approaches appear to be more effective since more comprehensive comparisons among the alternatives are necessary.

### 2.2.4 Behavioral Economics

Behavioral economics is a subdiscipline of economics formulated as a backlash against neoclassical economics whose main tenet is strong reliance on rationality of human decision makers (Camerer and Loewenstein, 2004; Simon, 1987). In the middle of the twentieth century, it was shown that many axioms based on human rationality are often violated by decision models. As well as people are limited in their information processing (Simon, 1982), people care more about the fairness over self-interest (ultimatum game) (Thaler, 1988); people weight risky outcomes in a nonlinear fashion (the prospect theory) (Kahneman and Tversky, 1979); and people value a thing more after they possess it (endowment effect) (Thaler, 1980). Such evidence led many economists to reappreciate the values of psychology to model and predict decision-making behaviors, which led the growth of behavioral economics as a strong subdiscipline (Pesendorfer, 2006).

However, behavioral economics is neither a totally new idea nor a drastic change in the core idea of neoclassical economics. Human psychology was already well understood by Adam Smith, who is the father of

modern economics. His books include not only *The Wealth of Nations* but also *Theory of Moral Sentiments*, which is less known but has profound insights about human psychology. It is interesting to see how eloquently he described loss-aversion in his book, "We suffer more . . . when we fall from a better to a worse situation, than we ever enjoy when we rise from a worse to a better" (Smith, 1875, p. 331). In addition, it is difficult to say that behavioral economics changed the core idea of neoclassical economics. That is, maximizing expected value is still worthwhile as a core framework to help understand many economic behaviors. Behavioral economics instead provides more capability and external validity by adding additional parameters in traditional economic models and theories.

Some criticize behavioral economics as well. First, it has been claimed that behavioral economics did not provide a unified and elegant theory of rational agents. However, a lack of a unified model not only reflects the diversity and complexity of human decision making but also conversely demonstrates the incompleteness of traditional economic models based on rationality (Kahneman, 2003). Second, some doubt that findings in behavioral economics have significant influences on real-life situations, such as policy making (Brannon, 2008). However, recent publications for the general public (Ariely, 2008; Levitt and Dubner, 2005; Thaler and Sunstein, 2008) have demonstrated direct implications on policy making and economics. More specifically, behavioral economics has been used to understand behaviors in macroeconomics, saving, labor economics, and finance. For example, many people do not mind decreases in their real wage, considering inflation, as long as there is no decrease in their nominal wage, without considering inflation (Shafir et al., 1997). This behavioral pattern is called "the money illusion."

Theory in behavioral economics has been covered largely in this section, but several new directions of behavioral economics have been observed, such as understanding emotion, using the method of neuropsychology to better understand human behavior. These directions have not fully explored, but these could provide interesting new insights for us. More comprehensive review for behavioral economics and behavioral finance can be found elsewhere (Barberis and Thaler, 2003; Camerer et al., 2004; Glaser et al., 2004).

## 2.3 Naturalistic Decision Models

In a dynamic and realistic environment, actions taken by a decision maker are made sequentially in time. Taking actions can change the environment, resulting in a new set of decisions. The decisions might be made under time pressure and stress by groups or by single decision makers. This process might be performed on a routine basis or might involve severe conflict. For example, either a group of soldiers or an individual officer might routinely identify marked vehicles as friends or foes. When a vehicle has unknown or ambiguous marking, the decision changes to a conflict-driven process. Naturalistic decision theory has emerged as a new field that focuses on such decisions in real-world environments (Klein, 1998; Klein et al., 1993). The notion that most decisions are made in a routine, nonanalytical way is the driving force of this approach. Areas where such behavior seems prominent include juror decision making, troubleshooting of complex systems, medical diagnosis, management decisions, and numerous other examples.

For many years, it has been recognized that decision making in natural environments often differs greatly between decision contexts (Beach, 1993; Hammond, 1993). In addressing this topic, the researchers involved often question the relevance and validity of both classical decision theory and behavioral research not conducted in real-world settings (Cohen, 1993). Numerous naturalistic models have been proposed (Klein et al., 1993). These models assume that people rarely weigh alternatives and compare them in terms of expected value or utility. Each model is also descriptive rather than prescriptive. Perhaps the most general conclusion that can be drawn from this work is that people use different decision strategies, depending on their experience, the task, and the decision context. Several of the models also postulate that people choose between decision strategies by trading off effectiveness against the effort required.

In the following discussion we address several models of dynamic and naturalistic decision making: (1) levels of task performance (Rasmussen, 1983), (2) recognition-primed decisions (Klein, 1989), (3) dominance structuring (Montgomery, 1989), and (4) explanation-based decision making (Pennington and Hastie, 1988).

### 2.3.1 Levels of Task Performance

There is growing recognition that most decisions are made on a routine basis in which people simply follow past behavior patterns (Rasmussen, 1983; Svenson, 1990; Beach, 1993). Rasmussen (1983) follows this approach to distinguish among skill-based, rule-based, and knowledge-based levels of task performance. Lehto (1991) further considers judgment-based behavior as a fourth level of performance. Performance is said to be at either a skill-based or a rule-based level when tasks are routine in nature. *Skill-based performance* involves the smooth, automatic flow of actions without conscious decision points. As such, skill-based performance describes the decisions made by highly trained operators performing familiar tasks. *Rule-based performance* involves the conscious perception of environmental cues, which trigger the application of rules learned on the basis of experience. As such, rule-based performance corresponds closely to recognition-primed decisions (Klein, 1989). *Knowledge-based performance* is said to occur during learning or problem-solving activity during which people cognitively simulate the influence of various actions and develop plans for what to do. The *judgment-based level* of performance occurs when affective reactions of a decision maker cause a change in goals or priorities between goals (Janis and Mann, 1977; Etzioni, 1988; Lehto, 1991). Distinctive types of errors in decision making occur at each of the four levels (Reason, 1990; Lehto, 1991).

At the skill-based level, errors occur due to perceptual variability and when people fail to shift up

to rule-based or higher levels of performance. At the rule-based level, errors occur when people apply faulty rules or fail to shift up to a knowledge-based level in unusual situations where the rules they normally use are no longer appropriate. The use of faulty rules leads to an important distinction between running and taking risks. Along these lines, Wagenaar (1992) discusses several case studies in which people following risky forms of behavior do not seem to be consciously evaluating the risk. Drivers, in particular, seem habitually to take risks. Wagenaar explains such behavior in terms of faulty rules derived on the basis of benign experience. In other words, drivers get away with providing small safety margins most of the time and consequently learn to run risks on a routine basis. Drucker (1985) points out several cases where organizational decision makers have failed to recognize that the generic principles they used to apply were no longer appropriate, resulting in catastrophic consequences.

At the knowledge-based level, errors occur because of cognitive limitations or faulty mental models or when the testing of hypotheses causes unforeseen changes to systems. At judgment-based levels, errors (or violations) occur because of inappropriate affective reactions, such as anger or fear (Lehto, 1991). As noted by Isen (1993), there also is growing recognition that positive affect can influence decision making. For example, positive affect can promote the efficiency and thoroughness of decision making but may cause people to avoid negative materials. Positive affect also seems to encourage risk-averse preferences. Decision making itself can be anxiety provoking, resulting in violations of rationality (Janis and Mann, 1977).

A study involving drivers arrested for drinking and driving (McKnight et al., 1995) provides an interesting perspective on how the sequential nature of naturalistic decisions can lead people into traps. The study also shows how errors can occur at multiple levels of performance. In this example, decisions made well in advance of the final decision to drive while impaired played a major role in creating situations where drivers were almost certain to drive impaired. For example, the driver may have chosen to bring along friends and therefore have felt pressured to drive home because the friends were dependent on him or her. This initial failure by drivers to predict the future situation could be described as a failure to shift up from a rule-based level to a knowledge-based level of performance. In other words, the driver never stopped to think about what might happen if he or she drank too much. The final decision to drive, however, would correspond to an error (or violation) at the judgment-based level if the driver's choice was influenced by an affective reaction (perceived pressure) to the presence of friends wanting a ride.

### 2.3.2 Recognition-Primed Decision Making

Klein (1998, 2004) developed the theory of recognition-primed decision making on the basis of observations of firefighters and other professionals in their naturalistic environments. He found that up to 80% of the decisions made by firefighters involved some sort of situation recognition, where the decision makers simply followed a past behavior pattern once they recognized the situation.

The model he developed distinguishes between three basic conditions. In the simplest case, the decision maker recognizes the situation and takes the obvious action. A second case occurs when the decision maker consciously simulates the action to check whether it should work before taking it. In the third and most complex case, the action is found to be deficient during the mental simulation and is consequently rejected. An important point of the model is that decision makers do not begin by comparing all the options. Instead, they begin with options that seem feasible based on their experience. This tendency, of course, differs from the SEU approach but is comparable to applying the satisficing decision rule (Simon, 1955) discussed earlier.

Situation assessment is well recognized as an important element of decision making in naturalistic environments (Klein et al., 1993). Recent research by Klein and his colleagues has examined the possibility of enhancing situation awareness through training (Klein and Wolf, 1995). Klein and his colleagues have also applied methods of cognitive task analysis to naturalistic decision-making problems. In these efforts they have focused on identifying (1) critical decisions, (2) the elements of situation awareness, (3) critical cues indicating changes in situations, and (4) alternative courses of action (Klein, 1995). Accordingly, practitioners of naturalistic decision making tend to focus on process-tracing methods and behavioral protocols (Ericsson and Simon, 1984) to document the processes people follow when they make decisions.[*]

### 2.3.3 Dominance Structuring

Dominance structuring (Montgomery, 1989; Montgomery and Willen, 1999) holds that decision making in real contexts involves a sequence of four steps. The process begins with a preediting stage in which alternatives are screened from further analysis. The next step involves selecting a promising alternative from the set of alternatives that survive the initial screening. A test is then made to check whether the promising alternative dominates the other surviving alternations. If dominance is not found, the information regarding the alternatives is restructured in an attempt to force dominance. This process involves both the bolstering and deemphasizing of information in a way that eliminates disadvantages of the promising alternative.

Empirical support can be found for each of the four stages of the bolstering process (Montgomery and Willen, 1999). Consequently, this theory may have value as a description of how people make nonroutine decisions.

### 2.3.4 Explanation-Based Decision Making

Explanation-based decision making (Oskarsson et al., 2009; Pennington and Hastie, 1986, 1988) assumes that people begin their decision-making process by

---

[*] Goldstein and Hogarth (1997) describe a similar trend in judgment and decision-making research.

constructing a mental model that explains the facts they have received. While constructing this explanatory model, people are also assumed to be generating potential alternatives to choose between. The alternatives are then compared to the explanatory model rather than to the facts from which it was constructed.

Pennington and Hastie have applied this model to juror decision making and obtained experimental evidence that many of its assumptions seem to hold. They note that juror decision making requires consideration of a massive amount of data that is often presented in haphazard order over a long time period. Jurors seem to organize this information in terms of stories describing causation and intent. As part of this process, jurors are assumed to evaluate stories in terms of their uniqueness, plausibility, completeness, or consistency. To determine a verdict, jurors then judge the fit between choices provided by the trial judge and the various stories they use to organize the information. Jurors' certainty about their verdict is assumed to be influenced both by evaluation of stories and by the perceived goodness of fit between the stories and the verdict.

## 3 GROUP DECISION MAKING

Much research has been done over the past 25 years or so on decision making by groups and teams. Most of this work has focused on groups as opposed to teams. In a team it is assumed that the members are working toward a common goal and have some degree of interdependence, defined roles and responsibilities, and task-specific knowledge (Orasanu and Salas, 1993). Team performance is a major area of interest in the field of naturalistic decision theory (Klein et al., 1993; Klein, 1998), as discussed earlier. Group performance has traditionally been an area of study in the fields of organizational behavior and industrial psychology. Traditional decision theory has also devoted some attention to group decision making (Raiffa, 1968; Keeney and Raiffa, 1976). In the following discussion we first discuss briefly some of the ways that group decisions differ from those made by isolated decision makers who need to consider only their own preferences. That is, ethics and social norms play a much more prominent role when decisions are made by or within groups. Attention will then shift to group processes and how they affect group decisions. In the last section we address methods of supporting or improving group decision making.

### 3.1 Ethics and Social Norms

When decisions are made by or within groups, a number of issues arise that have not been touched on in the earlier portions of this chapter. To start, there is the complication that preferences may vary between members of a group. It often is impossible to maximize the preferences of all members of the group, meaning that trade-offs must be made and issues such as fairness must be addressed to obtain acceptable group decisions. Another complication is that the return to individual decision makers can depend on the actions of others.

Game theory[*] distinguishes two common variations of this situation. In competitive games, individuals are likely to take "self-centered" actions that maximize their own return but reduce returns to other members of the group. Behavior of group members in this situation may be well described by the minimax decision rule discussed in Section 2.2.3. In cooperative games, the members of the group take actions that maximize returns to the group as a whole.

Members of groups may choose cooperative solutions that are better for the group as a whole for many different reasons (Dawes et al., 1988). Groups may apply numerous forms of coercion to punish members who deviate from the cooperative solutions. Group members may apply decision strategies such as reciprocal altruism. They also might conform because of their social conscience, a need for self-esteem, or feelings of group identity. Fairness considerations can in some case explain preferences and choices that seem to be in conflict with economic self-interest (Bazerman, 1998). Changes in the status quo, such as increasing the price of bottled water immediately after a hurricane, may be viewed as unfair even if they are economically justifiable based on supply and demand. People are often willing to incur substantial costs to punish "unfair" opponents and reward their friends or allies. The notion that costs and benefits should be shared equally is one fairness-related heuristic that people use (Messick, 1991). Consistent results were found by Guth et al. (1982) in a simple bargaining game where player 1 proposes a split of a fixed amount of cash and player 2 either accepts the offer or rejects it. If player 2 rejects the offer, both players receive nothing. Classical economics predicts that player 2 will accept any positive amount (i.e., player 2 should always prefer something to nothing). Consequently, player 1 should offer player 2 a very small amount greater than zero. The results showed that contrary to predictions of classical economics, subjects tended to offer a substantial proportion of the cash (the average offer was 30%). Some of the subjects rejected positive offers. Others accepted offers of zero. Further research, summarized by Bolton and Chatterjee (1996), confirms these findings that people seem to care about whether they receive their fair share.

Ethics clearly plays an important role in decision making. Some choices are viewed by nearly everyone as being immoral or wrong (i.e., violations of the law, dishonesty, and numerous other behaviors that conflict with basic societal values or behavioral norms). Many corporations and other institutions formally specify codes of ethics prescribing values such as honesty, fairness, compliance with the law, reliability, considerance or sensitivity to cultural differences, courtesy, loyalty, respect for the environment, and avoiding waste. It is easy to visualize scenarios where it is in the best interest of a decision maker to choose economically undesirable options (at least in the short term) to comply with ethical codes. According to Kidder (1995), the "really tough choices...don't center on right versus

---

[*] Friedman (1990) provides an excellent introduction to game theory.

wrong. They involve right versus right." Kidder refers to four dilemmas of right versus right that he feels qualify as paradigms: (1) truth versus loyalty (i.e., whether to divulge information provided in confidence), (2) individual versus community, (3) short term versus long term, and (4) justice versus mercy. At least three principles, which in some cases provide conflicting solutions, have been proposed for resolving ethical dilemmas. These include (1) utilitarianism, selecting the option with the best overall consequences; (2) rule based, following a rule regardless of its current consequences (i.e., waiting for a stop light to turn green even if no cars are coming); and (3) fairness, doing what you would want others to do for you.

Numerous social dilemmas also occur in which the payoffs to each participant result in individual decision strategies harmful to the group as a whole. The tragedy of the commons (Hardin, 1968) is illustrative of social dilemmas in general. For example, as discussed in detail by Baron (1998), consider the crash of the east coast commercial fishing industry, brought about by overfishing. Baron suggests that individual fishers may reason that if they do not catch the fish, someone else will. Each fisher then attempts to catch as many fish as possible, even if this will cause the fish stocks to crash. Despite the fact that cooperative solutions, such as regulating the catch, are obviously better than the current situation, individual fishers continue to resist such solutions. Regulations are claimed to infringe on personal autonomy, to be unfair, or to be based on inadequate knowledge.

Similar examples include littering, wasteful use of natural resources, pollution, social free riding, and corporations expecting governments or affected consumers to pay for the cost of accidents (e.g., oil spill and subprime mortgage crisis). These behaviors can all be explained in terms of the choices faced by the offending individual decision maker (Schelling, 1978). Simply put, the individual decision maker enjoys the benefits of the offensive behavior, as small as they may be, but the costs are incurred by the entire group.

## 3.2 Group Processes

A large amount of research has focused on groups and their behavior. Accordingly, many models have been developed that describe how groups make decisions. A common observation is that groups tend to move through several phases as they go through the decision-making process (Ellis and Fisher, 1994). One of the more classic models (Tuckman, 1965) describes this process with four words: forming, storming, norming, and performing. *Forming* corresponds to initial orientation, *storming* to conflict, *norming* to developing group cohesion and expressing opinions, and *performing* to obtaining solutions. As implied by Tuckman's choice of terms, there is a continual interplay between socioemotive factors and rational, task-oriented behavior throughout the group decision-making process. Conflict, despite its negative connotations, is a normal, expected aspect of the group decision process and can in fact serve a positive role (Ellis and Fisher, 1994). In the following discussion we first address causes and effects of group conflict, then shift to conflict resolution.

### 3.2.1 Conflict

Whenever people or groups have different preferences, conflict can occur. As pointed out by Zander (1994), conflict between groups becomes more likely when groups have fuzzy or potentially antagonistic roles or when one group is disadvantaged (or perceives that it is not being treated fairly). A lack of conflict-settling procedures and separation or lack of contact between groups can also contribute to conflict. Conflict becomes especially likely during a crisis and often escalates when the issues are perceived to be important or after resistance or retaliation occurs. Polarization, loyalty to one's own group, lack of trust, and cultural and socioeconomic factors are often contributing factors to conflict and conflict escalation.

Ellis and Fisher (1994) distinguish between affective and substantive forms of conflict. *Affective conflict* corresponds to emotional clashes between individuals or groups; *substantive conflict* involves opposition at the intellectual level. Substantive conflict is especially likely to have positive effects on group decisions by promoting better understanding of the issues involved. Affective conflict can also improve group decisions by increasing interest, involvement, and motivation among group members and, in some cases, cohesiveness. On the other hand, affective conflict may cause significant ill-will, reduced cohesiveness, and withdrawal by some members from the group process. Baron (1998) provides an interesting discussion of violent conflict and how it is related to polarized beliefs, group loyalty, and other biases.

Defection and the formation of coalitions is a commonly observed effect of conflict, or power struggles, within groups. Coalitions often form when certain members of the group can gain by following a common course of action at the expense of the long-run objectives of the group as a whole. Rapidly changing coalitions between politicians and political parties are obviously a fact of life. Another typical example is when a subgroup of technical employees leaves a corporation to form their own small company, producing a product similar to one they had been working on. Coalitions, and their formation, have been examined from decision-analytic and game theory perspectives (Raiffa, 1982; Bolton and Chatterjee, 1996). These approaches make predictions regarding what coalitions will form, depending on whether the parties are cooperating or competing, which have been tested in a variety of experiments (Bolton and Chatterjee, 1996). These experiments have revealed that the formation of coalitions is influenced by expected payoffs, equity issues, and the ease of communication. However, Bazerman (1998) notes that the availability heuristic, overconfidence, and sunk cost effects are likely to explain how coalitions actually form in the real world.

### 3.2.2 Conflict Resolution

Groups resolve conflict in many different ways. Discussion and argument, voting, negotiation, arbitration, and other forms of third-party intervention are all methods of resolving disputes. Discussion and argument are clearly

the most common methods followed within groups to resolve conflict. Other methods of conflict resolution normally play a complementary rather than a primary role in the decision process. That is, the latter methods are relied on when groups fail to reach consensus after discussion and argument or they simply serve as the final step in the process.

Group discussion and argument are often viewed as constituting a less than rational process. Along these lines, Brashers et al. (1994) state that the literature suggests "that argument in groups is a social activity, constructed and maintained in interaction, and guided perhaps by different rules and norms than those that govern the practice of ideal or rational argument. Subgroups speaking with a single voice appear to be a significant force . . . . Displays of support, repetitive agreement, and persistence all appear to function as influence mechanisms in consort with, or perhaps in place of, the quality or rationality of the arguments offered." Brashers et al. also suggest that members of groups appear uncritical because their arguments tend to be consistent with social norms rather than the rules of logic: "[S]ocial rules such as: (a) submission to higher status individuals, (b) experts' opinions are accepted as facts on all matters, (c) the majority should be allowed to rule, (d) conflict and confrontation are to be avoided whenever possible."

A number of approaches for conflict management have been suggested that attempt to address many of the issues raised by Brashers et al. These approaches include seeking consensus rather than allowing decisions to be posed as win–lose propositions, encouraging and training group members to be supportive listeners, deemphasizing status, depersonalizing decision making, and using facilitators (Likert and Likert, 1976). Other approaches that have been proposed include directing discussion toward clarifying the issues, promoting an open and positive climate for discussion, facilitating face-saving communications, and promoting the development of common goals (Ellis and Fisher, 1994).

Conflicts can also be resolved through voting and negotiation, as discussed further in Section 3.3. Negotiation becomes especially appropriate when the people involved have competing goals and some form of compromise is required. A typical example would be a dispute over pay between a labor union and management. Strategic concerns play a major role in negotiation and bargaining (Schelling, 1960). Self-interest on the part of the involved parties is the driving force throughout a process involving threats and promises, proposals and counterproposals, and attempts to discern how the opposing party will respond. Threats and promises are a means of signaling what the response will be to actions taken by an opponent and consequently become rational elements of a decision strategy (Raiffa 1982). Establishing the credibility of signals sent to an opponent becomes important.

Methods of attaining credibility include establishing a reputation, the use of contracts, cutting off communication, burning bridges, leaving an outcome beyond control, moving in small steps, and using negotiating agents (Dixit and Nalebuff, 1991). Given the fundamentally adversarial nature of negotiation, conflict may

move from a substantive basis to an affective, highly emotional state. At this stage, arbitration and other forms of third-party intervention may become appropriate, due to a corresponding tendency for the negotiating parties to take extreme, inflexible positions.

## 3.3 Group Performance and Biases

The quality of the decisions made by groups in a variety of different settings has been seriously questioned. Part of the issue here is the phenomenon of *groupthink*, which has been blamed for several disastrous public policy decisions (Janis, 1972; Hart et al., 1997). Eight symptoms of groupthink cited by Janis and Mann (1977) are the illusion of invulnerability, rationalization (discounting of warnings and negative feedback), belief in the inherent morality of the group, stereotyping of outsiders, pressure on dissenters within the group, self-censorship, illusion of unanimity, and the presence of mindguards who shield the group from negative information. Janis and Mann proposed that the results of groupthink include failure to consider all the objectives and alternatives, failure to reexamine choices and rejected alternatives, incomplete or poor search for information, failure to adequately consider negative information, and failure to develop contingency plans. Groupthink is one of the most cited characteristics of how group decision processes can go wrong. Given the prominence of groupthink as an explanation of group behavior, it is somewhat surprising that only a few studies have evaluated this theory empirically. Empirical evaluation of the groupthink effect and the development of alternative modeling approaches continue to be active areas of research (Hart et al., 1997).

Other research has attempted to measure the quality of group decisions in the real world against rational, or normative, standards. Viscusi (1991) cites several examples of apparent regulatory complacency and regulatory excess in government safety standards in the United States. He also discusses a variety of inconsistencies in the amounts awarded in product liability cases. Baron (1998) provides a long list of what he views as errors in public decision making and their very serious effects on society. These examples include collective decisions resulting in the destruction of natural resources and overpopulation, strong opposition to useful products such as vaccines, violent conflict between groups, and overzealous regulations, such as the Delaney clause. He attributes these problems to commonly held, and at first glance innocuous, intuitions such as do no harm, nature knows best, and be loyal to your own group, the need for retribution (an eye for an eye), and a desire for fairness.

A significant amount of laboratory research is available that compares the performance of groups to that of individual decision makers (Davis, 1992; Kerr et al., 1996). Much of the early work showed that groups were better than individuals on some tasks. Later research indicated that group performance is less than the sum of its parts. Groups tend to be better than individuals on tasks where the solution is obvious once it is advocated by a single member of the group (Davis, 1992; Kerr et al., 1996). Another commonly cited finding is that groups tend to be more willing than individuals, to select

risky alternatives, but in some cases the opposite is true. One explanation is that group interactions cause people within the group to adopt more polarized opinions (Moscovici, 1976). Large groups seem especially likely to reach polarized, or extreme, conclusions (Isenberg, 1986). Groups also tend to overemphasize the common knowledge of members, at the expense of underemphasizing the unique knowledge certain members have (Stasser and Titus, 1985; Gruenfeld et al., 1996). A more recent finding indicates that groups were more rational than individuals when playing the ultimatum game (Bornstein and Yaniv, 1998).

Duffy (1993) notes that teams can be viewed as information processes and cites team biases and errors that can be related to information-processing limitations and the use of heuristics, such as framing. Topics such as mediation and negotiation, jury decision making, and public policy are now being evaluated from the latter perspective (Heath et al., 1994). Much of this research has focused on whether groups use the same types of heuristics and are subject to the same biases of individuals. This research has shown (1) framing effects and preference reversals (Paese et al., 1993), (2) overconfidence (Sniezek, 1992), (3) use of heuristics in negotiation (Bazerman and Neale, 1983), and (4) increased performance with cognitive feedback (Harmon and Rohrbaugh, 1990). One study indicated that biasing effects of the representativeness heuristic were greater for groups than for individuals (Argote et al., 1986). The conclusion is that group decisions may be better than those of individuals in some situations but are subject to many of the same problems.

## 3.4  Prescriptive Approaches

A wide variety of prescriptive approaches have been proposed for improving group decision making. The approaches address some of the foregoing issues, including the use of agendas and rules of order, idea-generating techniques such as brainstorming, nominal group and Delphi techniques, decision structuring, and methods of computer-mediated decision making. As noted by Ellis and Fisher (1994), there is conflicting evidence regarding the effectiveness of such approaches. On the negative side, prescriptive approaches might stifle creativity in some situations and can be sabotaged by dissenting members of groups. On the positive side, prescriptive approaches make the decision process more orderly and efficient, promote rational analysis and participation by all members of the group, and help ensure implementation of group decisions. In the following discussion we review briefly some of these tools for improving group decision making.

### 3.4.1  Agendas and Rules of Order

Agendas and rules of order are often essential to the orderly functioning of groups. As noted by Welch (1994), an agenda "conveys information about the structure of a meeting: time, place, persons involved, topics to be addressed, perhaps suggestions about background material or preparatory work." Agendas are especially important when the members of a group are loosely coupled or do not have common expectations. Without an agenda, group meetings are likely to dissolve into chaos (Welch, 1994). Rules of order, such as *Robert's Rules of Order* (Robert, 1990), play a similarly important role, by regulating the conduct of groups to ensure fair participation by all group members, including absentees. Rules of order also specify voting rules and means of determining consensus. Decision rules may require unanimity, plurality, or majority vote for an alternative.

Attaining consensus poses an advantage over voting, because voting encourages the development of coalitions, by posing the decision as a win–lose proposition (Ellis and Fisher, 1994). Members of the group who voted against an alternative are often unlikely to support it. Voting procedures can also play an important role (Davis, 1992).

### 3.4.2  Idea Generation Techniques

A variety of approaches have been developed for improving the creativity of groups in the early stages of decision making. Brainstorming is a popular technique for quickly generating ideas (Osborn, 1937). In this approach, a small group (of no more than 10 people) is given a problem to solve. The members are asked to generate as many ideas as possible. Members are told that no idea is too wild and are encouraged to build on the ideas submitted by others. No evaluation or criticism of the ideas is allowed until after the brainstorming session is finished. Buzz group analysis is a similar approach, more appropriate for large groups (Ellis and Fisher, 1994). Here, a large group is first divided into small groups of four to six members. Each small group goes through a brainstorming-like process to generate ideas. They then present their best ideas to the entire group for discussion. Other commonly applied idea-generating techniques include focus group analysis and group exercises intended to inspire creative thinking through role playing (Ellis and Fisher, 1994; Clemen, 1996).

The use of brainstorming and the other idea-generating methods mentioned above will normally provide a substantial amount of, in some cases, creative suggestions, especially when participants build on each other's ideas. However, personality factors and group dynamics can also lead to undesirable results. Simply put, some people are much more willing than others to participate in such exercises. Group discussions consequently tend to center around the ideas put forth by certain more forceful individuals. Group norms, such as deferring to participants with higher status and power, may also lead to undue emphasis on the opinions of certain members.

### 3.4.3  Nominal Group and Delphi Technique

Nominal group technique (NGT) and the Delphi technique attempt to alleviate some of the disadvantages of working in groups (Delbecq et al., 1975). The nominal group technique consists of asking each member of a group to write down and think about his or her ideas independently. A group moderator then asks each member to present one or more of his or her ideas. Once all

of the ideas have been posted, the moderator allows discussion to begin. After the discussion is finished, each participant rates or ranks the ideas presented. The subject ratings are then used to develop a score for each idea. Nominal group technique is intended to increase participation by group members and is based on the idea that people will be more comfortable presenting their ideas if they have a chance to think about them first (Delbecq et al., 1975).

The Delphi technique allows participants to comment anonymously, at their leisure, on proposals made by other group members. Normally, the participants do not know who proposed the ideas they are commenting on. The first step is to send an open-ended questionnaire to members of the group. The results are then used to generate a series of follow-up questionnaires in which more specific questions are asked. The anonymous nature of the Delphi process theoretically reduces the effect of participant status and power. Separating the participants also increases the chance that members will provide opinions "uncontaminated" by the opinions of others.

### 3.4.4 Structuring Group Decisions

As discussed earlier in this chapter, the field of decision analysis has devised several methods for organizing or structuring the decision-making process. The rational reflection model (Siebold, 1992) is a less formal, six-step procedure that serves a similar function. Group members are asked first to define and limit the problem by identifying goals, available resources, and procedural constraints. After defining and limiting the problem, the group is asked to analyze the problem, collect relevant information, and establish the criteria that a solution must meet. Potential solutions are then discussed in terms of the agreed-upon decision criteria. After further discussion, the group selects a solution and determines how it should be implemented. The focus of this approach is on forcing the group to confine its discussion to the issues that arise at each step in the decision-making process. As such, this method is similar to specifying an agenda.

Raiffa (1982) provides a somewhat more formal decision-analytic approach for structuring negotiations. The approach begins by assessing (1) the alternatives to a negotiated settlement, (2) the interests of the involved parties, and (3) the relative importance of each issue. This assessment allows the negotiators to think analytically about mutually acceptable solutions. In certain cases, a bargaining zone is available. For example, an employer may be willing to pay more than the minimum salary acceptable to a potential employee. In this case, the bargaining zone is the difference between the maximum salary the employer is willing to pay and the minimum salary a potential employee is willing to accept. The negotiator may also think about means of expanding the available resources to be divided, potential trading issues, or new options that satisfy the interests of the concerned parties.

Other methods for structuring group preferences are discussed in Keeney and Raiffa (1976). The development of group utility functions is one such approach.

A variety of computer-mediated methods for structuring group decisions are also available.

## 4 DECISION SUPPORT AND PROBLEM SOLVING

The preceding sections of this chapter have much to say about how to help decision makers make better decisions. To summarize that discussion briefly: (1) classical decision theory provides optimal prescriptions for how decisions should be made, (2) decision analysis provides a set of tools for structuring decisions and evaluating alternatives, and (3) studies of human judgment and decision making, in both laboratory settings and naturalistic environments, help identify the strengths and weaknesses of human decision makers. These topics directly mirror important elements of decision support. That is, decision support should have an objective (i.e., optimal or satisfactory choices, easier choices, more justifiable choices, etc.). Also, it must have a means (i.e., decision analysis or other method of decision support) and it must have a current state (i.e., decision quality, effort expended, knowledge, etc., of the supported decision makers). The effectiveness of decision support can then be defined in terms of how well the means move the current state toward the objective.

The focus of this section is on providing an overview of commonly used methods of computer-based decision support[*] for individuals, groups, and organizations. Throughout this discussion, an effort is made to address the objectives of each method of support and its effectiveness. Somewhat surprisingly, less information is available on the effectiveness of these approaches than might be expected given their prevalence (see also Yates et al., 2003), so the latter topic is not addressed in a lot of detail.

The discussion begins with a brief introduction to the field of decision analysis. Attention then shifts to the topics of decision support systems (DSSs), expert systems, and neural networks. These systems can be designed to support the intelligence, design, or choice phases of decision making (Simon, 1977). The intelligence phase involves scanning and searching the environment to identify problems or opportunities. The design phase entails formulating models for generating possible courses of action. The choice phase refers to finding an appropriate course of action for the problem

---

[*] Over the years, many different approaches have been developed for aiding or supporting decision makers (see von Winterfeldt and Edwards, 1986; Yates et al., 2003). Some of these approaches have already been covered earlier in this chapter and consequently are not addressed further in this section. In particular, decision analysis provides both tools and perspectives on how to structure a decision and evaluate alternatives. Decision analysis software is also available and commonly used. In fact, textbooks on decision analysis normally discuss the use of spreadsheets and other software; software may even be made available along with the textbook (e.g., see Clemen, 1996). Debiasing, discussed earlier in this chapter, is another technique for aiding or supporting decision makers.

or opportunity. Hence, the boundary between the design and choice phases is often unclear. Decision support systems and expert systems can be used to support all three phases of decision making, whereas neural networks tend to be better suited for design and choice phases. For example, DSSs can be designed to help with interpreting economic conditions, while expert systems can diagnose problems. Neural networks can learn a problem domain, after which they can serve as a powerful aid for decision making.

Attention then shifts to methods of supporting decisions by groups and organizations. The latter discussion first addresses the use by groups of DSSs and other tools similar to those used by individuals. In the sections that follow, we address approaches specifically designed for use by groups before briefly discussing the implications of problem-solving research for decision-making research.

## 4.1 Decision Analysis

The application of classical decision theory to improve human decision making is the goal of decision analysis (Howard, 1968, 1988; Raiffa, 1968; Keeney and Raiffa, 1976). Decision analysis requires inputs from decision makers, such as goals, preference and importance measures, and subjective probabilities. Elicitation techniques have consequently been developed that help decision makers provide these inputs. Particular focus has been placed on methods of quantifying preferences, trade-offs between conflicting objectives, and uncertainty (Raiffa, 1968; Keeney and Raiffa, 1976). As a first step in decision analysis, it is necessary to do some preliminary structuring of the decision, which then guides the elicitation process. The following discussion first presents methods of structuring decisions and then covers techniques for assessing subjective probabilities, utility functions, and preferences.

### 4.1.1 Structuring Decisions

The field of decision analysis has developed many useful frameworks for representing what is known about a decision (Howard, 1968; von Winterfeldt and Edwards, 1986; Clemen, 1996). In fact, these authors and others have stated that the process of structuring decisions is often the greatest contribution of going through the process of decision analysis. Among the many tools used, decision matrices and trees provide a convenient framework for comparing decisions on the basis of expected value or utility. Value trees provide a helpful method of structuring the sometimes complex relationships among objectives, attributes, goals, and values and are used extensively in multiattribute decision-making problems. Event trees, fault trees, inference trees, and influence diagrams are useful for describing probabilistic relationships between events and decisions. Each of these approaches is discussed briefly below.

***Decision Matrices and Trees***   Decision matrices are often used to represent single-stage decisions (Figure 4). The simplicity of decision matrices is their primary advantage. They also provide a very convenient



**Figure 4**   Decision matrix representation of a single-stage decision.
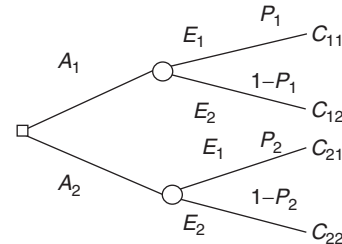


**Figure 5**   Decision tree representation of a single-stage decision.

format for applying the decision rules discussed in Section 2.1. Decision trees are also commonly used to represent single-stage decisions (Figure 5) and are particularly useful for describing multistage decisions (Raiffa, 1968). Note that in a multistage decision tree, the probabilities of later events are conditioned on the result of earlier events. This leads to the important insight that the results of earlier events provide information regarding future events.[*] Following this approach, decisions may be stated in conditional form. An optimal decision, for example, might be to do a market survey first, then market the product only if the survey is positive.

Analysis of a single- or multistage decision tree involves two basic steps, averaging out and folding back (Raiffa, 1968). These steps occur at chance and decision nodes, respectively.[†] *Averaging out* occurs when the expected value (or utility) at each chance node is calculated. In Figure 5 this corresponds to calculating the expected value of $A_1$ and $A_2$, respectively. *Folding back* refers to choosing the action with the greatest value expected at each decision node.

Decision trees thus provide a straightforward way of comparing alternatives in terms of expected value or SEU. However, their development requires significant simplification of most decisions and the provision of numbers, such as measures of preference and subjective probabilities, that decision makers may have difficulty determining. In certain contexts, decision makers struggling with this issue may find it helpful to develop value trees, event trees, or influence diagrams, as expanded on below.

---

[*] For example, the first event in a decision tree might be the result of a test. The test result then provides information useful in making the final decision.

[†] Note that the standard convention uses circles to denote chance nodes and squares to denote decision nodes (Raiffa, 1968).
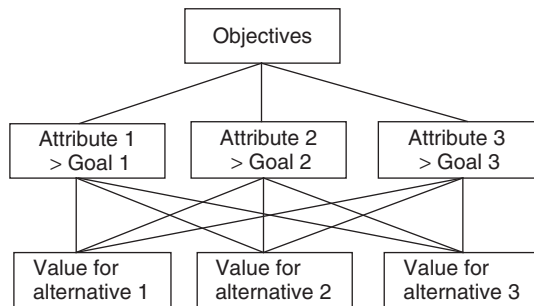
**Figure 6**   Generic value tree.



**Figure 7**   HRA event tree. (Adapted from Gertman and Blackman, 1994.)

***Value Trees***   Value trees hierarchically organize objectives, attributes, goals, and values (Figure 6). From this perspective, an objective corresponds to satisficing or maximizing a goal or set of goals. When there is more than one goal, the decision maker will have multiple objectives, which may differ in importance. Objectives and goals are both measured on a set of attributes. Attributes may provide (1) objective measures of a goal, such as when fatalities and injuries are used as a measure of highway safety; (2) subjective measures of a goal, such as when people are asked to rate the quality of life in the suburbs versus the city; or (3) proxy or indirect measures of a goal, such as when the quality of ambulance service is measured in terms of response time.

In generating objectives and attributes, it becomes important to consider their relevance, completeness, and independence. Desirable properties of attributes (Keeney and Raiffa, 1976) include:

1. *Completeness*: the extent to which the attributes measure whether an objective is met
2. *Operationality*: the degree to which the attributes are meaningful and feasible to measure
3. *Decomposability*: whether the whole is described by its parts
4. *Nonredundancy*: the fact that correlated attributes give misleading results
5. *Minimum size*: the fact that considering irrelevant attributes is expensive and may be misleading

Once a value tree has been generated, various methods can be used to assess preferences directly between the alternatives.

***Event Trees or Networks***   Event trees or networks show how a sequence of events can lead from primary events to one or more outcomes. Human reliability analysis (HRA) event trees are a classic example of this approach (Figure 7). If probabilities are attached to the primary events, it becomes possible to calculate the probability of outcomes, as illustrated in Section 4.1.2. This approach has been used in the field of risk assessment to estimate the reliability of human operators and other elements of complex systems (Gertman and Blackman, 1994).
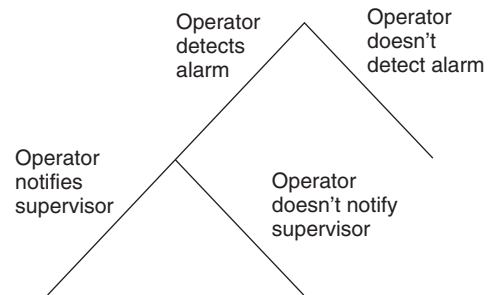
Fault trees work backward from a single undesired event to its causes (Figure 8). Fault trees are commonly used in risk assessment to help infer the chance of an accident occurring (Hammer, 1993; Gertman and Blackman, 1994). Inference trees relate a set of hypotheses at the top level of the tree to evidence depicted at the lower levels. The latter approach has been used by expert systems such as Prospector (Duda et al., 1979). Prospector applies a Bayesian approach to infer the presence of a mineral deposit from uncertain evidence.

***Influence Diagrams and Cognitive Mapping***
Influence diagrams are often used in the early stages of a decision to show how events and actions are related. Their use in the early stages of a decision is referred to as *knowledge* (or *cognitive*) *mapping* (Howard, 1988). Links in an inference diagram depict causal and temporal relations between events and decision stages.[*] A link leading from event $A$ to event $B$ implies that the probability of obtaining event $B$ depends on whether event $A$ has occurred. A link leading from a decision to an event implies that the probability of the event depends on the choice made at that decision stage. A link leading from an event to a decision implies that the decision maker knows the outcome of the event at the time the decision is made.

One advantage of influence diagrams in comparison to decision trees is that influence diagrams show the relationships between events more explicitly. Consequently, influence diagrams are often used to represent complicated decisions where events interactively influence the outcomes. For example, the influence diagram in Figure 9 shows that the true state of the machine affects both the probability of the warning signal and the consequence of the operator's decision. This linkage would be hidden within a decision tree.[†] Influence

---

[*] As for decision trees, the convention for influence diagrams is to depict events with circles and decisions with squares.
[†] The conditional probabilities in a decision tree would reflect this linkage, but the structure of the tree itself does not show the linkage directly. Also, the decision tree would use the flipped probability tree using $P$(warning) at the first stage and $P$(machine down|warning) at the second stage. It seems more natural for operators to think about the problem in terms of $P$(machine down) and $P$(warning|machine down), which is the way the influence diagram in Figure 8 depicts the relationship.
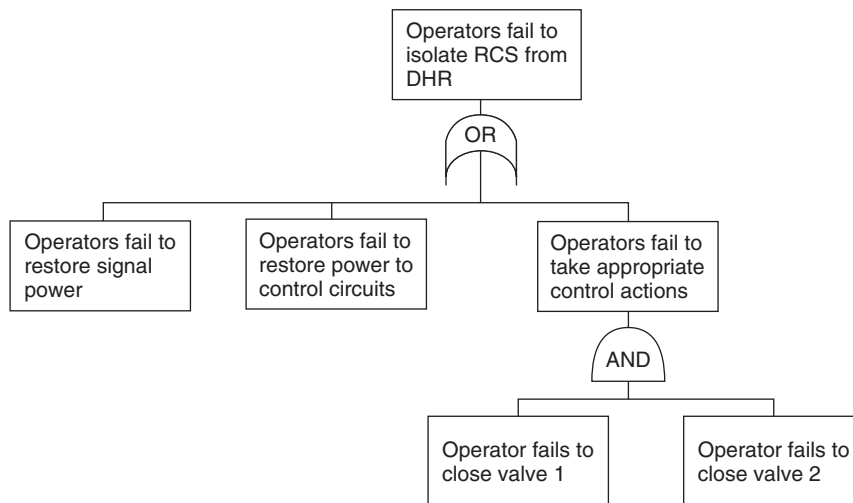
**Figure 8** Fault tree for operators. (Adapted from Gertman and Blackman, 1994.)
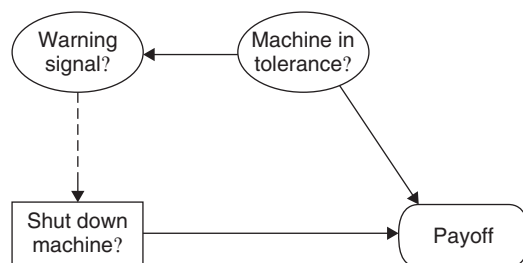


**Figure 9** Influence diagram representation of a single-stage decision.



**Figure 10** Standard gamble used in the variable probability method of eliciting utility functions.
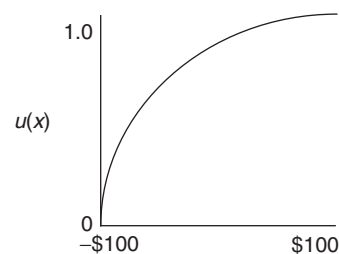


**Figure 11** Typical utility function.

diagrams have been used to structure medical decision-making problems (Holtzman, 1989) and are emphasized in modern texts on decision analysis (Clemen, 1996). Howard (1988) states that influence diagrams are the greatest advance he has seen in the communication, elicitation, and detailed representation of human knowledge. Part of the issue is that influence diagrams allow people who do not have deep knowledge of probability to describe complex conditional relationships with simple linkages between events. Once these linkages are defined, the decision becomes well defined and can be formally analyzed.

### 4.1.2 Utility Function Assessment

Standard methods for assessing utility functions (Raiffa, 1968) include (1) the variable probability method and (2) the certainty equivalent method. In the variable probability method, the decision maker is asked to give the value for the probability of winning at which they are indifferent between a gamble and a certain outcome (Figure 10). A utility function is then mapped out when the value of the certainty equivalent (CE) is changed over the range of outcomes. Returning to Figure 10, the value of $P$ at which the decision maker is indifferent

between the gamble and the certain loss of \$50 gives the value for $u(-\$50)$. In the utility function in Figure 11, the decision maker gave a value of about 0.5 in response to this question.

The certainty equivalent method uses lotteries in a similar way. The major change is that the probability of winning or losing the lottery is held constant while the amount won or lost is changed. In most cases the lottery provides an equal chance of winning and losing. The method begins by asking the decision maker to give a certainty equivalent for the original lottery ($CE_1$). The value chosen has a utility of 0.5. This follows since the utility of the best outcome is assigned a value of 1 and

the worst is given a utility of 0. The utility of the original gamble is therefore

$$u(\text{CE}_1) = pu(\text{best}) + (1-p)u(\text{worst})$$

$$= p(1) + (1-p)(0) = p = 0.5 \qquad (8)$$

The decision maker is then asked to give certainty equivalents for two new lotteries. Each uses the CE from the previous lottery as one of the potential prizes. The other prizes used in the two lotteries are the best and worst outcomes from the original lottery, respectively. The utility of the certainty equivalent $(\text{CE}_2)$ for the lottery using the best outcome and $\text{CE}_1$ is given by

$$u(\text{CE}_2) = pu(\text{best}) + (1-p)u(\text{CE}_1)$$

$$= p(1) + (1-p)(0.5) = 0.75 \qquad (9)$$

The utility of the certainty equivalent $(\text{CE}_3)$ given for the lottery using the worst outcome and $\text{CE}_1$ is given by

$$u(\text{CE}_3) = pu(\text{CE}_1) + (1-p)u(\text{worst})$$

$$= p(0.5) + (1-p)(0) = 0.25 \qquad (10)$$

This process is continued until the utility function is specified in sufficient detail. A problem with the certainty equivalent method is that errors are compounded as the analysis proceeds. This follows since the utility assigned in the first preference assessment [i.e., $u(\text{CE}_1)$] is used throughout the subsequent preference assessments. A second issue is that the CE method uses different ranges in the indifference lotteries, meaning that the CEs are compared against different reference values. This might create inconsistencies since, as discussed in Section 2.2, attitudes toward risk usually change depending on whether outcomes are viewed as gains or losses. The use of different reference points may, of course, cause the same outcome to be viewed as either a loss or a gain. Utilities may also vary over time. In Section 2.2.2, we discussed some of these issues further.

### 4.1.3 Preference Assessment

Methods for measuring strength of preference include indifference methods, direct assessment, and indirect measurement (Keeney and Raiffa, 1976; von Winterfeldt and Edwards, 1986). Indifference methods modify one of two sets of stimuli until subjects feel that they are indifferent between the two. Direct-assessment methods ask subjects to rate or otherwise assign numerical values to attributes, which are then used to obtain preferences for alternatives. Indirect-measurement techniques avoid decomposition and simply ask for preference orderings between alternatives. There has been some movement toward evaluating the effectiveness of particular methods for measuring preferences (Huber et al., 1993; Birnbaum et al., 1992). Each of these approaches are expanded upon below, and examples are given illustrating how they can be used.

***Indifference Methods***   Indifference methods are illustrated by the variable probability and certainty equivalent methods of eliciting utility functions presented in Section 2.1. There, indifference points were obtained by varying either probabilities or values of outcomes. Similar approaches have been applied to develop multiattribute utility or value functions. This approach involves four steps: (1) develop the single-attribute utility or value functions, (2) assume a functional form for the multiattribute function, (3) assess the indifference point between various multiattribute alternatives, and (4) calculate the substitution rate or relative importance of one attribute compared to the other. The single-attribute functions might be developed by indifference methods (i.e., the variable probability or certainty equivalent methods) or direct-assessment methods, as discussed later. Indifference points between multiattribute outcomes are obtained through an interactive process in which the values of attributes are increased or decreased systematically. Substitution rates are then obtained from the indifference points.

For example, consider the case for two alternative traffic safety policies, $A_1$ and $A_2$. Each policy has two attributes, $x =$ lives lost and $y =$ money spent. Assume that the decision maker is indifferent between $A_1$ and $A_2$, meaning the decision maker feels that $v(x_1, y_1) = v(20{,}000 \text{ deaths}; \$1 \text{ trillion})$ is equivalent to $v(x_2, y_2) = v(10{,}000 \text{ deaths}; \$1.5 \text{ trillion})$. For the sake of simplicity, assume an additive value function, where $v(x, y) = kv_x(x) + (1-k)v_y(y)$. Given this functional form, the indifference point $A_1 = A_2$ is used to derive the relation

$$(1-k)kv_x\,(20{,}000 \text{ deaths}) + kv_y\,(\$1 \times 10^{12})$$

$$= (1-k)v_x\,(10{,}000 \text{ deaths}) + kv_y\,(\$1.5 \times 10^{12}) \qquad (11)$$

This results in the substitution rate

$$\frac{k}{1-k} = \frac{v_x(20{,}000 \text{ deaths}) - v_x(10{,}000 \text{ deaths})}{v_y(\$1.5 \times 10^{12}) - v_y(\$1 \times 10^{12})} \qquad (12)$$

If $v_x = -x$ and $v_y = -y$, a value of approximately $2^{-5}$ is obtained for $k$. The procedure becomes somewhat more complex when nonadditive forms are assumed for the multiattribute function (Keeney and Raiffa, 1976).

***Direct-Assessment Methods***   Direct-assessment methods include curve-fitting and various numerical rating methods (von Winterfeldt and Edwards, 1986). Curve fitting is perhaps the simplest approach. Here, the decision maker first orders the various attributes and then simply draws a curve assigning values to them. For example, an expert might draw a curve relating levels of traffic noise (measured in decibels) to their level of annoyance (on a scale of $0-1$). Rating methods, as discussed earlier in reference to subjective probability assessment, include direct numerical measures on rating scales and relative ratings.

The analytic hierarchy process (AHP) provides one of the more implementable methods of this type (Saaty,

1990). In this approach, the decision is first structured as a value tree (Figure 6). Then each of the attributes is compared in terms of importance in a pairwise rating process. When entering the ratings, decision makers can enter numerical ratios (e.g., an attribute might be twice as important as another) or use the subjective verbal anchors mentioned earlier in reference to subjective probability assessment. The AHP program uses the ratings to calculate a normalized eigenvector assigning importance or preference weights to each attribute. Each alternative is then compared on the separate attributes. For example, two houses might first be compared in terms of cost and then be compared in terms of attractiveness. This results in another eigenvector describing how well each alternative satisfies each attribute. These two sets of eigenvectors are then combined into a single vector that orders alternatives in terms of preference. The subjective multiattribute rating technique (Smart) developed by Edwards (see von Winterfeldt and Edwards, 1986) provides a similar, easily implemented approach. Both techniques are computerized, making the assessment process relatively painless.

Direct-assessment approaches have been relied on extensively by product designers who would like to improve perceived product quality. For example, the automotive industry in the United States has for many years made extensive use of a structured approach for improving product quality called quality function deployment (QFD). QFD is defined as "converting the consumers' demands into quality characteristics and developing a design quality for the finished product by systematically deploying the relationships between the demands and the characteristics, starting with the quality of each functional component and extending the deployment to the quality of each part and process" (Akao, 2004, p. 5). The House of Quality in Figure 12 succinctly represents how QFD works though QFD is not a mere diagram, but more likely a quality improvement procedure. The matrix located at the center of Figure 12 combines the customer attributes (or consumers' demands) as rows and engineering characteristics as columns. By associating these two sets of attributes using a matrix, it is possible to identify which characteristics in a product should be improved in order to increase consumer satisfaction. On the top of the engineering characteristic, another diagonal matrix, which resembles a roof top of a house, shows the relationships between different engineering characteristics. Sometimes, two engineering characteristics can be improved together, but other times two engineering characteristics conflict (e.g., power vs. fuel efficiency of an automobile). Understanding the relationship between engineering characteristics could help informed decision making.

In order to capture the customer attributes (or consumers' needs), various methods have been used. First, the most important segment of users is usually defined, and analyzing existing customer databases and conducting a market survey using focus group, interview, and survey studies. As showing in Figure 12, consumers' needs are often described in their own words, such as "Easy to close from outside." These descriptions could be fuzzy, and fuzzy set theory could help quantify fuzziness, so that further quantitative analysis could be done (Lehto and Buck, 2007).

More recently, Web technologies have been used to efficiently collect customer ratings collectively, which is called collective filtering (e.g., Rashid et al., 2002; Schafer et al., 2001). These techniques have been actively applied to e-commerce websites, such as amazon.com and netflix.com.

***Indirect Measurement*** Indirect-measurement techniques avoid asking people to rate or rank directly the importance of factors that affect their preferences. Instead, subjects simply state or order their preferences for different alternatives. A variety of approaches can then be used to determine how individual factors influence preference.

Conjoint analysis provides one such approach for separating the effects of multiple factors when only their joint effects are known. Conjoint analysis is "a technique for measuring trade-offs for analyzing survey responses concerning preferences and intentions to buy, and it is a method for simulating how consumers might react to changes in current products or to new products introduced into an existing competitive array" (Green et al., 2001, p. S57). It has been successful in both academia and industry (Green et al., 2001) to understand preferences and intentions of consumers. For example, Marriott's Courtyard Hotels (Wind et al., 1989) and New York EZ-Pass system (Vavra et al., 1999) were designed using conjoint analysis approaches, both of which illustrated the utility of conjoint analysis.

There are different types of conjoint analysis. One of the simplest types is full-profile studies. In a full-profile study, profiles of a product (e.g., home) having relatively small number of attributes (four to five) are shown to a survey respondent, so that he or she can sort or rate the profiles based on attributes. The order or rating scores of profiles is used to investigate which attributes are more likely to influence the respondent's decision. Figure 13 shows examples of profiles.

Since each attribute could have multiple levels and profiles should cover all combinations comprehensively, even with a small number of attributes, the number of profiles that a respondent needs to compare with could be exponentially increased. For example, if each of five attributes has three levels, the number of profiles becomes 243 ($= 3^5$). This combinatorial nature restricts the number of attributes in full-profile studies. Common solutions to this problem are to use a partial set of profiles (Green and Krieger, 1990) or to ask a respondent what are more important attributes or which levels of attributes are more desired to decrease the number of profiles (Green and Krieger, 1987). More recent advances include adaptive conjoint analysis (Huber and Zwerina, 1996) and fast polyhedral adaptive conjoint estimation (Toubia et al., 2003), both of which cut down the number of questions in a conjoint analysis adaptively using respondent's responses. Related applications include the dichotomy-cut method, used to obtain decision rules for individuals and groups from ordinal rankings of multiattribute alternatives (Stanoulov, 1994).

However, in spite of its success and evolution of over three decades, conjoint analysis still has room

**Figure 12** Example House of Quality.

<table>
<tr><td>

**Card 1**

Annual Price: $20

Cash rebate: none

Retail purchase insurance: none

Rental car insurance: $30,000

Baggage insurance: $25,000

Airport club admission: $2 per visit

Medical–legal: no

Airport limousine: not offered

</td><td>

**Card 2**

Annual Price: $50

Cash rebate: 0.5%

Retail purchase insurance: none

Rental car insurance: $30,000

Baggage insurance: None

Airport club admission: $5 per visit

Medical–legal: yes

Airport limousine: 20% discount

</td></tr>
</table>

**Figure 13** Profile cards describe services that a credit card could offer. (Adapted from Green et al., 2001.)

for improvement (Bradlow, 2005). Respondents may change their preference structure, but this aspect has not been systematically considered in many conjoint analysis studies. Conjoint analysis often burdens research participants by asking too many questions though various techniques (e.g., adaptive conjoint analysis and choice-based conjoint analysis) have been suggested to cut down the number of questions to be answered. Findings in behavioral research have not been fully reflected in conjoint analysis, yet. In other words, consumers actually use various heuristics and strategies to cut down the number of candidates (or profiles in the context of conjoint analysis), but many studies in conjoint analysis do not accommodate these aspects.

The *policy-capturing approach* used in social judgment theory (Hammond et al., 1975; Hammond, 1993) is another indirect approach for describing human judgments of both preferences and probability. The policy-capturing approach uses multivariate regression or similar techniques to relate preferences to attributes for one or more decision makers. The equations obtained correspond to policies followed by particular decision makers. An example equation might relate medical symptoms to a physician's diagnosis. It has been argued that the policy-capturing approach measures the influence of factors on human judgments more accurately than do decomposition methods. Captured weights might be more accurate because decision makers may have little insight into the factors that affect their judgments (Valenzi and Andrews, 1973). People may also weigh certain factors in ways that reflect social desirability rather than influence on their judgments (Brookhouse et al., 1986). For example, people comparing jobs might rate pay as being lower in importance than intellectual challenge, whereas their preferences between jobs might be predicted entirely by pay. Caution must also be taken when interpreting regression weights as indicating importance, since regression coefficients are influenced by correlations between factors, their variability, and their validity (Stevenson et al., 1993).

There has been some movement toward evaluating the effectiveness of particular methods for measuring preferences (Birnbaum et al., 1992; Huber et al., 1993). However, the validity of direct versus indirect assessment is one area of continuing controversy. One conclusion that might be drawn is that it is not clear that any of the quantitative methods described above have adequate descriptors, factors, and methods to account for the dynamic characteristics (e.g., emerging consumer knowledge and reactions between competing opinions) of complex issues, such as making decisions about energy sources and consumption patterns responding to climate change and environmental concerns.

## 4.2 Individual Decision Support

The concept of DSSs dates back to the early 1970s. It was first articulated by Little (1970) under the term *decision calculus* and by Scott-Morton (1977) under the term *management decision systems*. DSSs are interactive computer-based systems that help decision makers utilize *data* and *models* to solve unstructured or semistructured problems (Scott-Morton, 1977; Keen and Scott-Morton, 1978). Given the unstructured nature of these problems, the goal of such systems is to *support*, rather than replace, human decision making.

The three key components of a DSS are (1) a model base, (2) a database, and (3) a user interface. The model base comprises quantitative models (e.g., financial or statistical models) that provide the analysis capabilities of DSSs. The database manages and organizes the data in meaningful formats that can be extracted or queried. The user interface component manages the dialogue or interface between the DSS and the users. For example, visualization tools can be used to facilitate communication between the DSS and the users.

DSSs are generally classified into two types: model driven and data driven. *Model-driven DSSs* utilize a collection of mathematical and analytical models for the decision analysis. Examples include forecasting and planning models, optimization models, and sensitivity analysis models (i.e., for asking "what-if" questions). The analytical capabilities of such systems are powerful because they are based on strong theories or models. On the other hand, *data-driven DSSs* are capable of analyzing large quantities of data to extract useful information. The data may be derived from transaction-processing systems, enterprise systems, data warehouses, or Web

warehouses. Online analytical processing and data mining can be used to analyze the data. Multidimensional data analysis enables users to view the same data in different ways using multiple dimensions. The dimensions could be product, salesperson, price, region, and time period. *Data mining* refers to a variety of techniques that can be used to find hidden patterns and relationships in large databases and to infer rules from them to guide decision making and predict future behavior. Data mining can yield information on associations, sequences, classifications, clusters, and forecasts (Laudon and Laudon, 2003). *Associations* are occurrences linked to a single event (e.g., beer is purchased along with diapers); *sequences* are events linked over time (e.g., the purchase of a new oven after the purchase of a house). *Classifications* refer to recognizing patterns and rules to categorize an item or object into its predefined group (e.g., customers who are likely to default on loans); *clustering* refers to categorizing items or objects into groups that have yet been defined (e.g., identifying customers with similar preferences). Data mining can also be used for forecasting (e.g., projecting sales demand).

Despite the popularity of DSSs not a lot of data are available documenting that they improve decision making (Yates et al., 2003). It does seem logical that DSSs should play a useful role in reducing biases (see Section 2.2.2) and otherwise improving decision quality. This follows because a well-designed DSS will increase both the amount and quality of information available to the decision maker. A well-designed DSS will also make it easier to analyze the information with sophisticated modeling techniques. Ease of use is another important consideration. As discussed earlier, Payne et al. (1993) identify two factors influencing the selection of a decision strategy: (1) cognitive effort required of a strategy in making the decision and (2) the accuracy of the strategy in yielding a "good" decision. Todd and Benbasat (1991, 1992) found that DSS users adapted their strategy selection to the type of decision aids available in such a way as to reduce effort. In other words, effort minimization is a primary or more important consideration to DSS users than is the quality of decisions. More specifically, the role of effort may have a direct impact on DSS effectiveness and must be taken into account in the design of DSSs.

In a follow-up study, Todd and Benbasat (1999) studied the moderating effect of incentives and cognitive effort required to utilize a more effortful decision strategy that would lead to a better decision outcome (i.e., additive compensatory vs. elimination by aspects; the former strategy requires more effort but leads to a better outcome). Although the results show that the level of incentives has no effect on decision strategy, the additive compensatory (i.e., 'better') strategy was used more frequently when its level of support was increased from no or little support to moderate or high support. The increased support decreased the amount of effort needed to utilize the additive compensatory strategy, thus inducing a strategy change. When designing DSSs, effort minimization should be given considerable attention, as it can drive the choice of decision strategy, which in turn influences the decision accuracy.

### 4.2.1 Expert Systems

Expert systems are developed to capture knowledge for a very specific and limited domain of human expertise. Expert systems can provide the following benefits: cost reduction, increased output, improved quality, consistency of employee output, reduced downtime, captured scarce expertise, flexibility in providing services, easier operation of equipment, increased reliability, faster response, ability to work with incomplete and uncertain information, improved training, increased ability to solve complex problems, and better use of expert time.

Organizations routinely use expert systems to enhance the productivity and skill of human knowledge workers across a spectrum of business and professional domains. They are computer programs capable of performing specialized tasks based on an understanding of how human experts perform the same tasks. They typically operate in narrowly defined task domains. Despite the name *expert systems*, few of these systems are targeted at replacing their human counterparts; *most of them are designed to function as assistants or advisers to human decision makers*. Indeed, the most successful expert systems—those that actually address mission-critical business problems—are not "experts" as much as "advisors" (LaPlante, 1990).

An expert system is organized in such a way that the knowledge about the problem domain is separated from general problem-solving knowledge. The collection of domain knowledge is called the *knowledge base*, whereas the general problem-solving knowledge is called the *inference engine*. The knowledge base stores domain-specific knowledge in the form of facts and rules. The inference engine operates on the knowledge base by performing logical inferences and deducing new knowledge when it applies rules to facts. Expert systems are also capable of providing explanations to users.

Examples of expert systems include the Plan Power system used by the Financial Collaborative for financial planning (Sviokla, 1989), Digital's XCON for computer configurations (Sviokla, 1990), and Baroid's MUDMAN for drilling decisions (Sviokla, 1986). As pointed out by Yates et al. (2003), the large number of expert systems that are now in actual use suggests that expert systems are by far the most popular form of computer-based decision support. However, as for DSSs, not a lot of data are available showing that expert systems improve decision quality. Ease of use is probably one of the main reasons for their popularity. This follows, because the user of an expert system can take a relatively passive role in the problem-solving process. That is, the expert system asks a series of questions which the user simply answers if he or she can. The ability of most expert systems to answer questions and explain their reasoning can also help users understand what the system is doing and confirm the validity of the system's recommendations. Such give and take may make users more comfortable with an expert system than they are with models that make sophisticated mathematical calculations that are difficult to verify.

### 4.2.2 Neural Networks

Neural networks consist of hardware or software that is designed to emulate the processing patterns of the biological brain. There are eight components in a neural network (Rumelhart et al., 1986):

1. A set of *processing units*
2. A *state of activation*
3. An *output function* for each unit
4. A *pattern of connectivity* among units
5. A *propagation rule* for propagating patterns of activities through the network
6. An *activation rule* for combining inputs impinging on a unit with its current state
7. A *learning* (or training) rule to modify patterns of connectivity
8. An *environment* within which the system must operate

A neural network comprises many interconnected processing elements that operate in parallel. One key characteristic of neural networks is their ability to learn. There are two types of learning algorithms: supervised learning and unsupervised learning. In supervised learning, the desired outputs for each set of inputs are known. Hence, the neural network learns by adjusting its weights in such a way that it minimizes the difference between the desired and actual outputs. Examples of supervised learning algorithms are back-propagation and the Hopfield network. Unsupervised learning is similar to cluster analysis in that only input stimuli are available. The neural network self-organizes itself to produce clusters or categories. Examples of unsupervised learning algorithms are adaptive resonance theory and Kohenen self-organizing feature maps.

By applying a training set such as historical cases, learning algorithms can be used to teach a neural network to solve or analyze problems. The outputs or recommendations from the system can be used to support human decision making. For example, neural networks have been developed to predict customer responses to direct marketing (Cui and Wong, 2004), to forecast stock returns (Olson and Mossman, 2003; Sapena et al., 2003; Jasic and Wood, 2004), to assess product quality in the metallurgical industry (Zhou and Xu, 1999), and to support decision making on sales forecasting (Kuo and Xue, 1998).

### 4.2.3 Visual Analytics

As the amount and complexity of available information ever grows, selecting and understanding relevant information become more and more challenging. For example, in the financial market, a decision maker should not only deal with numerous market indices for stock prices, bonds, futures, and so on, but also understand nonnumerical information about market trends and business rumors. Since such information is constantly created, and their historical aspects are also important, it is very challenging to keep up with such information

deluge and make an informed decision. Information deluge is actually a universal problem in other domains: physics and astronomy, environmental monitoring, disaster and emergency management, security, software analytics, biology, medicine, health, and personal information management (Keim et al., 2008b).

In order to deal with massive, complex, and heterogeneous information, attempts to utilize the highest bandwidth human sensory channel, *vision*, have been made. Thus, "visual analytics" has been recently proposed as a separate research field. A commonly accepted definition of visual analytics is "the science of analytical reasoning facilitated by interactive visual interfaces" (Thomas and Cook, 2005, p. 4). More specifically, Keim et al. (2008a, p. 4) detailed the goal of visual analytics as follows:

- Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data.
- Detect the expected and discover the unexpected.
- Provide timely, defensible, and understandable assessments.
- Communicate assessment effectively for action.

Obviously, visual analytics is largely overlapped with many disciplines, such as information visualization, human factors, data mining and management, decision making, and statistical analysis, to name a few. Keim et al. (2008b) especially pointed out the crucial role of human factors to understand interaction, cognition, perception, collaboration, presentation, and dissemination issues in employing visual analytics.

There have been some early successes in this endeavor. Jigsaw (Stasko et al., 2008) and IN-SPIRE (http://in-spire.pnl.gov/) are visual analytics tools to support investigative analysis on text data, such as investigative reports for potential terrorists. VisAware (Livnat et al., 2005) was built to raise situation awareness in the context of network intrusion detection. Map of the Market (Wattenberg, 1999) and FinDEx (Keim et al., 2006) are visualization techniques to analyze the stock market and assets. Figure 14 shows a screen shot of Map of the Market, which shows increases and decreases of stock prices in color encoding and market capitalization in size encoding. It also supports details on demand through simple interaction (e.g., a user can drill down to a specific market segment by selecting a pop-up menu item).

In spite of these interesting and successful examples, researchers in visual analytics run into several challenges. Some of visual analytic tools utilize quite complex visualization techniques which may not be intuitively understood by non-visualization-savvy users. The lack of comprehensive guidelines on how to create intuitive visualization techniques has long been a problem. Evaluating visual analytic tools has also been challenging (Plaisant, 2004). Visual analytics tasks tend to require high-level expertise and are dynamic, complex, and uncertain. Hence, investigating the effectiveness of visual analytic tools is very time-consuming and ambiguous. These problems are being further aggravated
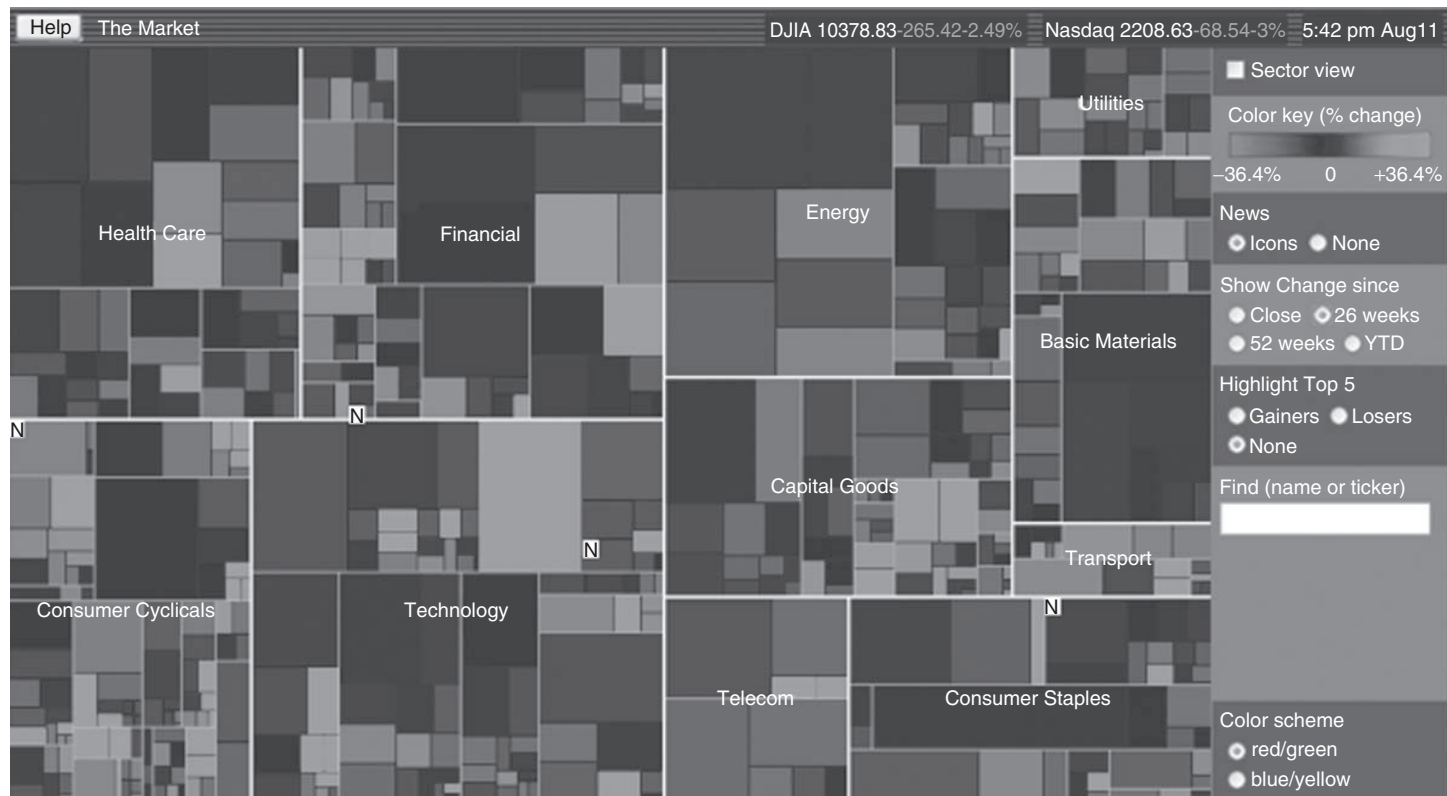
**Figure 14** Screen Shot of Map of the Market taken on July 15, 2010.

by an explosive growth of information. Though visual analytic approaches help users deal with larger data sets, the rapid growth in the volume of information is challenging to keep up. For a more comprehensive list of challenges, refer to Keim et al. (2008b), Thomas and Cook (2005), and Thomas and Kielman (2009).

### 4.2.4 Other Forms of Individual Decision Support

Other forms of individual decision support can be developed using fuzzy logic, intelligent agents, case-based reasoning, and genetic algorithms. *Fuzzy logic* refers to the use of membership functions to express imprecision and an approach to approximate reasoning in which the rules of inference are approximate rather than exact. Garavelli and Gorgoglione (1999) used fuzzy logic to design a DSS to improve its robustness under uncertainty, and Coma et al. (2004) developed a fuzzy DSS to support a design for assembly methodology. Collan and Liu (2003) combined fuzzy logic with agent technologies to develop a fuzzy agent–based DSS for capital budgeting. *Intelligent agents* use built-in or learned rules to make decisions. In a multiagent marketing DSS, the final solution is obtained through cooperative and competitive interactions among intelligent agents acting in a distributed mode (Aliev et al., 2000). Intelligent agents can also be used to provide real-time decision support on airport gate assignment (Lam et al., 2003). *Case-based reasoning*, which relies on past cases to derive at a decision, has been used by Lari (2003) to assist in making corrective and preventive actions for solving quality problems and by Belecheanu et al. (2003) to support decision making on new product development. *Genetic algorithms* are robust algorithms that can search through large spaces quickly by mimicking the Darwinian "survival of the fittest" law. They can be used to increase the effectiveness of simulation-based DSSs (Fazlollahi and Vahidov, 2001).

### 4.3 Group and Organizational Decision Support

Computer tools have been developed to assist in group and organizational decision making. Some of them implement the approaches discussed in Section 3. The spectrum of such tools ranges from traditional tools used in decision analysis, such as the analytic hierarchy process (Saaty, 1990; Basak and Saaty, 1993), to electronic meeting places or group DSSs (DeSanctis and Gallupe, 1987; Nunamaker et al., 1991), to negotiation support systems (Bui et al., 1990; Lim and Benbasat, 1993). We will discuss the use of individual decision support tools for group support, group DSSs, negotiation support systems, enterprise system support, and other forms of group and organizational support.

### 4.3.1 Using Individual Decision Support Tools for Group Support

Traditional single-user tools can be used to support groups in decision making. A survey by Satzinger and Olfman (1995) found that traditional single-user tools were perceived by groups to be more useful than group support tools. Sharda et al. (1988) assessed the effectiveness of a DSS for supporting business simulation games and found that groups with access to the DSS made significantly more effective decisions than their non-DSS counterparts. The DSS groups took more time to make their decisions than the non-DSS groups at the beginning of the experiment, but decision times converged in a later period. The DSS teams also exhibited a higher confidence level in their decisions than the non-DSS groups. Knowledge-based systems (or expert support systems) are effective in supporting group decision making, particularly so with novices than experts (Nah and Benbasat, 2004). Groups using the system also make better decisions than individuals provided with the same system (Nah et al., 1999). Hence, empirical findings have shown that traditional single-user tools can be effective in supporting group decision making.

### 4.3.2 Group Decision Support Systems

Group decision support systems (GDSSs) combine communication, computing, and decision support technologies to facilitate formulation and solution of unstructured problems by a group of people (DeSanctis and Gallupe, 1987). DeSanctis and Gallupe defined three levels of GDSS. Level 1 GDSSs provide technical features aimed at removing common communication barriers, such as large screens for instantaneous display of ideas, voting solicitation and compilation, anonymous input of ideas and preferences, and electronic message exchange among members. In other words, a level 1 GDSS is a communication medium only. Level 2 GDSSs provide decision modeling or group decision techniques aimed at reducing uncertainty and "noise" that occur in the group's decision process. These techniques include automated planning tools [e.g., project evaluation review technique (PERT), critical path method (CPM), Gantt], structured decision aids for the group process (e.g., automation of Delphi, nominal, or other idea-gathering and compilation techniques), and decision analytic aids for the task (e.g., statistical methods, social judgment models). Level 3 GDSSs are characterized by machine-induced group communication patterns and can include expert advice in the selecting and arranging of *rules* to be applied during a meeting. To date, there has been little research in level 3 GDSSs because of the difficulty and challenges in automating the process of group decision making.

GDSSs facilitate computer-mediated group decision making and provide several potential benefits (Brashers et al., 1994), including (1) enabling all participants to work simultaneously (e.g., they don't have to wait for their turn to speak, thus eliminating the need to compete for air time), (2) enabling participants to stay focused and be very productive in idea generation (i.e., eliminating production blocking caused by attending to others), (3) providing a more equal and potentially anonymous opportunity to be heard (i.e., reducing the negative effects caused by power distance), and (4) providing a more systematic and structured decision-making environment (i.e., facilitating a more linear

process and better control of the agenda). GDSSs also make it easier to control and manage conflict through the use of facilitators and convenient voting procedures.

The meta-analysis by Dennis et al. (1996) suggests that, in general, GDSSs improve decision quality, increase time to make decisions, and have no effect on participant satisfaction. They also found that larger groups provided with a GDSS had higher satisfaction and experienced greater improvement in performance than smaller groups with GSSs. The findings from McLeod's (1992) and Benbasat and Lim's (1993) meta-analyses show that GDSSs increase decision quality, time to reach decisions, and equality of participation but decrease consensus and satisfaction. To resolve inconsistencies in the GDSS literature (such as those relating to satisfaction), Dennis and his colleagues (Dennis et al., 2001; Dennis and Wixom, 2002) carried out further meta-analyses to test a fit–appropriation model and identify further moderators for these effects. The result shows that both fit (between GSS structures and task) and appropriation support (i.e., training, facilitation, and software restrictiveness) are necessary for GDSSs to yield an increased number of ideas generated, reduce the time taken for the task, and increase satisfaction of users (Dennis et al., 2001). The fit–appropriation profile is adapted from Zigurs and Buckland (1998).

Computer-supported collaborative systems provide features beyond GDSSs, such as project and calendar management, group authoring, audio and video conferencing, and group and organizational memory management. They facilitate collaborative work beyond simply decision making and are typically referred to as computer-supported collaborative work. These systems are particularly helpful for supporting group decision making in a distributed and asynchronous manner.

### 4.3.3 Negotiation Support Systems

Negotiation support systems (NSSs) are used to assist people in activities that are competitive or involve conflicts of interest. The need for negotiation can arise from differences in interest or in objectives or even from cognitive limitations. To understand and analyze a negotiation activity, eight elements must be taken into account (Holsapple et al., 1998): (1) the issue or matter of contention, (2) the set of participants involved, (3) participants' regions of acceptance, (4) participants' location (preference) within the region of acceptance, (5) strategies for negotiation (e.g., coalition), (6) participants' movements from one location to another, (7) rules of negotiation, and (8) assistance from an intervenor (e.g., mediator, arbitrator, or facilitator). NSSs should be designed with these eight components in mind by supporting these components.

The configuration of basic NSSs comprises two main components (Lim and Benbasat, 1993): (1) a DSS for each negotiating party and (2) an electronic linkage between these systems to enable electronic communication between the negotiators. Full-feature session-oriented NSSs should also offer group process structuring techniques, support for an intervenor, and documentation of the negotiation (Foroughi, 1998). Nego-Plan is an expert system shell that can be used

to represent negotiation issues and decompose negotiation goals to help analyze consequences of negotiation scenarios (Matwin et al., 1989; Holsapple and Whinston, 1996). A Web-based NSS called Inspire is used in teaching and training (Kersten and Noronha, 1999). Espinasse et al. (1997) developed a multiagent NSS architecture that can support a mediator in managing the negotiation process. To provide comprehensive negotiation support, NSSs should provide features of level 3 GDSSs, such as the ability to (1) perform analysis of conflict contingencies, (2) suggest appropriate process structuring formats or analytical models, (3) monitor the semantic content of electronic communications, (4) suggest settlements with high joint benefits, and (5) provide automatic mediation (Foroughi, 1998). In general, NSSs can support negotiation either by assisting participants or by serving as a participant (intervenor).

### 4.3.4 Enterprise Systems for Decision Support

Enterprise-wide support can be provided by enterprise systems (ESs) and executive support systems (ESSs) (Turban and Aronson, 2001). ESSs are designed to support top executives, whereas ESs can be designed to support top executives or to serve a wider community of users. ESSs are comprehensive support systems that go beyond flexible DSSs by providing communication capabilities, office automation tools, decision analysis support, advanced graphics and visualization capabilities, and access to external databases and information in order to facilitate business intelligence and environmental scanning. For example, intelligent agents can be used to assist in environmental scanning.

The ability to use ESs, also known as enterprise resource planning (ERP) systems, for decision support is made possible by data warehousing and online analytical processing. ESs integrate all the functions as well as the transaction processing and information needs of an organization. These systems can bring significant competitive advantage to organizations if they are integrated with supply chain management and customer relationship management systems, thus providing comprehensive information along the entire value chain to key decision makers and facilitating their planning and forecasting. Advanced planning and scheduling packages can be incorporated to help optimize production and ensure that the right materials are in the right warehouse at the right time to meet customers' demands (Turban and Aronson, 2001).

### 4.3.5 The Wisdom of Crowds

Surowiecki (2004) popularized the concept of the wisdom of crowds through his book *The Wisdom of Crowds: Why the Many Are Smarter Than the Few*, which argues that the aggregation of information or opinions of crowds could result in better decisions than those of expert individuals or groups. The example that Surowiecki used to open his book is Galton's surprising result at a weight-judging competition of a dressed ox at the annual show of the West of England Fat Stock and Poultry Exhibition (Galton, 1907). Galton analyzed the 787 collected votes for the competition and found

that the median of the votes was 1207 lb, which was just 9 lb off from the true value, 1198 lb, which showed the power of aggregated information. Similar evidence has been collected in many other cases, such as locating a lost submarine, predicting the winner in sports betting, and predicating the future in investigative organizations (Surowiecki, 2004).

However, simply collecting opinions of crowds does not guarantee a better decision. The wisdom of crowds breaks down in the following situations (Surowiecki, 2004): (1) when the crowd becomes homogeneous, crowds fail to collect information from diverse perspectives. (2) when an organization is too centralized or too divided, it fails to collect information from individual members who directly confront the situation, and the collected information cannot be communicated within the organization. (3) when individuals in the crowd imitate others' opinions or are emotionally influenced by others, only a few members in the crowd play as information collectors or decision makers. Thus, Surowiecki suggests that the wisdom of crowds functions properly when the crowd has the following characteristics: diversity of opinions, independence, decentralization, and a mechanism to aggregate information.

There have been efforts to construct functioning crowds more systematically. The prediction market (also known as information market, decision market, and event future) is a market to predict future events using a similar mechanism of the financial market (Wolfers and Zitzewitz, 2004). For example, a betting exchange, Tradesports.com, listed a security paying $100 if the head of the Defense Advanced Research Projects Agency (DARPA), Admiral John Poindexter, resigned by the end of August 2003. The price of the security reflected the possibility of the event, so it fluctuated as more information was collected. This prediction market provides a platform to collect information from individuals with proper incentives. Various studies also reported that these prediction markets are extremely accurate (Berg et al., 2008), and it has been actively applied to various areas, such as predicting influenza outbreaks (Holden, 2007).

### 4.3.6 Other Forms of Group and Organizational Decision Support

We have discussed how individual decision support tools, GDSSs, NSSs, ESSs, and ESs can facilitate and support group and organizational decision making. Other techniques drawn from the field of artificial intelligence, such as neural networks, expert systems, fuzzy logic, genetic algorithms, case-based reasoning, and intelligent agents, can also be used to enhance the decision support capabilities of these systems. It should also be noted that knowledge management practices can benefit groups and organizations by capitalizing on existing knowledge to create new knowledge, codifying existing knowledge in ways that are readily accessible to others, and facilitating knowledge sharing and distribution throughout an enterprise (Davenport and Prusak, 2000). Since knowledge is a key asset of organizations and is regarded as the only source of sustainable

competitive strength (Drucker, 1995), the use of technologies for knowledge management purposes is a high priority in most organizations. For example, knowledge repositories (e.g., intranets) can be created to facilitate knowledge sharing and distribution, focused knowledge environments (e.g., expert systems) can be developed to codify expert knowledge to support decision making, and knowledge work systems (e.g., computer-aided design, virtual reality simulation systems, and powerful investment workstations) can be used to facilitate knowledge creation. By making existing knowledge more available, these systems can help groups and organizations make more informed and better decisions.

### 4.4 Problem Solving

Though problem solving is so commonly used, defining it is not an easy task (Hunt, 1998). Problem solving can be "understood as the bridging of the gap between an initial state of affairs and a desired state where no predetermined operation or strategy is known to the individual" (Öllinger and Goel, 2010, p. 4), and most definitions of problem solving attempted so far include three core components: an initial state, a goal state, and paths between the two states (Mayer, 1983). The path is often unknown, so problem solving is largely an activity to search for the path. However, these descriptions and characterization do not clearly specify what problem solving is and is not. Problem solving deals with various topics, which include, but are not limited to, reading, writing, calculation, managerial problem solving, problem solving in electronics, game playing (e.g., chess), and problem solving for innovation and inventions (Sternberg and Frensch, 1991). Some problems are structured (e.g., Tower of Hanoi), but others are ill-structured (e.g., preparing good dinner for guests) (Reitman, 1964). Thus, we might need an even more inclusive definition of problem solving, as Anderson et al. (1985) suggested: any goal-directed sequence of cognitive operations.

As problem solving includes a wide spectrum of topics, clearly drawing the boundary between problem solving and decision making is almost meaningless. Though Simon et al. (1987) provided elegant separation of two fields of research (i.e., problem solving covers fixing agendas, settings goals, and designing actions while decision making covers evaluating and choosing), one can easily argue against the division. Virtually all decision-making activities could be problem solving since decision making is an activity from a state of not having a selection toward a state with a selection. Conversely, some activities of problem solving, choosing a path out of potential paths or generating alternatives, would be considered as activities of decision making. Thus, it would be more appropriate to see that decision making and problem solving are largely overlapped, and suggesting a theoretical distinction between problem solving and decision making may not be fruitful. Kepner and Tregoe (1976) even said that problem solving and decision making are often used interchangeably.

In spite of the fuzzy boundary between the two fields, decision making and problem solving have

distinctive lineages. While decision-making research has been largely led by economists, statisticians, and mathematicians until descriptive approaches become more prominent, problem solving has a relatively longer and distinctive history of research mainly done by psychologists (Simon et al., 1987). Due to this difference, researchers in problem solving introduced several interesting research methods.

An interesting and seminal contribution of research in problem solving in understanding human mind is the results of tight collaboration between psychologists and computer scientists. After seminal work done by Simon (1955), information-processing theory provided a foundation for this endeavor. The main contribution was the emphasis on detailed task analysis and the clear computation characterization of the problem space (Öllinger and Goel, 2010). Newell later published unified theory of cognition (UTC) (1994), which emphasized the importance of unified architecture that includes existing research results in psychology and cognitive science. UTC initiated the implementation and evolution of various cognitive architectures, such as SOAR (Laird et al., 1987), ACT-R (Anderson, 1993; Anderson and Lebiere, 1998), and EPIC (Meyer and Kieras, 1997a, 1997b). These cognitive architectures have expedited the research of problem solving and the human mind (Anderson and Douglass, 2001), and they are also applied to various areas (e.g., aviation, vehicle design, and human–computer interactions).

Another interesting approach is applying neuroimaging techniques, such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), to physiologically understand how people solve problems, especially ill-structured problems and insight problems. As briefly discussed, there are problems that do not have complete information in each component of problem solving (e.g., preparing a good meal for dinner and understanding creativity), which is called an ill-structured problem (Reitman, 1964). Goel (1995) argued that ill-structured problem solving involves different phases of problem solving (i.e., problem scoping, preliminary solutions, refinement, and detailing of solutions) and showed, using brain imaging techniques, that different parts of the brain are associated with these different types of computations required for these phases (Goel and Morrison, 2005). Insight problem solving has also fascinated many cognitive scientists and psychologists because of its interesting nature. Insight problems are often solved through very few steps in a path, but identifying the path turns out to be very difficult. The compound remote association (CRA) problem is an example of insight problems: "Each of the three words in (a) can form a compound word or two-word phrase with the solution word. The solution word can come before or after any of the problem words: boot, summer, ground" (Bowden et al., 2005, p. 324). The answer to this problem is "camp." Bowden et al. (2005) provided a neurophysiological account of the A-ha! moment using fMRI and EEG data while people solve CRA problems. They revealed that a certain region of the brain, the anterior superior temporal gyrus (aSTG), was highly activated right before experiencing

an epiphany (Lehrer, 2008). Neuroimaging approaches enable problem-solving researchers to take a close look at brain activity to unveil some of complicated and hidden cognitive activities while solving problems.

Over time, research in problem solving and decision making has largely overlapped, and approaches successfully employed in one domain are quickly adopted by the other. For example, information-processing theory has been one of important paradigms to driving the development of decision theory in the last half century (Payne and Bettman, 2004). Cognitive architectures, such as ACT-R, have been employed to understand biases and heuristics used in decision making (e.g., Altmann and Burns, 2005; Belavkin, 2006; Dickison and Taatgen, 2007). Neuroimaging techniques have also been employed to understand decision-making tasks (e.g., Ernst et al., 2002; Trepel et al., 2005). The interaction between the two research communities is expected to accelerate as the boundary of research questions is widened.

## 5  SUMMARY AND CONCLUSIONS

Beach (1993) discusses four revolutions in behavioral decision theory. The first took place when it was recognized that the evaluation of alternatives is seldom extensive. It is illustrated by use of the satisficing rule (Simon, 1955) and heuristics (Tversky and Kahneman, 1974; Gigerenzer et al., 1999) rather than optimizing. The second occurred when it was recognized that people choose between strategies to make decisions. It is marked by the development of contingency theory (Beach, 1990) and cognitive continuum theory (Hammond, 1980). The third is currently occurring. It involves the realization that people rarely make choices and instead rely on pre-learned procedures. This perspective is illustrated by the levels-of-processing approach (Rasmussen, 1983) and recognition-primed decisions (Klein, 1989). The fourth is just beginning. It involves recognizing that decision-making research must abandon a single-minded focus on the economic view of decision making and include approaches drawn from relevant developments and research in cognitive psychology, organizational behavior, and systems theory.

The discussion within this chapter parallels this view of decision making. The integrative model presented at the beginning of the chapter shows how the various approaches fit together as a whole. Each path through the model is distinguished by specific sources of conflict, the methods of conflict resolution followed, and the types of decision rules used to analyze the results of conflict resolution processes. The different paths through the model correspond to fundamentally different ways of making decisions, ranging from routine situation assessment-driven decisions to satisficing, analysis of single- and multiattribute expected utility, and even obtaining consensus of multiple decision makers in group contexts. Numerous other strategies and potential methods of decision support discussed in this chapter are also described by particular paths through the model.

This chapter goes beyond simply describing methods of decision making by pointing out reasons that people and groups may have difficulty making good decisions. These include cognitive limitations, inadequacies of various heuristics used, biases and inadequate knowledge of decision makers, and task-related factors such as risk, time pressure, and stress. The discussion also provides insight into the effectiveness of approaches for improving human decision making. The models of selective attention point to the value of providing only truly relevant information to decision makers. Irrelevant information might be considered simply because it is there, especially if it is highly salient. Methods of highlighting or emphasizing relevant information are therefore warranted. The models of selective information also indicate that methods of helping decision makers cope with working memory limitations will be of value. There also is reason to believe that providing feedback to decision makers in dynamic decision-making situations will be useful. Cognitive rather than outcome feedback is indicated as being particularly helpful when decision makers are learning. Training decision makers also seems to offer potentially large benefits. One reason for this conclusion is that the studies of naturalistic decision making revealed that most decisions are made on a routine, nonanalytical basis.

Studies of debiasing also partially support the potential benefits of training and feedback. On the other hand, the many failures to debias expert decision makers imply that decision aids, methods of persuasion, and other approaches intended to improve decision making are no panacea. Part of the problem is that people tend to start with preconceived notions about what they should do and show a tendency to seek out and bolster confirming evidence. Consequently, people may become overconfident with experience and develop strongly held beliefs that are difficult to modify, even if they are hard to defend rationally.

## REFERENCES

Ainslie, G. (1975), "Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control," *Psychological Bulletin*, Vol. 82, pp. 463–509.

Akao, Y. (2004), *Quality Function Deployment: Integrating Customer Requirements into Product Design*, Productivity Press, New York.

Aliev, R. A., Fazlollahi, B., and Vahidov, R. M. (2000), "Soft Computing Based Multi-Agent Marketing Decision Support System," *Journal of Intelligent and Fuzzy Systems*, Vol. 9, Nos. 1–2, pp. 1–9.

Allais, M. (1953), "Le comportement de l'homme rationel devant le risque: critique des postulateset axioms de l'école américaine," *Econometrica*, Vol. 21, pp. 503–546.

Altmann, E. M., and Burns, B. D. (2005), "Streak Biases in Decision Making: Data and a Memory Model," *Cognitive Systems Research*, Vol. 6, pp. 5–16.

Anderson, J. R. (1993), *Rules of the Mind*, Lawrence Erlbaum Associates, Mahwah, NJ.

Anderson, J. R., and Douglass, S. (2001), "Tower of Hanoi: Evidence for the Cost of Goal Retrieval," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 27, No. 6, pp. 1331–1346.

Anderson, J. R., and Lebiere, C. (1998), *The Atomic Components of Thought*, Lawrence Erlbaum Associates, Mahwah, NJ.

Anderson, J. R., Boyle, C. F., and Reiser, B. J. (1985), "Intelligent Tutoring Systems," *Science*, Vol. 228, No. 4698, pp. 456–462.

Anderson, N. H. (1981), *Foundations of Information Integration Theory*, Academic, New York.

Argote, L., Seabright, M. A., and Dyer, L. (1986), "Individual Versus Group: Use of Base-Rate and Individuating Information," *Organizational Behavior and Human Decision Making Processes*, Vol. 38, pp. 65–75.

Ariely, D. (2008), Predictably Irrational: *The Hidden Forces That Shape Our Decisions*, Harper, New York.

Arkes, H. R. (1981), "Impediments to Accurate Clinical Judgment and Possible Ways to Minimize Their Impact," *Journal of Consulting and Clinical Psychology*, Vol. 49, pp. 323–330.

Arkes, H. R., and Blumer, C. (1985), "The Psychology of Sunk Cost," *Organizational Behavior and Human Decision Processes*, Vol. 35, pp. 124–140.

Arkes, H. R., and Hutzel, L. (2000), "The Role of Probability of Success Estimates in the Sunk Cost Effect," *Journal of Behavioral Decision Making*, Vol. 13, No. 3, pp. 295–306.

Armor, D. A., and Taylor, S. E. (2002), "When Predictions Fail: The Dilemma of Unrealistic Optimism," in *Heuristics and Biases: The Psychology of Intuitive Judgement*, T. Gilovich, D. W. Griffin, and D. Kahneman, Eds., Cambridge University Press, Cambridge, pp. 334–347.

Balzer, W. K., Doherty, M. E., and O'Connor, R. O., Jr. (1989), "Effects of Cognitive Feedback on Performance," *Psychological Bulletin*, Vol. 106, pp. 41–43.

Barberis, N., and Thaler, R. (2003), "A Survey of Behavioral Finance," in *Financial Markets and Asset Pricing*, Vol. 1, G. M. Constantinides, M. Harris, and R. M. Stulz, Eds., Elsevier, Amsterdam, The Netherlands, pp. 1053–1128.

Bar-Hillel, M. (1973), "On the Subjective Probability of Compound Events," *Organizational Behavior and Human Performance*, Vol. 9, pp. 396–406.

Baron, J. (1985), *Rationality and Intelligence*, Cambridge University Press, Cambridge.

Baron, J. (1998), *Judgment Misguided: Intuition and Error in Public Decision Making*, Oxford University Press, New York.

Basak, I., and Saaty, T. (1993), "Group Decision Making Using the Analytic Hierarchy Process," *Mathematical and Computer Modeling*, Vol. 17, pp. 101–109.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. (2001), "Bad Is Stronger Than Good," *Review of General Psychology*, Vol. 5, No. 4, pp. 323–370.

Bazerman, M. (1998), *Judgment in Managerial Decision Making*, 4th ed., Wiley, New York.

Bazerman, M. H., and Neale, M. A. (1983), "Heuristics in Negotiation: Limitations to Effective Dispute Resolution," in *Negotiating in Organizations*, M. H. Bazerman and R. Lewicki, Eds., Sage, Beverly Hills, CA.

Beach, L. R. (1990), *Image Theory: Decision Making in Personal and Organizational Contexts*, Wiley, Chichester, West Sussex, England.

Beach, L. R. (1993), "Four Revolutions in Behavioral Decision Theory," in *Leadership Theory and Research*, M. M. Chemers and R. Ayman, Eds., Academic, San Diego, CA.

Beach, L. R., and Mitchell, T. R. (1978), "A Contingency Model for the Selection of Decision Strategies," *Academy of Management Journal*, Vol. 3, pp. 439–449.

Belavkin, R. V. (2006), "Towards a Theory of Decision-Making without Paradoxes," in *Proceedings of the Seventh International Conference on Cognitive Modeling*, Trieste, Italy, pp. 38–43.

Belecheanu, R., Pawar, K. S., Barson, R. J., and Bredehorst, B. (2003), "The Application of Case-Based Reasoning to Decision Support in New Product Development," *Integrated Manufacturing Systems*, Vol. 14, No. 1, pp. 36–45.

Bell, D. (1982), "Regret in Decision Making Under Uncertainty," *Operations Research*, Vol. 30, pp. 961–981.

Benbasat, I., and Lim, L. H. (1993), "The Effects of Group, Task, Context, Technology Variables on the Usefulness of Group Support Systems: A Meta-Analysis of Experimental Studies," *Small Group Research*, Vol. 24, No. 4, pp. 430–462.

Berg, J., Forsythe, R., Nelson, F., and Rietz, T. (2008), "Results from a Dozen Years of Election Futures Markets Research," in *Handbook of Experimental Economics Results*, Vol. 1, C. R. Plott and V. L. Smith, Eds., Elsevier, Amsterdam, The Netherlands, pp. 742–751.

Bergus, G. R., Levin, I. P., and Santee, A. S. (2002), "Presenting Risks and Benefits to Patients: The Effects of Information Order on Decision Making," *Journal of General Internal Medicine*, Vol. 17, No. 8, pp. 612–617.

Bernoulli, D. (1738), *Exposition of a New Theory of the Measurement of Risk*, Imperial Academy of Science, St. Petersburg, Russia.

Bettman, J. R., Luce, M. F., and Payne, J. W. (1998), "Constructive Consumer Choice Processes," *Journal of Consumer Research*, Vol. 25, No. 3, pp. 187–217.

Birnbaum, M. H., and Mellers, B. A. (1983), "Bayesian Inference: Combining Base Rates with Opinions of Sources Who Vary in Credibility," *Journal of Personality and Social Psychology*, Vol. 37, pp. 792–804.

Birnbaum, M. H., Coffey, G., Mellers, B. A., and Weiss, R. (1992), "Utility Measurement: Configural-Weight Theory and the Judge's Point of View," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 18, pp. 331–346.

Bock, R. D., and Jones, L. V. (1968), *The Measurement and Prediction of Judgment and Choice*, Holden-Day, San Francisco, CA.

Bolton, G. E., and Chatterjee, K. (1996), "Coalition Formation, Communication, and Coordination: An Exploratory Experiment," in *Wise Choices: Decisions, Games, and Negotiations*, R. J. Zeckhauser, R. C. Keeney, and J. K. Sebanies (Eds.), Harvard University Press, Boston, MA.

Bornstein, G., and Yaniv, I. (1998), "Individual and Group Behavior in the Ultimatum Game: Are Groups More Rational Players?" *Experimental Economics*, Vol. 1, pp. 101–108.

Bowden, E. M., Jung-Beeman, M., Fleck, J., and Kounios, J. (2005), "New Approaches to Demystifying Insight," *Trends in Cognitive Sciences*, Vol. 9, No. 7, pp. 322–328.

Bradlow, E. T. (2005), "Current Issues and a 'Wish List' for Conjoint Analysis," *Applied Stochastic Models in Business and Industry*, Vol. 21, Nos. 4–5, pp. 319–323.

Brannon, I. (2008), "Review: Nudge: Improving Decisions about Health, Wealth, and Happiness," *The Cato Journal*, Vol. 28, No. 3, pp. 562–565.

Brashers, D. E., Adkins, M., and Meyers, R. A. (1994), "Argumentation and Computer-Mediated Group Decision Making," in *Group Communication in Context*, L. R. Frey, Ed., Lawrence Erlbaum Associates, Mahwah, NJ.

Brehmer, B. (1981), "Models of Diagnostic Judgment," in *Human Detection and Diagnosis of System Failures*, J. Rasmussen and W. Rouse, Eds., Plenum, New York.

Brehmer, B., and Joyce, C. R. B. (1988), *Human Judgment: The SJT View*, North-Holland, Amsterdam.

Brookhouse, J. K., Guion, R. M., and Doherty, M. E. (1986), "Social Desirability Response Bias as One Source of the Discrepancy between Subjective Weights and Regression Weights," *Organizational Behavior and Human Decision Processes*, Vol. 37, pp. 316–328.

Brunswik, E. (1952), *The Conceptual Framework of Psychology*, University of Chicago Press, Chicago.

Buck, J. R. (1989), *Economic Risk Decisions in Engineering and Management*, Iowa State University Press, Ames, IA.

Budescu, D., and Weiss, W. (1987), "Reflection of Transitive and Intransitive Preferences: A Test of Prospect Theory," *Organizational Behavior and Human Performance*, Vol. 39, pp. 184–202.

Bui, T. X., Jelassi, T. M. and Shakun, M. F. (1990), "Group Decision and Negotiation Support Systems," *European Journal of Operational Research*, Vol. 46, No. 2, pp. 141–142.

Busemeyer, J. R., and Townsend, J. T. (1993), "Decision Field Theory: A Dynamic–Cognitive Approach to Decision Making in an Uncertain Environment," *Psychological Review*, Vol. 100, pp. 432–459.

Camerer, C. F., and Loewenstein, G. (2004), "Behavioral Economics: Past, Present, Future," in *Advances in Behavioral Economics*, C. F. Camerer, G. Loewenstein, and M. Rabin, Eds., Princeton University Press, Princeton, NJ. pp. 3–51.

Camerer, C. F., Loewenstein, G., and Rabin, M. (2004), *Advances in Behavioral Economics*, Princeton University Press, Princeton, NJ.

Caverni, J. P., Fabre, J. M., and Gonzalez, M. (1990), *Cognitive Biases*, North-Holland, Amsterdam.

Chemers, M. M., and Ayman, R., Eds. (1993), *Leadership Theory and Research*, Academic, San Diego, CA.

Cheng, P. E., and Holyoak, K. J. (1985), "Pragmatic Reasoning Schemas," *Cognitive Psychology*, Vol. 17, pp. 391–416.

Christensen-Szalanski, J. J., and Willham, C. F. (1991), "The Hindsight Bias: A Meta-Analysis," *Organizational Behavior and Human Decision Processes*, Vol. 48, pp. 147–168.

Clemen, R. T. (1996), *Making Hard Decisions: An Introduction to Decision Analysis*, 2nd ed., Duxbury, Belmont, CA.

Cohen, M. S. (1993), "The Naturalistic Basis of Decision Biases," in *Decision Making in Action: Models and Methods*, G. A. Klein, J. Orasanu, R. Calderwood, and E. Zsambok, Eds., Ablex, Norwood, NJ, pp. 51–99.

Collan, M., and Liu, S. (2003), "Fuzzy Logic and Intelligent Agents: Towards the Next Step of Capital Budgeting Decision Support," *Industrial Management and Data Systems*, Vol. 103, No. 6, pp. 410–422.

Coma, O., Mascle, O., and Balazinski, M. (2004), "Application of a Fuzzy Decision Support System in a Design for Assembly Methodology," *International Journal of Computer Integrated Manufacturing*, Vol. 17, No. 1, pp. 83–94.

Connolly, T., Ordonez, L. D., and Coughlan, R. (1997), "Regret and Responsibility in the Evaluation of Decision Outcomes," *Organizational Behavior and Human Decision Processes*, Vol. 70, pp. 73–85.

Cosmides, L., and Tooby, J. (1996), "Are Humans Good Intuitive Statisticians After All? Rethinking Some Conclusions from the Literature on Judgment under Uncertainty", *Cognition*, Vol. 58, pp. 1–73.

Cui, G., and Wong, M. L. (2004), "Implementing Neural Networks for Decision Support in Direct Marketing," *International Journal of Market Research*, Vol. 46, No. 2, pp. 235–254.

Davenport, T. H., and Prusak, L. (2000), *Working Knowledge: How Organizations Manage What They Know*, Harvard Business School Press, Boston.

Davis, J. H. (1992), "Some Compelling Intuitions about Group Consensus Decisions, Theoretical and Empirical Research, and Interperson Aggregation Phenomena: Selected Examples, 1950–1990," *Organizational Behavior and Human Decision Processes*, Vol. 52, pp. 3–38.

Dawes, R. M., and Mulford, M. (1996), "The False Consensus Effect and Overconfidence: Flaws in Judgement or Flaws in How We Study Judgement?" *Organizational Behavior and Human Decision Processes*, Vol. 65, pp. 201–211.

Dawes, R. M., van de Kragt, A. J. C., and Orbell, J. M. (1988), "Not Me or Thee but We: The Importance of Group Identity in Eliciting Cooperation in Dilemma Situations: Experimental Manipulations," *Acta Psychologia*, Vol. 68, pp. 83–97.

Delbecq, A. L., Van de Ven, A. H., and Gustafson, D. H. (1975), *Group Techniques for Program Planning*, Scott, Foresman, Glenview, IL.

Dennis, A. R., and Wixom, B. H. (2002), "Investigators the Moderators of the Group Support Systems Use with Meta-Analysis," *Journal of Management Information Systems*, Vol. 18, No. 3, pp. 235–258.

Dennis, A. R., Haley, B. J., and Vandenberg, R. J. (1996), "A Meta-Analysis of Effectiveness, Efficiency, and Participant Satisfaction in Group Support Systems Research," *Proceedings of the International Conference on Information Systems*, Cleveland, OH, pp. 278–289.

Dennis, A. R., Wixom, B. H., and Vandenberg, R. J. (2001), "Understanding Fit and Appropriation Effects in Group Support Systems via Meta-Analysis," *MIS Quarterly*, Vol. 25, No. 2, pp. 167–193.

DeSanctis, G., and Gallupe, R. B. (1987), "A Foundation for the Study of Group Decision Support Systems," *Management Science*, Vol. 33, No. 5, pp. 589–609.

Dickison, D., and Taatgen, N. A. (2007), "ACT-R Models of Cognitive Control in the Abstract Decision Making Task," in *Proceedings of the Eighth International Conference on Cognitive Modeling*, Ann Arbor, MI, pp. 79–84.

Dixit, A., and Nalebuff, B. (1991), "Making Strategies Credible," in *Strategy and Choice*, R. J. Zechhauser, Ed., MIT Press, Cambridge, MA, pp. 161–184.

Doherty, M. E. (2003), "Optimists, Pessimists, and Realists," in *Emerging Perspectives on Judgment and Decision Research*, S. Schnieder and J. Shanteau, Eds., Cambridge University Press, New York, pp. 643–679.

Dorris, A. L., and Tabrizi, J. L. (1978), "An Empirical Investigation of Consumer Perception of Product Safety," *Journal of Products Liability*, Vol. 2, pp. 155–163.

Dougherty, M. R. P., Gronlund, S. D., and Gettys, C. F. (2003), "Memory as a Fundamental Heuristic for Decision Making," in *Emerging Perspectives on Judgment and Decision Research*, S. Schnieder and J. Shanteau, Eds., Cambridge University Press, New York, pp. 125–164.

Driskell, J. E., and Salas, E., Eds. (1996), *Stress and Human Performance*, Lawrence Erlbaum Associates, Mahwah, NJ.

Drucker, P. F. (1985), *The Effective Executive*, Harper & Row, New York.

Drucker, P. (1995), "The Information Executives Truly Need," *Harvard Business Review*, Vol. 73, No. 1, pp. 54–62.

Du Charme, W. (1970), "Response Bias Explanation of Conservative Human Inference," *Journal of Experimental Psychology*, Vol. 85, pp. 66–74.

Duda, R. O., Hart, K., Konolige, K., and Reboh, R. (1979), "A Computer-Based Consultant for Mineral Exploration," Technical Report, SRI International, Stanford, CA.

Duffy, L. (1993), "Team Decision Making Biases: An Information Processing Perspective," in *Decision Making in Action: Models and Methods*, G. A. Klein, J. Orasanu, R. Calderwood, and E. Zsambok, Eds., Ablex, Norwood, NJ.

Ebbinghaus, H. (1913), *Memory: A Contribution to Experimental Psychology*, Teachers College, Columbia University, New York.

Edland, E., and Svenson, O. (1993), "Judgment and Decision Making under Time Pressure," in *Time Pressure and Stress in Human Judgment and Decision Making*, O. Svenson and A. J. Maule, Eds., Plenum, New York.

Edwards, W. (1954), "The Theory of Decision Making," *Psychological Bulletin*, Vol. 41, pp. 380–417.

Edwards, W. (1968), "Conservatism in Human Information Processing," in *Formal Representation of Human Judgment*, B. Kleinmuntz, Ed., Wiley, New York, pp. 17–52.

Edwards, W., and Fasolo, B. (2001), "Decision Technology," *Annual Review of Psychology*, Vol. 52, No. 1, pp. 581–606.

Einhorn, H. J., and Hogarth, R. M. (1978), "Confidence in Judgment: Persistence of the Illusion of Validity," *Psychological Review*, Vol. 70, pp. 193–242.

Einhorn, H. J., and Hogarth, R. M. (1981), "Behavioral Decision Theory: Processes of Judgment and Choice," *Annual Review of Psychology*, Vol. 32, pp. 53–88.

Ellis, D. G., and Fisher, B. A. (1994), *Small Group Decision Making: Communication and the Group Process*, 4th ed., McGraw-Hill, New York.

Ellsberg, D. (1961), "Risk, Ambiguity, and the Savage Axioms," *Quarterly Journal of Economics*, Vol. 75, pp. 643–699.

Elster, J., Ed. (1986), *The Multiple Self*, Cambridge University Press, Cambridge.

Embrey, D. E. (1984), *SLIM-MAUD: An Approach to Assessing Human Error Probabilities Using Structured Expert Judgment*, NUREG/CR-3518, Vols. 1 and 2, U.S. Nuclear Regulatory Commission, Washington, DC.

Ericsson, K. A., and Simon, H. A. (1984), *Protocol Analysis: Verbal Reports as Data*, MIT Press, Cambridge, MA.

Ernst, M., Bolla, K., Mouratidis, M., Contoreggi, C., Matochik, J. A., Kurian, V., Cadet, J. L., et al. (2002), "Decision-Making in a Risk-taking Task: A PET Study," *Neuropsychopharmacology*, Vol. 26, No. 5, pp. 682–691.

Espinasse, B., Picolet, G., and Chouraqui, E. (1997), "Negotiation Support Systems: A Multi-Criteria and Multi-Agent Approach," *European Journal of Operational Research*, Vol. 103, No. 2, pp. 389–409.

Estes, W. (1976), "The Cognitive Side of Probability Learning," *Psychological Review*, Vol. 83, pp. 37–64.

Etzioni, A. (1988), "Normative-Affective Factors: Toward a New Decision-Making Model," *Journal of Economic Psychology*, Vol. 9, pp. 125–150.

Evans, J. B. T. (1989), *Bias in Human Reasoning: Causes and Consequences*, Lawrence Erlbaum Associates, London.

Evans, J. B. T., and Pollard, P. (1985), "Intuitive Statistical Inferences about Normally Distributed Data," *Acta Psychologica*, Vol. 60, pp. 57–71.

Fallesen, J. J., and Pounds, J. (2001), "Identifying and Testing a Naturalistic Approach for Cognitive Skill Training," in *Linking Expertise and Naturalistic Decision Making*, E. Salas and G. Klien, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 55–70.

Fazlollahi, B., and Vahidov, R. (2001), "Extending the Effectiveness of Simulation-Based DSS through Genetic Algorithms," *Information and Management*, Vol. 39, No. 1, pp. 53–64.

Feather, N. T. (1966), "Effects of Prior Success and Failure on Expectations of Success and Failure," *Journal of Personality and Social Psychology*, Vol. 3, pp. 287–298.

Fedrizzi, M., Kacprzyk, J., and Yager, R. R., Eds. (1994), *Decision Making under Dempster–Shafer Uncertainties*, Wiley, New York.

Fischhoff, B. (1982), "For Those Condemned to Study the Past: Heuristics and Biases in Hindsight," in *Judgment under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic, and A. Tversky, Eds., Cambridge University Press, Cambridge.

Fischhoff, B., and MacGregor, D. (1982), "Subjective Confidence in Forecasts," *Journal of Forecasting*, Vol. 1, pp. 155–172.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1977), "Knowing with Certainty: The Appropriateness of Extreme Confidence," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 3, pp. 552–564.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1978), "Fault Trees: Sensitivity of Estimated Failure Probabilities to Problem Representation," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 4, pp. 330–344.

Fishburn, P. C. (1974), "Lexicographic Orders, Utilities and Decision Rules: A Survey," *Management Science*, Vol. 20, No. 11, pp. 1442–1471.

Flin, R., Salas, E., Strub, M., and Martin, L., Eds. (1997), *Decision Making under Stress: Emerging Themes and Applications*, Ashgate, Aldershot, Hampshire, England.

Fong, G. T., Krantz, D. H., and Nisbett, R. E. (1986), "The Effects of Statistical Training on Thinking about Everyday Problems," *Cognitive Psychology*, Vol. 18, pp. 253–292.

Foroughi, A. (1998), "Minimizing Negotiation Process Losses with Computerized Negotiation Support Systems," *Journal of Applied Business Research*, Vol. 14, No. 4, pp. 15–26.

Fox, C. R., and Hadar, L. (2006), "Decisions from Experience = Sampling Error + Prospect Theory: Reconsidering Hertwig, Barron, Weber & Erev (2004)," *Judgment and Decision Making*, Vol. 1, No. 2, pp. 159–161.

Friedman, J. W. (1990), *Game Theory with Applications to Economics*, Oxford University Press, New York.

Frisch, D., and Clemen, R. T. (1994), "Beyond Expected Utility: Rethinking Behavioral Decision Research," *Psychological Bulletin*, Vol. 116, No. 1, pp. 46–54.

Galton, F. (1907), "Vox Populi," *Nature*, Vol. 75, No. 1949, pp. 450–451.

Garavelli, A. C., and Gorgoglione, M. (1999), "Fuzzy Logic to Improve the Robustness of Decision Support Systems under Uncertainty," *Computers and Industrial Engineering*, Vol. 27, Nos. 1–2, pp. 477–480.

Gertman, D. I., and Blackman, H. S. (1994), *Human Reliability and Safety Analysis Data Handbook*, Wiley, New York.

Gigerenzer, G. (2008), "Why Heuristics Work," *Perspectives on Psychological Science*, Vol. 3, No. 1, pp. 20–29.

Gigerenzer, G., and Hoffrage, U. (1995), "How to Improve Bayesian Reasoning without Instruction: Frequency Formats," *Psychological Review*, Vol. 102, pp. 684–704.

Gigerenzer, G., Hoffrage, U., and Kleinbolting, H. (1991), "Probabilistic Mental Models: A Brunswikian Theory of Confidence," *Psychological Review*, Vol. 98, pp. 506–528.

Gigerenzer, G., Todd, P., and the ABC Research Group (1999), *Simple Heuristics That Make Us Smart*, Oxford University Press, New York.

Glaser, M., Nöth, M., and Weber, M. (2004), "Behavioral Finance," in *Blackwell Handbook of Judgment and Decision Making*, D. J. Koehler and N. Harvey, Eds., Wiley-Blackwell, Hoboken, NJ, pp. 527–546.

Goel, V. (1995), *Sketches of Thought*, MIT Press, Cambridge, MA.

Goel, V., and Morrison, R. G. (2005), "Cognitive Neuroscience of Deductive Reasoning," in *The Cambridge Handbook of Thinking and Reasoning*, K. J. Holyoak, Ed., Cambridge University Press, Cambridge.

Goldstein, W. M., and Hogarth, R. M. (1997), "Judgment and Decision Research: Some Historical Context," in *Research on Judgment and Decision Making: Currents, Connections, and Controversies*, W. M. Goldstein and R. M. Hogarth, Eds., Cambridge University Press, Cambridge, pp. 3–65.

Gordon, J., and Shortliffe, E. H. (1984), "The Dempster–Schafer Theory of Evidence," in *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, B. G. Buchanan and E. H. Shortliffe, Eds., Addison-Wesley, Reading, MA, pp. 272–294.

Green, P. E., and Krieger, A. M. (1987), "A Consumer-Based Approach to Designing Product Line Extensions," *Journal of Product Innovation Management*, Vol. 4, No. 1, pp. 21–32.

Green, P. E., and Krieger, A. M. (1990), "A Hybrid Conjoint Model for Price-Demand Estimation," *European Journal of Operational Research*, Vol. 44, No. 1, pp. 28–38.

Green, P. E., Krieger, A. M., and Wind, Y. (2001), "Thirty Years of Conjoint Analysis: Reflections and Prospects," *Interfaces*, Vol. 31, No. 3, pp. S56–S73.

Gruenfeld, D. H., Mannix, E. A., Williams, K. Y., and Neale, M. A. (1996), "Group Composition and Decision Making: How Member Familiarity and Information Distribution Affect Process and Performance," *Organizational Behavior and Human Decision Making Processes*, Vol. 67, No. 1, pp. 1–15.

Guth, W., Schmittberger, R., and Schwarze, B. (1982), "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization*, Vol. 3, pp. 367–388.

Hamilton, V., and Warburton, D. M., Eds. (1979), *Human Stress and Cognition*, Wiley, New York.

Hammer, W. (1993), *Product Safety Management and Engineering*, 2nd ed., American Society of Safety Engineers, Chicago.

Hammond, K. R. (1980), "Introduction to Brunswikian Theory and Methods," in *Realizations of Brunswick's Experimental Design*, K. R. Hammond and N. E. Wascoe, Eds., Jossey-Bass, San Francisco, CA.

Hammond, K. R. (1993), "Naturalistic Decision Making from a Brunswikian Viewpoint: Its Past, Present, Future," in *Decision Making in Action: Models and Methods*,

G. A. Klein, J. Orasanu, R. Calderwood, and E. Zsambok, Eds., Ablex, Norwood, NJ, pp. 205–227.

Hammond, K. R. (1996), *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice*, Oxford University Press, New York.

Hammond, K. R, Stewart, T. R., Brehmer, B., and Steinmann, D. O. (1975), "Social Judgment Theory," in *Human Judgment and Decision Processes*, M. F. Kaplan and S. Schwartz, Eds., Academic, New York, pp. 271–312.

Hammond, K. R., Hamm, R. M., Grassia, J., and Pearson, T. (1987), "Direct Comparison of the Efficacy of Intuitive and Analytical Cognition in Expert Judgment," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 17, pp. 753–770.

Hardin, G. (1968), "The Tragedy of the Commons," *Science*, Vol. 162, pp. 1243–1248.

Harmon, J., and Rohrbaugh, J. (1990), "Social Judgement Analysis and Small Group Decision Making: Cognitive Feedback Effects on Individual and Collective Performance," *Organizational Behavior and Human Decision Processes*, Vol. 46, pp. 34–54.

Hart, P., Stern, E. K., and Sundelius, B. (1997), *Beyond Groupthink: Political Group Dynamics and Foreign Policy-Making*, University of Michigan Press, Ann Arbor, MI.

Hasher, L., and Zacks, R. T. (1984), "Automatic Processing of Fundamental Information: The Case of Frequency of Occurrence," *American Psychologist*, Vol. 39, pp. 1372–1388.

Heath, C. (1995), "Escalation and De-escalation of Commitment in Response to Sunk Costs: The Role of Budgeting in Mental Accounting," *Organizational Behavior and Human Decision Processes*, Vol. 62, pp. 38–54.

Heath, L., Tindale, R. S., Edwards, J., Posavac, E. J., Bryant, F. B., Henderson-King, E., Suarez-Balcazar, Y., and Myers, J. (1994), *Applications of Heuristics and Biases to Social Issues*, Plenum, New York.

Hertwig, R., Hoffrage, U., and Martignon, L. (1999). "Quick Estimation: Letting the Environment Do the Work," in *Simple Heuristics That Make Us Smart*, G. Gigerenzer, P. Todd, and the ABC Research Group, Eds., Oxford University Press, New York; pp. 209–234.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004), "Decisions from Experience and the Effect of Rare Events in Risky Choice," *Psychological Science*, Vol. 15, No. 8, pp. 534–539.

Hogarth, R. M., and Einhorn, H. J. (1992), "Order Effects in Belief Updating: The Belief-Adjustment Model," *Cognitive Psychology*, Vol. 24, pp. 1–55.

Holden, C. (2007), "Bird Flu Futures," *Science*, Vol. 315, No. 5817, pp. 1345b.

Holsapple, C. W., and Whinston, A. B. (1996), *Decision Support Systems: A Knowledge-Based Approach*, West Publishing, St. Paul, MN.

Holsapple, C. W., Lai, H., and Whinston, A. B. (1998), "A Formal Basis for Negotiation Support System Research," *Group Decision and Negotiation*, Vol. 7, No. 3, pp. 203–227.

Holtzman, S. (1989), *Intelligent Decision Systems*, Addison-Wesley, Reading, MA.

Howard, R. A. (1968), "The Foundations of Decision Analysis," *IEEE Transactions on Systems, Science, and Cybernetics*, Vol. 4, pp. 211–219.

Howard, R. A. (1988), "Decision Analysis: Practice and Promise," *Management Science*, Vol. 34, pp. 679–695.

Huber, J., and Zwerina, K. (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, Vol. 33, No. 3, pp. 307–317.

Huber, J., Wittink, D. R., Fiedler, J. A., and Miller, R. (1993), "The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice," *Journal of Marketing Research*, Vol. 30, pp. 105–114.

Hunt, E. (1998), "Problem Solving," in *Thinking and Problem Solving*, R. J. Sternberg, Ed., Elsevier, pp. 215–232.

Isen, A. M. (1993), "Positive Affect and Decision Making," in *Handbook of Emotions*, M. Lewis and J. M. Haviland, Eds., Guilford, New York, pp. 261–277.

Isenberg, D. J. (1986), "Group Polarization: A Critical Review and Meta Analysis," *Journal of Personality and Social Psychology*, Vol. 50, pp. 1141–1151.

Iverson, G., and Luce, R. D. (1998), "The Representational Measurement Approach to Psychophysical and Judgmental Problems," in *Measurement, Judgement, and Decision Making*, M. H. Birnbaum, Ed., Academic, San Diego, CA, pp. 1–79.

Jacoby, J. (1977), "Information Load and Decision Quality: Some Contested Issues," *Journal of Marketing Research*, Vol. 14, pp. 569–573.

Janis, I. L. (1972), *Victims of Groupthink*, Houghton Mifflin, Boston, MA.

Janis, I. L., and Mann, L. (1977), *Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment*, Free Press, New York.

Jasic, T., and Wood, D. (2004), "The Profitability of Daily Stock Market Indices Trades Based on Neural Network Predictions: Case Study for the S&P 5000, the DAX, the TOPIX and the FTSE in the Period 1965–1999," *Applied Financial Economics*, Vol. 14, No. 4, pp. 285–297.

Johansen, R. (1988), *Groupware: Computer Support for Business Teams*, Free Press, New York.

Johnson, E. J., and Payne, J. W. (1985), "Effort and Accuracy in Choice," *Management Science*, Vol. 31, No. 4, pp. 395–414.

Jonides, J., and Jones, C. M. (1992), "Direct Coding for Frequency of Occurrence," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 18, pp. 368–378.

Juslin, P., and Olsson, H. (1999), "Computational Models of Subjective Probability Calibration," in *Judgment and Decision Making: Neo-Brunswikian and Process-Tracing Approaches*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 67–95.

Kahneman, D. (2003), "Maps of Bounded Rationality: Psychology for Behavioral Economics," *American Economic Review*, Vol. 93, No. 5, pp. 1449–1475.

Kahneman, D., and Tversky, A. (1973), "On the Psychology of Prediction," *Psychological Review*, Vol. 80, pp. 251–273.

Kahneman, D., and Tversky, A. (1979), "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, Vol. 47, pp. 263–291.

Kahneman, D., Slovic, P., and Tversky, A., Eds. (1982), *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.

Keen, P. G. W., and Scott-Morton, M. S. (1978), *Decision Support Systems: An Organizational Perspective*, Addison-Wesley, Reading, MA.

Keeney, R. L., and Raiffa, H. (1976), *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, Wiley, New York.

Keim, D., Andrienko, G., Fekete, J., Görg, C., Kohlhammer, J., and Melançon, G. (2008a), "Visual Analytics: Definition, Process, and Challenges," in *Information Visualization*, pp. 154–175.

Keim, D., Mansmann, F., Schneidewind, J., Thomas, J., and Ziegler, H. (2008b), "Visual Analytics: Scope and Challenges," in *Visual Data Mining*, pp. 76–90.

Keim, D. A., Nietzschmann, T., Schelwies, N., Schneidewind, J., Schreck, T., and Ziegler, H. (2006), "A Spectral Visualization System for Analyzing Financial Time Series Data," in *Proceedings of the Eurographics/IEEE-VGTC Symposium on Visualization (EuroVis)*, Lisbon, Portugal.

Kepner, C. H., and Tregoe, B. B. (1976), *The Rational Manager: A Systematic Approach to Problem Solving and Decision Making*, Kepner-Tregoe, Princeton, NJ.

Keren, G. (1990), "Cognitive Aids and Debiasing Methods: Can Cognitive Pills Cure Cognitive Ills?" in *Cognitive Biases*, J. P. Caverni, J. M. Fabre, and M. Gonzalez, Eds., North-Holland, Amsterdam.

Kerr, L. N., MacCoun, R. J., and Kramer, G. P. (1996), "Bias in Judgment: Comparing Individuals and Groups," *Psychological Review*, Vol. 103, pp. 687–719.

Kersten, G. E., and Noronha, S. J. (1999), "WWW-Based Negotiation Support Systems: Design, Implementation, and Use," *Decision Support Systems*, Vol. 25, No. 2, pp. 135–154.

Kidder, R. M. (1995), *How Good People Make Tough Choices*, Harper Collins, New York.

Kieras, D. E. (1985), "The Role of Prior Knowledge in Operating Equipment from Written Instructions," Report 19 (FR-85/ONR-19), Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI.

Klayman, J., and Ha, Y. W. (1987), "Confirmation, Disconfirmation, and Information in Hypothesis Testing," *Journal of Experimental Psychology: Human Learning and Memory*, Vol. 94, No. 2, pp. 211–228.

Klein, G. A. (1989), "Recognition-Primed Decisions," in *Advances in Man–Machine System Research*, W. Rouse, Ed., JAI Press, Greenwich, CT.

Klein, G. A. (1995), "The Value Added by Cognitive Analysis," in *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*, San Diego, CA, pp. 530–533.

Klein, G. A. (1996), "The Effect of Acute Stressors on Decision Making," in *Stress and Human Performance*, J. E. Driskell, and E. Salas, Eds., Lawrence Erlbaum Associates, Mahwah, NJ.

Klein, G. A. (1998), *Sources of Power: How People Make Decisions*, MIT Press, Cambridge, MA.

Klein, G. A. (2004), *The Power of Intuition: How to Use Your Gut Feelings to Make Better Decisions at Work*, A Currecy Book/Double Day, New York.

Klein, G. A., and Wolf, S. (1995), "Decision-Centered Training," in *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*, San Diego, CA, pp. 1249–1252.

Klein, G. A., Orasanu, J., Calderwood, R., and Zsambok, E., Eds. (1993), *Decision Making in Action: Models and Methods*, Ablex, Norwood, NJ.

Kleinmuntz, B. (1984), "The Scientific Study of Clinical Judgment in Psychology and Medicine," *Clinical Psychology Review*, Vol. 4, pp. 111–126.

Koehler, J. J. (1996), "The Base Rate Fallacy Reconsidered: Descriptive, Normative, and Methodological Challenges," *Behavioral and Brain Sciences*, Vol. 19, pp. 1–53.

Koehler, D. J., and Harvey, N. (2004), *Blackwell Handbook of Judgment and Decision Making*, 1st ed., Wiley-Blackwell, Hoboken, NJ.

Koriat, A., Lichtenstein, S., and Fischhoff, B. (1980), "Reasons for Confidence," *Journal of Experimental Psychology: Human Learning and Memory*, Vol. 6, pp. 107–118.

Kraus, N. N., and Slovic, P. (1988), "Taxonomic Analysis of Perceived Risk: Modeling Individual and Group Perceptions within Homogenous Hazards Domains," *Risk Analysis*, Vol. 8, pp. 435–455.

Krauss, S., Martignon, L., and Hoffrage, U. (1999), "Simplifying Bayesian Inference: The General Case," in *Model-Based Reasoning in Scientific Discovery*, L. Magnani, N. Nersessian, and P. Thagard, Eds., Plenum, New York, pp. 165–179.

Kuhberger, A. (1995), "The Framing of Decisions: A New Look at Old Problems," *Organizational Behavior and Human Decision Processes*, Vol. 62, pp. 230–240.

Kuo, R. J., and Xue, K. C. (1998), "A Decision Support System for Sales Forecasting through Fuzzy Neural Networks with Asymmetric Fuzzy Weights," *Decision Support Systems*, Vol. 24, No. 2, pp. 105–126.

Lam, S. -H., Cao, J. -M., and Fan, H. (2003), "Development of an Intelligent Agent for Airport Gate Assignment," *Journal of Air Transportation*, Vol. 8, No. 1, pp. 103–114.

Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987), "Soar: An Architecture for General Intelligence," *Artificial Intelligence*, Vol. 33, No. 1, pp. 1–64.

LaPlante, A. (1990), "Bring in the Expert: Expert Systems Can't Solve All Problems, But They're Learning," *InfoWorld*, Vol. 12, No. 40, pp. 55–64.

Lari, A. (2003), "A Decision Support System for Solving Quality Problems Using Case-Based Reasoning," *Total Quality Management and Business Excellence*, Vol. 14, No. 6, pp. 733–745.

Lathrop, R. G. (1967), "Perceived Variability," *Journal of Experimental Psychology*, Vol. 23, pp. 498–502.

Laudon, K. C., and Laudon, J. P. (2003), *Management Information Systems: Managing the Digital Firm*, 8th ed., Prentice Hall, Upper Saddle River, NJ.

Lehrer, J. (2008, July 28), "The Eureka Hunt," *New Yorker*, pp. 40–45.

Lehto, M. R. (1991), "A Proposed Conceptual Model of Human Behavior and Its Implications for Design of Warnings," *Perceptual and Motor Skills*, Vol. 73, pp. 595–611.

Lehto, M. R., and Buck, J. R. (2007), *Introduction to Human Factors and Ergonomcs for Engineers*, CRC Press, Boca Raton, FL.

Lehto, M. R., and Papastavrou, J. (1991), "A Distributed Signal Detection Theory Model: Implications to the Design of Warnings," in *Proceedings of the 1991 Automatic Control Conference*, Boston, MA, pp. 2586–2590.

Lehto, M. R., James, D. S., and Foley, J. P. (1994), "Exploratory Factor Analysis of Adolescent Attitudes toward Alcohol and Risk," *Journal of Safety Research*, Vol. 25, pp. 197–213.

Lehto, M. R., Papastavrou, J. P., Ranney, T. A., and Simmons, L. (2000), "An Experimental Comparison of Conservative vs. Optimal Collision Avoidance System Thresholds," *Safety Science*, Vol. 36, No. 3, pp. 185–209.

Levin, L. P. (1975), "Information Integration in Numerical Judgements and Decision Processes," *Journal of Experimental Psychology: General*, Vol. 104, pp. 39–53.

Levitt, S. D., and Dubner, S. J. (2005), *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*, William Morrow & Co., New York.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., and Coombs, B. (1978), "Judged Frequency of Lethal Events," *Journal of Experimental Psychology: Human Learning and Memory*, Vol. 4, pp. 551–578.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1982), "Calibration of Probabilities: The State of the Art to 1980," in *Judgment under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic, and A. Tversky, Eds., Cambridge University Press, Cambridge, pp. 306–334.

Likert, R., and Likert, J. G. (1976), *New Ways of Managing Conflict*, McGraw-Hill, New York.

Lim, L. -H., and Benbasat, I. (1993), "A Theoretical Perspective of Negotiation Support Systems," *Journal of Management Information Systems*, Vol. 9, No. 3, pp. 27–44.

Little, J. D. C. (1970), "Models and Managers: The Concept of a Decision Calculus," *Management Science*, Vol. 16, No. 8, pp. 466–485.

Livnat, Y., Agutter, J., Moon, S., and Foresti, S. (2005), "Visual Correlation for Situational Awareness," in *Proceedings of IEEE Symposium on Information Visualization*, Los Alamitos, CA, pp. 95–102.

Luce, R. D., and Raiffa, H. (1957), *Games and Decisions*, Wiley, New York.

Mallet, L., Vaught, C., and Brnich, M. J., Jr. (1993), "Sociotechnical Communication in an Underground Mine Fire: A Study of Warning Messages during an Emergency Evacuation," *Safety Science*, Vol. 16, pp. 709–728.

Mandler, D. D. (1979), "Thought Processes, Consciousness and Stress," in *Human Stress and Cognition: An Information Processing Approach*, V. Hamilton and D. M. Warburton, Eds., Wiley, West Sussex, Chichester, England.

Martignon, L., and Krauss, S. (2003), "Can L'Homme Eclaire be Fast and Frugal? Reconciling Bayesianism and Bounded Rationality," in *Emerging Perspectives on Judgment and Decision Research*, S. Schnieder and J. Shanteau, Eds., Cambridge University Press, New York, pp. 108–122.

Martocchio, J. J., Webster, J., and Baker, C. R. (1993), "Decision-Making in Management Information Systems Research: The Utility of Policy Capturing Methodology," *Behaviour and Information Technology*, Vol. 12, pp. 238–248.

Matwin, S., Szpakowicz, S., Koperczak, Z., Kersten, G. E., and Michalowski, W. (1989), "Negoplan: An Expert System Shell for Negotiation Support," *IEEE Expert*, Vol. 4, No. 4, pp. 50–62.

Maule, A. J., and Hockey, G. R. J. (1993), "State, Stress, and Time Pressure," in *Time Pressure and Stress in Human Judgment and Decision Making*, O. Svenson and A. J. Maule, Eds., Plenum, New York, pp. 27–40.

McGuire, W. J. (1966), "Attitudes and Opinions," *Annual Review of Psychology*, Vol. 17, pp. 475–514.

McKnight, A. J., Langston, E. A., McKnight, A. S., and Lange, J. E. (1995), "The Bases of Decisions Leading to Alcohol Impaired Driving," in *Proceedings of the 13th International Conference on Alcohol, Drugs, and Traffic Safety*, C. N. Kloeden and A. J. McLean, Eds., Adelaide, Australia, August 13–18, pp. 143–147.

McLeod, P. L. (1992), "An Assessment of the Experimental Literature on Electronic Support of Group Work: Results of a Meta-Analysis," *Human–Computer Interaction*, Vol. 7, pp. 257–280.

Messick, D. M. (1991), "Equality as a Decision Heuristic," in *Psychological Issues in Distributive Justice*, B. Mellers, Ed., Cambridge University Press, New York.

Mayer, R. E. (1983), *Thinking, Problem Solving, Cognition*, W. H. Freeman, New York.

Meyer, D. E., and Kieras, D. E. (1997a), "A Computational Theory of Executive Cognitive Processes and Multiple-Task Performance: I. Basic Mechanisms," *Psychological Review*, Vol. 104, No. 1, pp. 3–65.

Meyer, D. E., and Kieras, D. E. (1997b), "A Computational Theory of Executive Cognitive Processes and Multiple-Task Performance: I. Basic Mechanisms," *Psychological Review*, Vol. 104, No. 1, pp. 3–65.

Mockler, R. L., and Dologite, D. G. (1991), "Using Computer Software to Improve Group Decision-Making," *Long Range Planning*, Vol. 24, pp. 44–57.

Montgomery, H. (1989), "From Cognition to Action: The Search for Dominance in Decision Making," in *Process and Structure in Human Decision Making*, H. Montgomery and O. Svenson, Eds., Wiley, Chichester, West Sussex, England.

Montgomery, H., and Willen, H. (1999), "Decision Making and Action: The Search for a Good Structure," in *Judgment and Decision Making: Neo-Brunswikian and Process-Tracing Approaches*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 147–173.

Moscovici, S. (1976), *Social Influence and Social Change*, Academic, London.

Murphy, A. H., and Winkler, R. L. (1974), "Probability Forecasts: A Survey of National Weather Service Forecasters," *Bulletin of the American Meteorological Society*, Vol. 55, pp. 1449–1453.

Myers, J. L., Suydam, M. M., and Gambino, B. (1965), "Contingent Gains and Losses in a Risky Decision Situation," *Journal of Mathematical Psychology*, Vol. 2, pp. 363–370.

Nah, F., and Benbasat, I. (2004), "Knowledge-Based Support in a Group Decision Making Context: An Expert–Novice Comparison," *Journal of Association for Information Systems*, Vol. 5, No. 3, pp. 125–150.

Nah, F. H., Mao, J., and Benbasat, I. (1999), "The Effectiveness of Expert Support Technology for Decision Making: Individuals Versus Small Groups," *Journal of Information Technology*, Vol. 14, No. 2, pp. 137–147.

Naruo, N., Lehto, M., and Salvendy, G. (1990), "Development of a Knowledge Based Decision Support System for Diagnosing Malfunctions of Advanced Production Equipment," *International Journal of Production Research*, Vol. 28, pp. 2259–2276.

Navon, D. (1979), "The Importance of Being Conservative," *British Journal of Mathematical and Statistical Psychology*, Vol. 31, pp. 33–48.

Newell, A. (1994), *Unified Theories of Cognition*, Harvard University Press, Cambridge, MA.

Newell, A., and Simon, H. A. (1972), *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, NJ.

Nisbett, R., and Ross, L. (1980), *Human Inference: Strategies and Shortcomings of Social Judgment*, Prentice-Hall, Englewood Cliffs, NJ.

Nunamaker, J. F., Dennis, A. R., Valacich, J. S., Vogel, D. R., and George, J. F. (1991), "Electronic Meeting Systems to Support Group Work: Theory and Practice at Arizona," *Communications of the ACM*, Vol. 34, No. 7, pp. 40–61.

Oldenburger, K., Lehto, X., Feinberg, R., Lehto, M., and Salvendy, G. (2007), "Critical Purchasing Incidents in

E-Business," *Behaviour and Information Technology*, Vol. 27, No. 1, pp. 63–77.

Öllinger, M., and Goel, V. (2010), "Problem Solving," in *Towards a Theory of Thinking*, B. Glatzeder, V. Goel, and A. von Müller, Eds., Springer, Berlin/Heidelberg, Germany, pp. 3–21.

Olson, D., and Mossman, C. (2003), "Neural Network Forecasts of Canadian Stock Returns Using Accounting Ratios," *International Journal of Forecasting*, Vol. 19, No. 3, pp. 453–465.

Orasanu, J. (1997), "Stress and Naturalistic Decision Making: Strengthening the Weak Links," in *Decision Making under Stress: Emerging Themes and Applications*, R. Flin, E. Salas, M. Strub, and L. Martin, Eds., Ashgate, Aldershot, Hampshire, England.

Orasanu, J., and Salas, E. (1993), "Team Decision Making in Complex Environments," in *Decision Making in Action: Models and Methods*, G. A. Klein, J. Orasanu, R. Calderwood, and E. Zsambok, Eds., Ablex, Norwood, NJ.

Osborn, F. (1937), *Applied Imagination*, Charles Scribner & Sons, New York.

Oskarsson, A. T., Van Boven, L., McClelland, G. H., and Hastie, R. (2009), "What's Next? Judging Sequences of Binary Events," *Psychological Bulletin*, Vol. 135, No. 2, pp. 262–285.

Paese, P. W., Bieser, M., and Tubbs, M. E. (1993), "Framing Effects and Choice Shifts in Group Decision Making," *Organizational Behavior and Human Decision Processes*, Vol. 56, pp. 149–165.

Payne, J. W. (1980), "Information Processing Theory: Some Concepts and Methods Applied to Decision Research," in *Cognitive Processes in Choice and Decision Research*, T. S. Wallsten, Ed., Lawrence Erlbaum Associates, Hillsdale, NJ.

Payne, J. W., and Bettman, J. R. (2004), "Walking with the Scarecrow: The Information-Processing Approach to Decision Research," in *Blackwell Handbook of Judgment and Decision Making*, D. J. Koehler, N. Harvey, Eds., Wiley-Blackwell, Hoboken, NJ, pp. 110–132.

Payne, J. W., Bettman, J. R., and Johnson, E. J. (1988), "Adaptive Strategy Selection in Decision Making.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 14, No. 3, pp. 534–552.

Payne, J. W., Bettman, J. R., and Johnson, E. J. (1993), *The Adaptive Decision Maker*, Cambridge University Press, Cambridge.

Pennington, N., and Hastie, R. (1986), "Evidence Evaluation in Complex Decision Making," *Journal of Personality and Social Psychology*, Vol. 51, No. 2, pp. 242–258.

Pennington, N., and Hastie, R. (1988), "Explanation-Based Decision Making: Effects of Memory Structure on Judgment," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 14, No. 3, pp. 521–533.

Pesendorfer, W. (2006), "Behavioral Economics Comes of Age: A Review Essay on 'Advances in Behavioral Economics,'" *Journal of Economic Literature*, Vol. 44, No. 3, pp. 712–721.

Peterson, C. R., and Beach L. R. (1967), "Man as an Infinitive Statistician," *Psychological Bulletin*, Vol. 68, pp. 29–46.

Pitz, G. F. (1980), "The Very Guide of Life: The Use of Probabilistic Information for Making Decisions," in *Cognitive Processes in Choice and Decision Behavior*, T. S. Wallsten, Ed., Lawrence Erlbaum Associates, Mahwah, NJ.

Plaisant, C. (2004), "The Challenge of Information Visualization Evaluation," in *Proceedings of the Working Conference on Advanced Visual Interfaces*, ACM, Gallipoli, Italy, pp. 109–116.

Pliske, R., and Klein, G. (2003), "The Naturalistic Decision-Making Perspective," in *Emerging Perspectives on Judgment and Decision Research*, S. Schnieder and J. Shanteau, Eds., Cambridge University Press, New York, pp. 108–122.

Pliske, R. M., McCloskey, M. J., and Klein, G. (2001), "Decision Skills Training: Facilitating Learning from Experience," in *Linking Expertise and Naturalistic Decision Making*, E. Salas and G. Klien, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 37–53.

Raiffa, H. (1968), *Decision Analysis*, Addison-Wesley, Reading, MA.

Raiffa, H. (1982), *The Art and Science of Negotiation*, Harvard University Press, Cambridge, MA.

Rashid, A. M., Albert, I., Cosley, D., Lam, S. K., McNee, S. M., Konstan, J. A., and Riedl, J. (2002), "Getting to Know You: Learning New User Preferences in Recommender Systems," in *Proceedings of the 7th International Conference on Intelligent User Interfaces*, ACM, San Francisco, CA, pp. 127–134.

Rasmussen, J. (1983), "Skills, Rules, Knowledge: Signals, Signs, and Symbols and Other Distinctions in Human Performance Models," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 13, No. 3, pp. 257–267.

Reason, J. (1990), *Human Error*, Cambridge University Press, Cambridge.

Reitman, W. R. (1964), "Heuristic Decision Procedures, Open Constraints, and the Structure of Ill-Defined Problems," in *Human Judgments and Optimality*, M. W. Shelly and G. L. Bryan, Eds., Wiley, New York, pp. 282–315.

Rethans, A. J. (1980), "Consumer Perceptions of Hazards," in *PLP-80 Proceedings*, pp. 25–29.

Reyna, V. F., and Brainerd, C. F. (1995), "Fuzzy-Trace Theory: An Interim Synthesis," *Learning and Individual Differences*, Vol. 7, pp. 1–75.

Robert, H. M. (1990), *Robert's Rules of Order Newly Revised*, 9th ed., Scott, Foresman, Glenview, IL.

Rumelhart, D. E., and McClelland, J. L. (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, MIT Press, Cambridge, MA.

Saaty, T. L. (1990), *Multicriteria Decision Making: The Analytic Hierarchy Process*, RWS Publications, Pittsburgh, PA.

Sage, A. (1981), "Behavioral and Organizational Considerations in the Design of Information Systems and Processes for Planning and Decision Support," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 11, pp. 61–70.

Sapena, O., Botti, V., and Argente, E. (2003), "Application of Neural Networks to Stock Prediction in 'Pool' Companies," *Applied Artificial Intelligence*, Vol. 17, No. 7, pp. 661–673.

Satish, U., and Streufert, S. (2002), "Value of a Cognitive Simulation in Medicine: Towards Optimizing Decision Making Performance of Healthcare Personnel," *Quality and Safety in Health Care*, Vol. 11, No. 2, pp. 163–167.

Satzinger, J., and Olfman, L. (1995), "Computer Support for Group Work: Perceptions of the Usefulness of Support Scenarios and Specific Tools," *Journal of Management Systems*, Vol. 11, No. 4, pp. 115–148.

Savage, L. J. (1954), *The Foundations of Statistics*, Wiley, New York.

Schafer, J. B., Konstan, J. A., and Riedl, J. (2001), "E-Commerce Recommendation Applications," *Data Mining and Knowledge Discovery*, Vol. 5, No. 1, pp. 115–153.

Schelling, T. (1960), *The Strategy of Conflict*, Harvard University Press, Cambridge, MA.

Schelling, T. (1978), *Micromotives and Macrobehavior*, W.W. Norton, New York.

Scott-Morton, M. S. (1977), *Management Decision Systems: Computer-Based Support for Decision Making*, Harvard University Press, Cambridge, MA.

Sedlmeier, P., Hertwig, R., and Gigerenzer, G. (1998), "Are Judgments of the Positional Frequencies of Letters Systematically Biased due to Availability?" *Journal of Experimental Psychology: Learning, Memory and Cognition*, Vol. 24, No. 3, pp. 754–770.

Selye, H. (1936), "A Syndrome Produced by Diverse Noxious Agents," *Nature*, Vol. 138, p. 32.

Selye, H. (1979), "The Stress Concept and Some of Its Implications," in *Human Stress and Cognition*, V. Hamilton and D. M. Warburton, Eds., Wiley, New York.

Shafer, G. (1976), *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, NJ.

Shafir, E., Diamond, P., and Tversky, A. (1997), "Money Illusion," *Quarterly Journal of Economics*, Vol. 112, No. 2, pp. 341–374.

Sharda, R., Barr, S. H., and McDonnell, J. C. (1988), "Decision Support System Effectiveness: A Review and an Empirical Test," *Management Science*, Vol. 34, No. 2, pp. 139–159.

Siebold, D. R. (1992), "Making Meetings More Successful: Plans, Formats, and Procedures for Group Problem-Solving," in *Small Group Communication*, 6th ed., R. Cathcart and L. Samovar, Eds., Wm. C. Brown, Dubuque, IA, pp. 178–191.

Simon, H. A. (1955), "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics*, Vol. 69, pp. 99–118.

Simon, H. A. (1977), *The New Science of Management Decisions*, Prentice-Hall, Englewood Cliffs, NJ.

Simon, H. A. (1982), *Models of Bounded Rationality*, MIT Press, Cambridge, MA.

Simon, H. A. (1983), "Alternative Visions of Rationality," in *Reason in Human Affairs*, H. A. Simon, Ed., Stanford University Press, Palo Alto, CA, pp. 3–36.

Simon, H. A. (1987), "Behavioural Economics," *The New Palgrave: A Dictionary of Economics*, Vol. 1, pp. 221–225.

Simon, H. A., Dantzig, G. B., Hogarth, R., Plott, C. R., Raiffa, H., Schelling, T. C., Shepsle, K. A., et al. (1987), "Decision Making and Problem Solving," *Interfaces*, Vol. 17, No. 5, pp. 11–31.

Singleton, W. T., and Hovden, J. (1987), *Risk and Decisions*, Wiley, New York.

Slovic, P. (1978), "The Psychology of Protective Behavior," *Journal of Safety Research*, Vol. 10, pp. 58–68.

Slovic, P. (1987), "Perception of Risk," *Science*, Vol. 236, pp. 280–285.

Slovic, P., Fischhoff, B., and Lichtenstein, S. (1977), "Behavioral Decision Theory," *Annual Review of Psychology*, Vol. 28, pp. 1–39.

Smith, A. (1875), *The Theory of Moral Sentiments: To Which Is Added, a Dissertation on the Origin of Languages*, George Bell & Sons, London, UK.

Sniezek, J. A. (1992), "Groups under Uncertainty: An Examination of Confidence in Group Decision Making," *Organizational Behavior and Human Decision Processes*, Vol. 52, pp. 124–155.

Sniezek, J. A., and Henry, R. A. (1989), "Accuracy and Confidence in Group Judgment," *Organizational Behavior and Human Decision Processes*, Vol. 43, pp. 1–28.

Sniezek, J. A., Wilkins, D. C., Wadlington, P. L., and Baumann, M. R. (2002), "Training for Crisis Decision-Making: Psychological Issues and Computer-Based Solutions," *Journal of Management Information Systems*, Vol. 18, No. 4, pp. 147–168.

Stanoulov, N. (1994), "Expert Knowledge and Computer-Aided Group Decision Making: Some Pragmatic Reflections," *Annals of Operations Research*, Vol. 51, pp. 141–162.

Stasko, J., Görg, C., and Spence, R. (2008), "Jigsaw: Supporting Investigative Analysis through Interactive Visualization," *Information Visualization*, Vol. 7, No. 2, pp. 118–132.

Stasser, G., and Titus, W. (1985), "Pooling of Unshared Information in Group Decision Making: Biased Information Sampling during Discussion," *Journal of Personality and Social Psychology*, Vol. 48, pp. 1467–1478.

Sternberg, R. J., and Frensch, P. A. (1991), *Complex Problem Solving: Principles and Mechanisms*, Lawrence Erlbaum Associates, Mahwah, NJ.

Stevenson, M. K., Busemeyer, J. R., and Naylor, J. C. (1993), "Judgment and Decision-Making Theory," in *Handbook of Industrial and Organizational Psychology*, 2nd ed., Vol. 1, M. D. Dunnette and L. M. Hough, Eds., Consulting Press, Palo Alto, CA.

Stone, M. (1960), "Models for Reaction Time," *Psychometrika*, Vol. 25, pp. 251–260.

Strat, T. M. (1994), "Decision Analysis Using Belief Functions," in *Decision Making under Dempster–Shafer Uncertainties*, M. Fedrizzi, J. Kacprzyk, and R. R. Yager, Eds., Wiley, New York.

Stukey, E., and Zeckhauser, R. (1978), "Decision Analysis," in *A Primer for Policy Analysis*, W.W. Norton, New York, pp. 201–254.

Surowiecki, J. (2004), *The Wisdom of Crowds: Why the Many Are Smarter Than the Few*, Doubleday, New York, NY.

Svenson, O. (1990), "Some Propositions for the Classification of Decision Situations," in *Contemporary Issues in Decision Making*, K. Borcherding, O. Larichev, and D. Messick, Eds., North-Holland, Amsterdam, pp. 17–31.

Svenson, O., and Maule, A. J. (1993), *Time Pressure and Stress in Human Judgment and Decision Making*, Plenum, New York.

Sviokla, J. J. (1986), "PlanPower, Xcon and Mudman: An In-Depth Analysis into Three Commercial Expert Systems in Use," unpublished Ph.D. dissertation, Harvard University, Cambridge, MA.

Sviokla, J. J. (1989), "Expert Systems and Their Impact on the Firm: The Effects of PlanPower Use on the Information Processing Capacity of the Financial Collaborative," *Journal of Management Information Systems*, Vol. 6, No. 3, pp. 65–84.

Sviokla, J. J. (1990), "An Examination of the Impact of Expert Systems on the Firm: The Case of XCON," *MIS Quarterly*, Vol. 14, No. 2, pp. 127–140.

Swain, A. D., and Guttman, H. (1983), *Handbook for Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications*, NUREG/CR-1278, U.S. Nuclear Regulatory Commission, Washington, DC.

Tanner, W. P., and Swets, J. A. (1954), "A Decision Making Theory of Visual Detection," *Psychological Review*, Vol. 61, pp. 401–409.

Thaler, R. H. (1980), "Toward a Positive Theory of Consumer Choice," *Journal of Economic Behavior & Organization*, Vol. 1, No. 1, pp. 39–60.

Thaler, R. H. (1988), "Anomalies: The Ultimatum Game," *Journal of Economic Perspectives*, Vol. 2, No. 4, pp. 195–206.

Thaler, R. H., and Sunstein, C. R. (2008), *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press, New Haven, CT.

Thomas, J. J., and Cook, K. A. (2005), "Illuminating the Path: The Research and Development Agenda for Visual Analytics," *IEEE Computer Society*, Washington DC.

Thomas, J. J., and Kielman, J. (2009), "Challenges for Visual Analytics," *Information Visualization*, Vol. 8, No. 4, pp. 309–314.

Todd, P., and Benbasat, I. (1991), "An Experimental Investigation of the Impact of Computer Based Decision Aids on Decision Making Strategies," *Information Systems Research*, Vol. 2, No. 2, pp. 87–115.

Todd, P., and Benbasat, I. (1992), "The Use of Information in Decision Making: An Experimental Investigation of the Impact of Computer Based DSS on Processing Effort," *MIS Quarterly*, Vol. 16, No. 3, pp. 373–393.

Todd, P., and Benbasat, I. (1993), "An Experimental Investigation of the Relationship between Decision Makers, Decision Aids and Decision Making Effort," *INFOR*, Vol. 31, No. 2, pp. 80–100.

Todd, P., and Benbasat, I. (1999), "Evaluating the Impact of DSS, Cognitive Effort, and Incentives on Strategy Selection," *Information Systems Research*, Vol. 10, No. 4, pp. 356–374.

Todd, P., and Benbasat, I. (2000), "Inducing Compensatory Information Processing through Decision Aids That Facilitate Effort Reduction: An Experimental Assessment," *Journal of Behavioral Decision Making*, Vol. 13, No. 1, pp. 91–106.

Torrance, E. P. (1953), "The Behavior of Small Groups under the Stress Conditions of 'Survival,'" *American Sociological Review*, Vol. 19, pp. 751–755.

Toubia, O., Simester, D. I., Hauser, J. R., and Dahan, E. (2003), "Fast Polyhedral Adaptive Conjoint Estimation," *Marketing Science*, Vol. 22, No. 3, pp. 273–303.

Trepel, C., Fox, C. R., and Poldrack, R. A. (2005), "Prospect Theory on the Brain? Toward a Cognitive Neuroscience of Decision Under Risk," *Cognitive Brain Research*, Vol. 23, No. 1, pp. 34–50.

Tuckman, B. W. (1965), "Development Sequence in Small Groups," *Psychological Bulletin*, Vol. 63, pp. 289–399.

Turban, E., and Aronson, J. E. (2001), *Decision Support Systems and Intelligent Systems*, 6th ed., Prentice Hall, Upper Saddle River, NJ.

Tversky, A. (1969), "Intransitivity of Preferences," *Psychological Review*, Vol. 76, pp. 31–48.

Tversky, A. (1972), "Elimination by Aspects: A Theory of Choice," *Psychological Review*, Vol. 79, pp. 281–289.

Tversky, A., and Kahneman, D. (1973), "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology*, Vol. 5, pp. 207–232.

Tversky, A., and Kahneman, D. (1974), "Judgment under Uncertainty: Heuristics and Biases," *Science*, Vol. 185, pp. 1124–1131.

Tversky, A., and Kahneman, D. (1981), "The Framing of Decisions and the Psychology of Choice," *Science*, Vol. 211, pp. 453–458.

Valenzi, E., and Andrews, I. R. (1973), "Individual Differences in the Decision Processes of Employment Interviews," *Journal of Applied Psychology*, Vol. 58, pp. 49–53.

Vavra, T. G., Green, P. E., and Krieger, A. M. (1999), "Evaluating EZ-Pass," *Marketing Research*, Vol. 11, No. 3, pp. 5–16.

Viscusi, W. K. (1991), *Reforming Products Liability*, Harvard University Press, Cambridge, MA.

von Neumann, J., and Morgenstern, O. (1947), *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ.

von Winterfeldt, D., and Edwards, W. (1986), *Decision Analysis and Behavioral Research*, Cambridge University Press, Cambridge.

Wagenaar, W. A. (1992), "Risk Taking and Accident Causation," in *Risk-Taking Behavior*, J. F. Yates, Ed., Wiley, New York, pp. 257–281.

Wallenius, J., Dyer, J. S., Fishburn, P. C., Steuer, R. E., Zionts, S., and Deb, K. (2008), "Multiple Criteria Decision Making, Multiattribute Utility Theory: Recent Accomplishments and What Lies Ahead," *Management Science*, Vol. 54, No. 7, pp. 1336–1349.

Wallsten, T. S. (1972), "Conjoint-Measurement Framework for the Study of Probabilistic Information Processing," *Psychological Review*, Vol. 79, pp. 245–260.

Wallsten, T. S. (1976), "Using Conjoint-Measurement Models to Investigate a Theory about Probabilistic Information Processing," *Journal of Mathematical Psychology*, Vol. 14, pp. 144–185.

Wallsten, T. S. (1995), "Time Pressure and Payoff Effects on Multidimensional Probabilistic Inference," in *Time Pressure and Stress in Human Judgment*, O. Svenson and J. Maule, Eds., Plenum, New York.

Wallsten, T. S., Zwick, R., Kemp, S., and Budescu, D. V. (1993), "Preferences and Reasons for Communicating Probabilistic Information in Verbal and Numerical Terms," *Bulletin of the Psychonomic Society*, Vol. 31, pp. 135–138.

Wattenberg, M. (1999), "Visualizing the Stock Market," in *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, ACM, Pittsburgh, PA, pp. 188–189.

Weber, E. (1994), "From Subjective Probabilities to Decision Weights: The Effect of Asymmetric Loss Functions on the Evaluation of Uncertain Outcomes and Events," *Psychological Bulletin*, Vol. 115, pp. 228–242.

Weber, E., Anderson, C. J., and Birnbaum, M. H. (1992), "A Theory of Perceived Risk and Attractiveness," *Organizational Behavior and Human Decision Processes*, Vol. 52, pp. 492–523.

Wegner, D. (1987), "Transactive Memory: A Contemporary Analysis of Group Mind," in *Theories of Group Behavior*, B. Mullen and G. R. Goethals, Eds., Springer-Verlag, New York, pp. 185–208.

Weinstein, N. D. (1979), "Seeking Reassuring or Threatening Information about Environmental Cancer," *Journal of Behavioral Medicine*, Vol. 2, pp. 125–139.

Weinstein, N. D. (1980), "Unrealistic Optimism about Future Life Events," *Journal of Personality and Social Psychology*, Vol. 39, pp. 806–820.

Weinstein, N. D. (1987), "Unrealistic Optimism about Illness Susceptibility: Conclusions from a Community-Wide Sample," *Journal of Behavioral Medicine*, Vol. 10, pp. 481–500.

Weinstein, N. D., and Klein, W. M. (1995), "Resistance of Personal Risk Perceptions to Debiasing Interventions," *Health Psychology*, Vol. 14, pp. 132–140.

Welch, D. D. (1994), *Conflicting Agendas: Personal Morality in Institutional Settings*, Pilgrim Press, Cleveland, OH.

Welford, A. T. (1976), *Skilled Performance*, Scott, Foresman, Glenview, IL.

Wickens, C. D. (1992), *Engineering Psychology and Human Performance*, HarperCollins, New York.

Wind, J., Green, P. E., Shifflet, D., and Scarbrough, M. (1989), "'Courtyard by Marriott': Designing a Hotel Facility with Consumer-Based Marketing Models," *Interfaces*, Vol. 19, No. 1, pp. 25–47.

Winkler, R. L., and Murphy, A. H. (1973), "Experiments in the Laboratory and the Real World," *Organizational Behavior and Human Performance*, Vol. 10, pp. 252–270.

Wolfers, J., and Zitzewitz, E. (2004), "Prediction Markets," *Journal of Economic Perspectives*, Vol. 18, No. 2, pp. 107–126.

Wright, P. (1975), "Consumer Choice Strategies: Simplifying vs. Optimizing," *Journal of Marketing Research*, Vol. 12, No. 1, pp. 60–67.

Yakes, R. M., and Dodson J. O. (1908), "The Relative Strength of Stimulus to Rapidity of Habit Formation," *Journal of Comparative and Neurological Psychology*, Vol. 18, pp. 459–482.

Yates, J. F., Ed. (1992), *Risk-Taking Behavior*, Wiley, New York.

Yates, J. F., Veinott, E. S., and Patalano, A. L. (2003), "Hard Decisions, Bad Decisions: On Decision Quality and Decision Aiding," in *Emerging Perspectives on Judgment and Decision Research*, S. Schnieder and J. Shanteau, Eds., Cambridge University Press, New York, pp. 13–63.

Zander, A. (1994), *Making Groups Effective*, 2nd ed., Jossey-Bass, San Francisco.

Zey, M., Ed. (1992), *Decision Making: Alternatives to Rational Choice Models*, Sage, London.

Zhou, M., and Xu, S. (1999), "Dynamic Recurrent Neural Networks for a Hybrid Intelligent Decision Support System for the Metallurgical Industry," *Expert Systems*, Vol. 16, No. 4, pp. 240–247.

Zigurs, I., and Buckland, B. K. (1998), "A Theory of Task/Technology Fit and Group Support Systems Effectiveness," *MIS Quarterly*, Vol. 22, No. 3, pp. 313–334.

Zimmer, A. (1983), "Verbal Versus Numerical Processing of Subjective Probabilities," in *Decision Making under Uncertainty*, R. W. Scholtz, Ed., North-Holland, Amsterdam.

# CHAPTER **8**

# MENTAL WORKLOAD AND SITUATION AWARENESS

**Michael A. Vidulich**
**Warfighter Interface Division**
**Air Force Research Laboratory**
**Wright-Patterson Air Force Base, Ohio**

**Pamela S. Tsang**
**Department of Psychology**
**Wright State University**
**Dayton, Ohio**

## 1 INTRODUCTION

Mental workload and situation awareness (SA) have certainly been topics of great interest to, and ongoing debate among, practitioners and researchers in human factors and cognitive engineering. By most accounts, the concept of mental workload that first permeated the literature in the 1970s (Leplat and Welford, 1978; Moray, 1979) and the concept of SA that first came on the scene in the late 1980s (e.g., Endsley, 1995; Pew, 1994) have matured to the point of their widespread applications in domains that range from office work to medicine, transportation, and military operations, to mention just a few (e.g., Proctor and Vu, 2010). Although there remain skeptics who argue that the constructs of mental workload and SA are nothing more than folk models lacking scientific merit (Dekker and Woods, 2002; Dekker and Hollnagel, 2004), Parasuraman et al. (2008) forcefully countered this on two grounds: (a) the constructs of mental workload and SA are distinguishable from other cognitive constructs such as attention or memory and

(b) the distinction is based on a large scientific base of empirical studies and theoretical treatments of the constructs (see Gopher, 1994; Vidulich, 2003; Endsley, 2006; Tenny and Pew, 2006; Tsang and Vidulich, 2006; Durso et al., 2007; Durso and Alexander, 2010). While the work is far from complete, research on their theoretical underpinnings and methodological exactitude continues, albeit at a slower pace than when the concepts first emerged (Proctor and Vu, 2010). The goal of this chapter is to review the current developments in the understanding of the constructs of mental workload and SA and their applications in selected domains.

Human engineering seeks to understand and improve human interactions with machines to perform tasks. This goal can be especially difficult to achieve in dynamic complex systems that characterize much of modern work. A good example of the problem can be seen in considering the human operator's response to automation. Many complex tasks, such as those involved in monitoring and managing a process control plant

(Moray, 1997; Wickens and Hollands, 2000) or remotely controlling multiple uninhabited vehicles (Parasuraman et al., 2009), would not be possible without the assistance of automation. However, many researchers (e.g., Kessel and Wickens, 1982; Bainbridge, 1987; Moray, 1986; Wiener and Curry, 1980; Tsang and Vidulich, 1989; Adams et al., 1991; Prewett et al., 2010) have pointed out that automated assistance can come at a high price. Moray (1986) noted that, as computer automation did more, the operator would do less and therefore experience less mental workload (Moray, 1986, p. 40-5):

> Is there a price for the advantages? It could be said that the information processing demands become so alien to the operator that if called upon to reenter the control loop such reentry is no longer possible. . . . The system will be poorly understood and the operator will lack practice in exercising control, so the possibility of human error in emergencies will increase.

Naturally there is an issue of how well the requirements of using any machine match the capabilities of the human operators. Some mismatches should be easy for an observer to discern, especially physical mismatches. For example, Fitts and Jones (1947/1961a) found that 3% of "pilot errors" in a corpus of 460 were due to the pilot being physically unable to reach the required control. But, other mismatches may not be so obvious, especially mental mismatches. Fitts and Jones (1947/1961b) found that 18% of errors in reading and interpreting aircraft instruments were due to problems associated with integrating the information from multirevolution indicators (e.g., altitude displays with different pointers for one, tens, and hundreds of feet). An outside observer would not necessarily know by watching the pilot that such a misinterpretation had occurred.

Given the impossibility of seeing the mental processes of an operator performing a task, the need to know set off a myriad of research and applied activities to explore and use constructs like mental workload and SA to shed light on the effectiveness of the coupling of the human operator and the machine with which the operator interacts. Vidulich (2003) echoed the argument of other researchers that mental workload and SA had taken on the quality of metameasures that encapsulate the demand on and the quality of the mental processes involved (Hardiman et al., 1996; Selcon et al., 1996). These measures are particularly useful at times when more specific information is simply not available. Other times, they may actually be preferable and sufficient. For example, an interface that allows the task to be performed with a more comfortable level of mental workload and better SA would be preferable to one that did not. Recently, Parasuraman et al. (2008) advocated that mental workload and SA are among a small number of human cognition and performance constructs that have the highly useful properties of being both predictive of performance in complex human–machine systems and diagnostic of the operator's cognitive state. Consequently, measures of both mental workload and SA can provide insight to practitioners seeking to improve the performance of human–machine systems.

To appreciate the potential roles workload and SA might play in supporting system development, visions of two future systems in aviation will be considered. Fallows (2001a,b) presented an intriguing vision of the future of air travel. Noting the increasing bottlenecks and delays inherent in the existing airline industry, Fallows expected that a simple scaling up of the existing system with more planes and more runways at existing airports would not be a practical or economically viable approach to keep pace with the projected increases in air travel. Fallows envisaged that the increased reliability of aircraft mechanical systems combined with innovative research on cockpit interfaces will not only revitalize general aviation but also lead to the emergence of a much more extensive light-jet air taxi industry. Such a development would naturally lead to more pilots flying that lack the extensive training of current professional pilots. But Fallows, along with the Advanced General Aviation Transport Experiments (AGATE) consortium of the National Aeronautics and Space Administration (NASA), the Federal Aviation Administration (FAA), the general aviation industry and a number of universities expected that future general aviation cockpits will take advantage of advanced technologies such as those of highway-in-the-sky displays, synthetic vision systems, and decision-making aids to reduce the mental workload and increase the SA of the pilot in order to maintain an acceptable level of safety. Although the vision that light-jet air travel would be commonplace has not materialized to date, considerable innovation is being demonstrated and some noteworthy accomplishments have been achieved (Fallows 2008a,b,c).

On a much broader scale, NextGen is a transformational modernization plan for the National Airspace System aimed at meeting the demands of significant growth in air traffic projected through 2025 (Joint Planning and Development Office, 2007). Sweeping changes in equipage supported by advance technologies are anticipated to fundamentally change the roles and responsibilities of the pilots and air traffic controllers. Reduced flight time, delays, and fuel expenditures are all expected to profit from such changes. A number of researchers have identified mental workload and SA to be among the most critical human factors considerations in the planning and assessment of NextGen systems for shaping a safer and more effective airspace (e.g., Durso and Manning, 2008; Langan-Fox et al., 2009; Sheridan, 2009). However, these researchers also have expressed concerns about the limitations of extant measurement tools and encouraged continued mental workload and SA research.

Technological changes constantly bring about a practical need to know about their impact on the human operator and on the safe and effective operation of the human–machine system. There seems to be a consensus that the concepts of both mental workload and SA are vital for building safe and effective systems that best accommodate the human's cognitive strengths while supporting human frailty. There is also the growing recognition that the utility of one concept does not replace or diminish the utility of the other. In

fact, parallel studies of both are not only good for applied evaluations but could also help sharpen their respective definitions and stimulate new understanding (Endsley, 1993; Wickens, 2001; Vidulich, 2003; Parasuraman, et al., 2008; Durso and Alexander, 2010; Vidulich et al., 2010).

## 2 THEORETICAL UNDERPINNINGS OF MENTAL WORKLOAD AND SITUATION AWARENESS

In 1994, Pew stated that *situation awareness* had replaced *workload* as the buzzword of the 1990s. But can the concept of situation awareness replace that of mental workload? Hendy (1995) and Wickens (1995, 2001) argued that the two concepts are clearly distinct but are also intricately related to each other. This idea has seemed to have congealed in the workload and SA literature and in practice (e.g., Durso and Alexander, 2010). Although the two concepts are affected by many of the same human variables (such as limited processing capacity and the severe limit of working memory) and system variables (such as task demands, system constraints, and technological support), one concept does not supplant, but rather complements, the other. For example, Ma and Kaber (2005) used both workload and SA measures to assess the individual and combined impact of adaptive cruise control and cell phone use on a simulated car-driving task. They found that the two types of measures combined to create a more complete and compelling picture of the benefits and risks associated with the use of such technologies during automobile driving.

Figure 1 provides a conceptual sketch of the relationship between mental workload and SA and is not intended to be a complete representation of all the processes involved. There are two main components in this figure: the attention and mental workload loop and the memory and SA loop. The ensuing portrayal will make clear that mental workload and SA are intricately intertwined, as one affects and is affected by the other. Although convention would bias us in thinking that elements on top or on the left in the figure might have temporal precedence over those at the bottom or on the right, this is not necessarily the case with the dynamic interplay between workload and SA. For example, task demands could be initiated by an external task (such as the need to respond to an air traffic controller's request) as well as by an internal decision to engage in solving a nagging problem. Despite the seemingly discrete and linear depiction of the relations among the elements in the figure, the elements are actually thought of to be mutually interacting adaptively in response to both exogenous demands and endogenous states (e.g., Hockey, 1997, 2008; Kramer and Parasuraman, 2007).

### 2.1 Attention and Workload

Since the 1970s, much has been debated and written about the concept of mental workload (e.g., Welford, 1978; Moray, 1979; Gopher and Donchin, 1986; O'Donnell and Eggemeier, 1986; Adams et al., 1991; Huey and Wickens, 1993; Gopher, 1994; Kramer et al., 1996; Tsang and Wilson, 1997; Vidulich, 2003; Hockey, 2008; Durso and Alexander, 2010). Parasuraman et al. (2008) succinctly defined mental workload as "the relation between the function relating the mental resources demanded by a task and those resources available to be supplied by the human operator" (pp. 145–146). This supply-and-demand notion is portrayed by the attention–workload loop in Figure 1.

There are two main determinants of mental workload: exogenous task demands as specified by factors such as task difficulty, task priority, situational contingencies (represented by the World in Figure 1), and endogenous supply of attentional or processing resources to support information processing such as perceiving, updating memory, planning, decision making, and response processing. Further, this supply is modulated by individual differences such as one's skill set and expertise. The ultimate interest in measuring
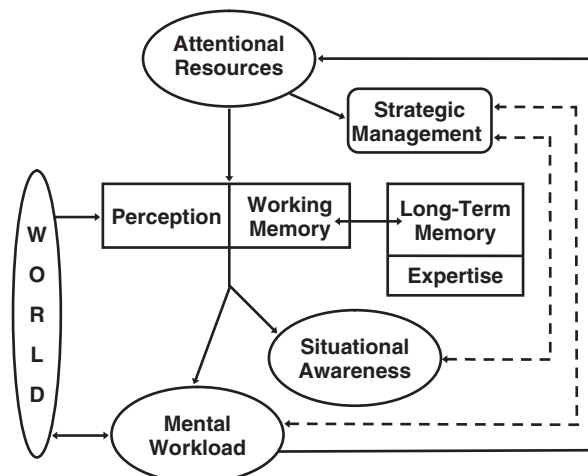


**Figure 1** Conceptual framework illustrating the relationship between mental workload and SA.

workload, of course, is how it might affect system performance represented via the feedback loop back to the world in Figure 1. Mental workload can be expressed in subjective experience, performance, and physiological manifestations. A host of assessment techniques now have been developed and are used in both laboratories and applied settings (see Kramer and Parasuraman, 2007; Gawron, 2008). A selected few will be described in Section 3 for illustrative purposes.

Although there are a number of theoretical accounts of attention, the one readily embraced and adopted in the workload literature is the energetics account (e.g., Hockey et al., 1986; Gopher, 1994; Wickens, 2001; Durso and Alexander, 2010). Central to the present discussion is the notion that attentional resources are demanded for task processing, but they are of limited supply. Performance improves monotonically with increasing investment of resources up to the limit of resource availability (Norman and Bobrow, 1975). An important implication of this relationship is that performance could be the basis of inference for the amount of resources used and remained. The latter, referred to as *spare capacity*, could serve as reserve fuel or energy for emergencies and unexpected added demands. Further, attentional resources are subject to voluntary and strategic allocation. According to Kahneman (1973), attention is allocated via a closed feedback loop with continuous monitoring of the efficacy of the allocation policy that is governed by enduring dispositions (of lasting importance, such as one's own name and well-learned rules), momentary intentions (pertinent to the task at hand), and evaluation of the performance (involving self-monitoring of the adequacy of performance in relation to task demands). Among the most convincing support for the limited, energetic, and allocatable property of attentional resources are the reciprocity effects in performance and certain neuroindices observed between time-shared tasks. As the demand or priority of one task changes, the increase in performance, P300 amplitude, or brain activity level in one task has been observed to be accompanied by a decrease in the corresponding measures in the other task (e.g., Gopher et al., 1982; Wickens et al., 1983; Kramer et al., 1987; Sirevaag et al., 1989; Fowler, 1994; Tsang et al., 1996; Parasuraman and Caggiano, 2002; Just, Carpenter, and Miyake, 2003; Tsang, 2007; Low et al., 2009). Because attention can be deployed flexibly, researchers advocate the need to examine the allocation policy in conjunction with the joint performance in a multitask situation in order to assess the workload and spare capacity involved (e.g., Gopher, 1994; Proctor and Vu, 2010).

By the late 1970s, the notion of Kahneman's undifferentiated or all-purpose attentional resource was challenged by a body of data that suggested multiple specialized resources for different types of processing (see Allport et al., 1972; Kinsbourne and Hicks, 1978; Navon and Gopher, 1979; Friedman and Polson, 1981; Wickens, 1984). Based on an expansive systematic review of the interference pattern in the dual-task data available at the time, Wickens (1980, 1987) proposed a multiple resource model. According to this model, attentional resources are defined along three dichotomous dimensions: (1) stages of processing with perceptual/central processing requiring resources different from those used for response processing, (2) processing codes with spatial processing requiring resources different from those used for verbal processing, and (3) input/output modalities with visual and auditory processing requiring different processing resources and manual and speech responses also requiring different processing resources. A recent revision of the model included a fourth dimension of visual channels, distinguishing between focal and ambient vision (Wickens, 2002, 2008a).

The energetic and specificity aspects of the attentional resources are receiving converging support from subjective (e.g., Tsang and Velazquez, 1996; Rubio et al., 2004), performance (e.g., Tsang et al., 1996; Wickens, 2002), and neurophysiological (e.g., Just et al., 2003; Parasuraman, 2003; Kramer and Parasuraman, 2007) measures. First, parametric manipulation of task demands as well as changes in task priorities have been found to produce systematic and graded changes in the level of subjective workload ratings, performances, and level of neuronal activation. Second, these measures have been found to be sensitive to the competition for specific resource demands. Further, increased neuronal activation associated with different types of processing (e.g., spatial processing and verbal processing) are found to be somewhat localized in different cortical regions. An important application of this model is its prediction of multiple-task performance that is common in many complex work environments (Wickens, 2008a). The higher the similarity in resource demands among the task components, the more severe the competition for similar resources, the less spare capacity, and the higher the level of workload that would result. The other side of the coin is that it would be less feasible to dynamically reallocate resources among task components that utilize highly dissimilar resources. Consequently, the characterization of the processing demand of a task will need to take into account both the intensity aspect (how much resources) and the structural aspect (which resources).

As mentioned above, the supply or availability of processing resources is subject to individual differences such as one's ability and skill level. Just et al. (2003) reviewed a set of behavioral and neurophysiological data that lend support to the notion that a higher level of skill or ability effectively constitutes a larger resource supply. For example, Parks et al. (1988) used a verbal fluency task that required subjects to generate as many words as possible that began with a given stimulus letter. Those with a higher verbal ability were found to exhibit a lower level of positron emission tomography (PET) measures of brain activity. Just et al. proposed that the difference between the more and less proficient subjects lay in the proportion of resources they needed to perform the task. In another study, Haier et al. (1992) found that weeks of practice in the spatial computer game Tetris led to improved performance and a reduced amount of PET-measured activity. Just et al. proposed that practice improved the subjects' procedural knowledge, and the newly acquired, more efficient procedures entailed a lower level of resource use. In practice, a reduced level

of resource requirement by one task would translate to increased spare resources for processing other tasks.

## 2.2 Memory and Situation Awareness

Like the concept of mental workload (and many other psychological concepts, such as intelligence), SA is difficult to define precisely (e.g., Durso and Gronlund, 1999). Pew (1994) defined a situation as "a set of environmental conditions and system states with which the participant is interacting that can be characterized uniquely by a set of information, knowledge and resource options" (p. 18). The most commonly referenced working definition for SA came from Endsley (1990, p. 1–3): SA is "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future." This definition connotes both perception of the now and present and connection with knowledge gained in the past. As Figure 1 denotes, SA is most closely linked to the perceptual and the working memory processes. Certainly, it is not sufficient just for the information relevant to the situation to be available; the information needs to be perceived by the operator, and perception entails far more than the detection of signals or changes. For pattern recognition, object categorization, and comprehension of meaning to occur, contact with knowledge is necessary. But knowledge stored in long-term memory (LTM) s accessible only through short-term or working memory. It is noteworthy that Baddeley (1990) introduced the term working memory to emphasize that short-term memory is far more than a temporary depository for information. Rather, it is an active, effortful process involved in maintaining information and is subject to capacity as well as attentional limits.

Adams et al. (1995) made a distinction between the process and product of SA: "*product* refers to the state of awareness with respect to information and knowledge, whereas *process* refers to the various perceptual and cognitive activities involved in constructing, updating, and revising the state of awareness" (p. 88). To elaborate, the product of SA is a distillation of the ongoing processing of the interchange between information perceived from the now and present (working memory) and knowledge and experience gained from the past (long-term memory). As will be made clear below, this distinction between the process and product of SA has important implications on the interaction of mental workload and SA and on the appropriate assessment techniques.

Just as, given the same objective task demand, mental workload could vary due to individual differences in resource supply as a result of skill and ability differences, given the same situation, SA could vary due to individual differences in knowledge and experience that is amassed in one's LTM. The extant view of the nature of expertise further expounds on the role of memory in determining the content and process of SA.

Expertise is mostly learned, acquired through many hours of deliberate practice (e.g., Glaser, 1987; Chi et al., 1988; Druckman and Bjork, 1991; Adams and Ericsson, 1992; Ericsson, 1996). A fundamental difference between novices and experts is the amount of acquired domain-specific knowledge (e.g., Charness and Tuffiash, 2008). For example, Chase and Ericsson (1982) and Staszewski (1988) reported three people who, after extensive practice, developed a digit span in the neighborhood of 100 numbers. Being avid runners, the subjects associated the random digits to running-related facts that already existed in their LTM (e.g., dates of the Boston Marathon). These subjects had a normal short-term memory span when the studies began and after practice demonstrated the normal span when materials other than digits were tested. Charness (1995) pointed out that such escapes from normal limits also have been observed in perceptual processing. For example, Reingold et al. (2001) found that when chess symbols (as opposed to letters designating chess pieces) were used, highly skilled players could make their decision in some cases without moving their eyes from the initial fixation point at the center of the display. In contrast, weaker players had to make direct fixations. When letter symbols instead of chess piece symbols were used, even the experts were forced to fixate on the pieces directly much more often. Charness pointed out that these observations show that experts can both accurately encode a situation and prepare an appropriate response much more quickly than their less skilled counterparts, but only in the domain of their expertise.

In addition to having acquired declarative knowledge (factual information), experts have a large body of procedural (how-to) knowledge. With practice, many procedural rules (productions) become concatenated into larger rules that can produce action sequences efficiently (Druckman and Bjork, 1991). Importantly, the expertise advantage goes beyond a quantitative difference. The organization of knowledge is fundamentally different between experts and novices. An expert's knowledge is highly organized and well structured, so that retrieving information is much facilitated. In a recent study, Meade et al. (2009) presented expert pilots, novice pilots, and nonpilots with aviation scenarios and had them recall the scenarios alone or in collaboration with a fellow participant at the same expertise level. Whereas the nonexperts were disrupted by collaboration, the experts benefited from it. The benefits were attributed to the expert's superior, highly organized, domain-specific knowledge and their ability to acknowledge and elaborate on contributions to the joint memory performance from others.

Ericsson and Kintsch (1995) further proposed that a long-term working memory (LTWM) emerges as expertise develops and is a defining feature of advanced skill (Ericsson and Delaney, 1998). Whereas working memory has severe capacity and temporal limits, LTWM is hypothesized to have a larger capacity that persists for a period of minutes (or even hours). With an organizational structure that already would have been built in LTM, even very briefly seen, seemingly random, incoming information might be organized similarly. Retrieval cues can then be devised and used to access information in LTM quickly. To illustrate, Ericsson and Kintsch (1995) described the medical diagnosis process that requires one to store numerous individual facts in

working memory. Medical experts were found to be better able to recall critical information at a higher conceptual level that subsumed specific facts and to produce more effective diagnosis. A very important function of LTWM appears to be providing working memory support for reasoning about and evaluating diagnostic alternatives (Norman et al., 1989; Patel and Groen, 1991). That is, expert problem solving is more than just quick retrieval of stored solutions to old problems. Expertise is also associated with effective application of a large amount of knowledge in reasoning to cope with novel problems (Charness, 1989; Horn and Masunaga, 2000).

Finally, experts show metacognitive capabilities that are not present in novices (Druckman and Bjork, 1991). These capabilities include knowing what one knows and does not know, planning ahead, efficiently apportioning one's time and attentional resources, and monitoring and editing one's efforts to solve a problem (Glaser, 1987). In short, experts' large body of organized knowledge enables them to readily see meaningful patterns, make inferences from partial information, constrain search, frame the problem, apprehend the situation, update perception of the current situation continuously, and anticipate future events, including eventual retrieval conditions (Glaser, 1987; Charness, 1995). Importantly, an accurate account of the current situation allows an experienced operator to rapidly retrieve the appropriate course of action directly from memory, enabling swift response execution.

## 2.3  Mental Workload and Situation Awareness

A number of researchers have emphasized that the concepts of mental workload and SA are intricately intertwined (e.g., Wickens, 2002; Vidulich, 2003; Wickens et al., 2008). In this section we attempt to sharpen their distinction and to examine their interactions more closely. Wickens (2001, p. 446) contrasts the two concepts in the following way: "Mental workload is fundamentally an *energetic* construct, in which the quantitative properties ('how much') are dominant over the qualitative properties ('what kind'), as the most important element. In contrast, situation awareness is fundamentally a *cognitive* concept, in which the critical issue is the operator's accuracy of ongoing understanding of the situation (i.e., a qualitative property)." In practice, one assesses the amount and type of workload and the quality (scope, depth, and accuracy) of the content of SA (Vidulich, 2003). Recently, Durso et al., (2007) and Durso and Sethumadhavan (2008) emphasized the need to assess the cognitive processes involved in attaining SA as well.

Both the level of workload and the quality of SA are shaped by exogenous and endogenous factors. Exogenous factors are inherent in the situation (e.g., task demands and situation complexity and uncertainty). Endogenous factors are inherent in a person's ability and skill. The same level of task demands could impose different levels of workload on the operator, depending on her ability or skill level. As discussed above, a high skill level is functionally equivalent to having a larger processing resource supply. A moderate crosswind could be a challenge for a student pilot trying to land a plane but a rather routine task for a seasoned pilot. An overly sensitive warning alarm could be exceedingly disruptive to assessment of the situation by a new operator but could safely be ignored by an experienced operator with intimate knowledge of the workings of the system. Although calibrating the exogenous demands is not always straightforward, their influence on workload is obvious. Less apparent is the endogenous influences on the interplay between the level of workload and the quality of SA.

To the extent that workload is caused, and SA supported, by many of the same cognitive processes, they are enabled by, and subject to the limits of, many of the same processes. The more demanding the task, the more complex the situation and the more "work" is required to get the job done and the situation assessed. By our definition, the higher the level of workload, the more attention is needed for task performance and the less is left for keeping abreast of the situation. The SA process could actually compete with task performance for the limited resource supply, and therefore a high level of workload could lead to poor SA. On the other hand, SA could be improved by working harder (e.g., more frequent sampling and updating of information). That is, a high-level workload is sometimes necessary to maintain a good SA. Thus, a high level of workload could be associated with either a low or high degree of SA (Endsley, 1993). But poor SA may or may not impose more workload. One could simply not be doing the work necessary to attain and maintain SA, and if one is not aware of the dire situation that one is in and takes no action to correct the situation, no additional work would be initiated. Although a low degree of SA is never desirable, an awareness of one's lack of SA could start a course of action that could increase the level of workload in the process of attaining or restoring SA. The ideal scenario is one where a high degree of SA would support more efficient use of resources and thereby produce a low level of workload. In short, mental workload and SA could support each other as well as compete with each other.

Strategic management is proposed to be needed for the balancing act of maintaining adequate SA without incurring excessive workload. Strategic management is also referred to as executive control and is a much discussed topic in the literature. One point of contention is what exactly constitutes executive control, since a host of higher level cognitive functions have been included under its rubric. The coordinating of multiple tasks (including the allocation of limited processing resources), planning, chunking or reorganizing information to increase the amount of materials that can be remembered, and the inhibiting of irrelevant information have all been labeled as part of executive control. As Figure 1 indicates, strategic management is skills based and is highly dependent on one's apprehension of the situation. For example, a beginner tennis player would be content to have made contact with the tennis ball and would not have the spare resources or the knowledge to ponder game strategies. After having mastered the basic strokes (which have become more automatic), however, the strategic component would take on more central importance. But strategic management is not attention free. Even though declarative and procedural knowledge

develops as expertise develops and are used to support performance, there are components in many complex performances that are never automatic. High-performing athletes, chess players, musicians, and command and control officers expend considerable effort to perform at the level that they display.

Recent neurophysiological evidence provides some support for the notion that executive control is a distinct construct and consumes processing resources. Just et al. (2003) pointed out that the executive system is identified primarily with the prefrontal cortex, which does not receive direct sensory input but has widespread connections with a number of cortical areas associated with various types of processing (e.g., spatial and verbal processing). A number of functional magnetic resonance imaging studies have shown a higher level of activation in the prefrontal cortex in (1) a problem-solving task that requires more planning than one that requires less (Baker et al., 1996), (2) a working memory task that requires more updating of a larger amount of information (Braver et al., 1997), and (3) dual-task performance (a semantic category judgment and a mental rotation task) versus single-task performance (D'Exposito et al., Zarahn, and Aguire, 1999). These results show that the activation in the prefrontal cortex varies systematically with task demands. Further, neuropsychological patients with lesions in the frontal lobe show impairments in planning and other higher level cognitive functions (Shallice, 1988).

Returning to Figure 1, strategic management competes directly with all the processes that generate mental workload for processing resources. But strategic management could optimize performance by planning and by smartly allocating the limited resources to the processes that need resources the most to meet system requirements. An efficacious strategic management would, of course, require high-quality SA. In a later section, we discuss potential human factors support (such as display support, automation aids, training) that would improve the potential of attaining this ideal scenario of a high level of SA without an exceedingly high level of workload.

## 3 METRICS OF MENTAL WORKLOAD AND SITUATION AWARENESS

There are three major categories of measures of mental workload and SA based on the nature of the data collected: performance, subjective ratings, and psychophysiological measures. There are several properties that should be considered when selecting measures of cognitive states or activities: sensitivity, diagnosticity, intrusiveness, validity, reliability, ease of use, and operator acceptance (e.g., Tsang and Wilson, 1997; Wickens and Hollands, 2000). In addition, Durso and Alexander (2010) and Tenny et al. (1992) caution that SA measures should be designed and selected based on whether the process or the product of SA is of interest. As outlined below, different measures have different strengths and weaknesses and a thoughtful combination of measures can lead to a more complete picture. Practitioners are encouraged to consult more in-depth coverage of the

metrics to develop a good understanding of the properties of the different measures such that the most appropriate choice(s) can be made (Gopher and Donchin, 1986; O'Donnell and Eggemeier, 1986; Lysaght et al., 1989; Vidulich et al., 1994a; Bryne, 1995; Wickens and Hollands, 2000; Vidulich, 2003; Salmon et al., 2006; Kramer and Parasuraman, 2007; Gawron, 2008).

### 3.1 Performance Measures

System designers are typically most concerned with system performance. Some might say that the workload or SA experienced by an operator can be important only if it affects system performance. Consequently, performance-based measures might be the most valuable to system designers. There are two main categories of performance-based workload measures: primary-task performance and secondary-task performance. SA assessment also has made use of primary-task performance. In addition, SA researchers have often employed recall-based memory probe performance and real-time performance. Although primary-task performance is the measure that is most strongly linked to the system designer's goal of optimizing system performance, Vidulich (2003) pointed out that the secondary-task method of workload assessment and the memory probe method of SA assessment are prototypical measures of the theoretical concepts behind workload and SA.

### 3.1.1 Primary-Task Performance

***Primary-Task Workload Assessment*** The primary-task method of workload assessment consists of monitoring the operator's performance of interest and noting what changes occur as task demands are varied. Some common measures are accuracy, response times, and signal detection performance. Some examples of more domain-specific measures are movement time for a specific computer mouse design, average brake time to traffic lights while texting, and deviation from a command altitude in a flight task. The primary-task methodology is grounded in the framework presented above. Since human operators have a finite capacity to deal with the task demand, as that task demand continues to increase, task performance would be expected to deteriorate as the task demand exceeds resource availability. For example, an automobile driver might have more difficulty maintaining a proper course as the weather becomes windy, and if the wind increases even more when the road is slippery, the driver may fail completely to keep the car in the proper lane.

It should be noted that mental workload is not the only factor that can influence operator performance (Gopher and Donchin, 1986; Wickens and Hollands, 2000). For example, direct performance measures often do not reflect variation in resource investment. One could incur a higher level of workload by trying harder to eschew performance deterioration as task demand increases. Alternatively, two individuals may produce the same level of performance but experience different levels of workload due to differences in skill or strategies used. They therefore would have different amount of spare capacity for other tasks. Of interest, Salvendy and his colleagues found that including a factor that

reflected a person's skill, attitude, and personality contributed significantly to the predictive value of their projective modeling technique (Bi and Salvendy, 1994; Xie and Salvendy, 2000a,b). In addition, performance could be limited by poor technological interface or by the poor quality of the data available. One does not have to try very hard to read in the dark to produce a poor comprehension score. Although the primary-task performance is, clearly, very important to system evaluators as a test of whether design goals have been achieved, primary-task performance by itself oftentimes does not provide an adequate metric of an operator's mental workload. First, primary-task performance may not be diagnostic of the source of workload. For example, a high error rate can be caused by many possible task or system factors. Second, in highly automated systems whereby the human operator takes on the monitoring and supervisory role rather than that of an active controller, directly observable performance measures may simply not be available.

***Primary-Task SA Assessment***   Despite the problems in using primary-task performance as a workload measure, it has become a common tool for assessing the impact of human–machine interface modifications intended to improve SA. For example, Vidulich (2000) found that it was common for researchers to propose an interface alteration that would improve SA and test it by determining if performance improved when the alteration was in place. The logic behind primary-task performance-based measures of SA is well illustrated by the Andre et al. (1991) study of aircraft cockpit display design. Andre et al. used the pilot's ability to recover from disorienting events as a direct measure of how well the attitude information provided by the cockpit supported the pilots' SA of current and future attitudes of the aircraft. Their results showed that, SA, as measured by flight performance and recovery from disorientation

events, was best maintained with a planar outside-in display among the alternatives studied.

### 3.1.2  Secondary-Task Measures of Workload

The secondary-task measure has been considered the prototypical measure of mental workload (e.g., Ogden et al., 1970; Gopher, 1994; Vidulich, 2003). A system evaluator would usually desire to assess primary-task performance even if it was not being interpreted as a mental workload indicator. In contrast, a secondary task is usually only incorporated in a system assessment for assessing mental workload. The secondary-task technique is considered to be a procedure that is optimally suited to reflect the commonly accepted concept of mental workload described above. Workload is often assessed to determine whether the human operator is working within a tolerable information-processing capacity while performing the required task. It follows logically that if there is unused capacity, the operator could perform another task. For example, it is expected that spare capacity would be very valuable in emergencies or when under stress (Wickens, 2001; Hockey et al., 2003).

With the secondary-task method, the operator is required to perform a second task concurrently with the primary task of interest. It is explained to the operators that the primary task is more important and the primary-task performance must be performed to the best of their ability whether or not it is performed with the secondary task. Operators are to use only their spare capacity to perform the secondary task. Since the primary and secondary tasks would compete for the limited processing resources, changes in the primary-task demand should result in changes in the secondary-task performance as more or less resources become available for the secondary task.

Figure 2*a* illustrates possible changes in the joint performance of the primary and secondary tasks within a
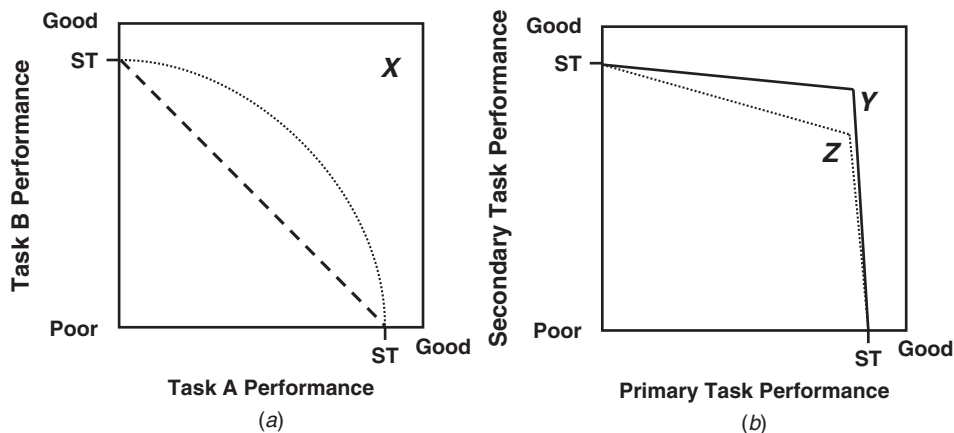


**Figure 2**  (*a*) Hypothetical performance operating characteristic (POC). Tasks A and B are two tasks that have been performed both independently and together. ST = level of single-task performance. The dashed and dotted lines illustrate possible joint performance when the two tasks are performed together. *X*, perfect time-sharing. (*b*) Hypothetical POC. The primary task was performed with two different interfaces. *Y* and *Z*, joint performances observed when the secondary task is performed with the two versions of the primary task.

performance operating characteristic (POC) representation. If the two tasks did not compete for any resources, perfect time-sharing could be observed. This is represented by the "X" on the figure, which would indicate that both tasks were performed at their respective single-task levels even when performed together. Such perfect time-sharing is rare but has been observed (e.g., Allport et al., 1972). The two performance lines in the figure, being much closer to the origin of the graph than the perfect time-sharing point, indicate substantial interference in dual-task conditions. The dashed line shows a perfect performance tradeoff pattern between the two tasks. As one task's performance improves, a comparable degradation is observed in the other. Performance tradeoffs of this sort would typically show up in dual-task studies that manipulate the relative priorities of the two tasks. The POC reflects the subject's allocation strategy for distributing attention between the time-shared tasks as the relative priorities of the two tasks varied. The dotted line shows the two task's joint performance being somewhat better than the perfect tradeoff case. This would be expected to occur if the two tasks required at least some different types of information-processing resources.

Figure 2b illustrates what can be expected in a situation in which the primary task's performance must be defended because of its priority or criticality. For example, as important as it is to communicate with air traffic control, the pilot's task to aviate has paramount priority. Suppose the primary-task performance ($x$ axis) of two possible interfaces, $Y$ and $Z$, is being evaluated. In this hypothetical example, the subjects have done a good job of following the priority instructions and are protecting their primary-task performance, maintaining it at a very high level—near that of the single task. Notice that the primary-task performance with both interfaces is equally high (points $Y$ and $Z$ on the POC). However, interface $Y$'s secondary-task performance is substantially better than interface $Z$'s secondary-task performance. This result would be interpreted as interface $Y$ inflicting less workload on the operator while performing the primary task than would interface $Z$.

An important consideration in the selection of a secondary task is the type of task demand of both the primary and secondary tasks. According to the logic of multiple-resource theory, secondary-task performance will be a sensitive workload measure of the primary-task demand only if the two tasks compete for the same processing resources. The greater the dissimilarity of the resource demands of the time-shared tasks, the lower the degree of the interference there would be between the two tasks. Although a low degree of interference usually translates to a higher level of performance (which of course is desirable), this is not compatible with the goal of workload assessment. A fundamental assumption of the secondary-task method is that the secondary task will compete with the primary task for limited processing resources. It is the degree of interference that is used for inferring the level of workload. Care must therefore be taken to assure that the secondary task selected demands resources similar to those of the primary task. Fortunately, many secondary tasks have been developed and calibrated for use in different evaluations, providing

a database (e.g., Gawron, 2008) for an evaluator to select the secondary tasks that would be most diagnostic of the primary task's resource demands.

The secondary-task measure also offers some practical advantages for workload assessment in comparison to primary-task performance assessment. The secondary-task measure can be assessed in environments where primary-task performance is difficult to obtain or is not available. This is often the case in many operational settings, such as automobiles, ships, and airplanes, that do not have performance-recording capability. Also, with highly automated systems in which the primary role of the operator is that of monitoring and supervising, little observable primary-task performance would be available for analysis. Finally, as noted above, primary-task measures may not be sensitive to very low workload levels because operators could increase their efforts to maintain a stable level of performance (e.g., O'Donnell and Eggemeier, 1986). Adding a secondary task will increase the overall task demand to a level that performance measures may be more sensitive.

One drawback of the secondary-task method is that the addition of an extraneous task to the operational environment may not only add to the workload but also fundamentally change the processing of the primary task. The resulting workload metric would then be nothing more than an experimental artifact. The *embedded secondary-task technique* was proposed to circumvent this difficulty (Shingledecker, 1984; Vidulich and Bortolussi, 1988). With this method, a naturally occurring part of the overall task is used as the secondary task. In some situations, such as piloting a jet fighter aircraft, task shedding is an accepted and taught strategy that is used when primary-task workload becomes excessive. Tasks that can be shed can perhaps serve as naturally lower priority embedded secondary tasks in a less intense workload evaluation situation. However, a naturally lower priority operational task may not always be available. Another drawback is that using the secondary-task method requires considerable background knowledge and experience to properly conduct a secondary-task evaluation and to interpret the results. For example, care must be taken to control for the operator's attention allocation strategy, so as to assure that the operator is treating the primary task as a high priority task (e.g., Damos, 1991). The use of secondary tasks may also entail additional software and hardware development.

Despite the challenges of using the secondary-task procedure, it is still used profitably for system assessment. For example, Ververs and Wickens (2000) used a set of secondary tasks to assess a simulated flight path following and taxiing performance with different sets of head-up display (HUD) symbology. One set of symbology presented a "tunnel in the sky" for the subjects to follow during landing approaches. The other display was a more traditional presentation of flight director information. The tunnel display reduced the subject's flight path error during landing. Subjects also responded more quickly and accurately to secondary-task airspeed changes and were more accurate at detecting intruders on the runway. However, the other display

was associated with faster detections of the runway intruder and quicker identification of the runway. The authors concluded that although the tunnel display produced a lower workload during the landing task, it also caused cognitive tunneling that reduced sensitivity to unexpected outside events. In a study of driving workload, Baldauf et al. (2009) successfully employed a time perception secondary task to distinguish between workload levels inflicted by driving tasks of differing levels of complexity. The length of produced time intervals increased with increasing driving complexity, as did the subjects' electrodermal activity and subjective workload ratings. In a study of office workload, Leyman et al. (2004) used a secondary task to assess the workload associated with simulated office tasks of varying complexity. The subjects in the experiment typed a practiced paragraph as the secondary task that was time-shared with a skilled-based random word memory task of varying list lengths, a more cognitively complex, rule-based geographical reasoning task, or the most cognitively complex, knowledge-based scheduling task. The secondary-task typing errors increased with increasing cognitive complexity. Further, the secondary-task results correlated with the perceived workload and validated the workload assessments of a new electromyographic (EMG) office workload measure. Leyman et al. proposed that this kind of information can be used to better organize work activities in office environments to increase productivity and reduce stress. In another study, Wästlund et al. (2008) used a reaction time secondary task to assess the benefits of better matching the text page layout to the computer screen size on reading comprehension and mental workload. The results indicated that the better layout reduced mental workload while maintaining reading comprehension level.

### 3.1.3 Memory Probe Measures of Situation Awareness

The first popular and standardized procedure for assessing SA was the memory probe technique. It can be considered the prototypical SA measurement tool. The memory probe technique attempts to assess at least part of the contents of memory at a specific time during task performance, so it assesses the product of SA. As represented by the Situation Awareness Global Assessment Technique (SAGAT, Endsley, 1988, 1990), the memory probe procedure consists of unexpectedly stopping the subject's task, blanking the displays, and asking the subject to answer questions to assess his or her knowledge of the current situation. The questions asked are typically drawn from a large set of questions that correspond to experimenter's assessment of the SA requirements for task performance. The subject's answers are compared to the true situation to determine the SAGAT score. Vidulich (2000) found this SAGAT-style approach with unpredictable measurement times and random selection of queries from large sets of possible questions to be generally sensitive to interface manipulations designed to affect SA. In contrast, as memory probes were made more specific or predictable, the sensitivity to interface manipulations appeared to be diminished. For example,

Vidulich et al. (1994b) used a memory probe procedure in which the memory probe, if it appeared, was at a predictable time and the same question was always used. This procedure failed to detect a beneficial effect of display augmentation to highlight targets in a simulated air-to-ground attack, even though there was a significant benefit in task performance (i.e., more targets destroyed) and a significant increase in SA ratings. In contrast, Vidulich et al. (1995) used a SAGAT-like approach with many different questions that were asked during unpredictable trial stoppages. In this case, the memory probe data showed a significant SA benefit of the presence of a tactical situation display.

In another study, Strater et al. (2001) examined U.S. Army platoon leaders in simulated Military Operations on Urbanized Terrain (MOUT) exercises. The platoon leaders varied from relatively inexperienced lieutenants to relatively experienced captains. SAGAT data were collected in a scenario that had the soldiers assaulting an enemy position and a scenario that involved defending a position. SAGAT probe questions were developed that could be used in either scenario. Results showed that the soldiers were more sensitive to different information depending on the scenario type. For example, the soldiers were more sensitive to the location of adjacent friendly forces in the assault scenario but they were more sensitive to the location of exposed friendly elements in the defend scenario. Consistent with the notion of a close link between SA and expertise described above, significant effects of soldier experience level were detected. Experienced soldiers were more sensitive to enemy locations and strength than were inexperienced soldiers. The authors suggested that the data collected from such experimentation could be used to improve training efficacy by identifying better information-seeking behaviors for the novices.

Although the memory probe procedure is attractive due to its assessing the information possessed by the subject at a specific moment in time, it does have practical constraints that limit its applicability. First, it can be highly intrusive to task performance when the task has to be stopped unexpectedly in order to query subject's knowledge of the current state. Although Endsley (1988) has demonstrated that the performance of a simulated air-to-air combat task in trials that included SAGAT stoppages did not significantly differ from trials that did not, the effect such stoppage could have on the cognitive processes involved was not clear. Second, there are assessment environments where such stoppages are impossible (e.g., actual airplane flight tests). Third, the number of questions required to provide an accurate picture of the operator's SA and that must be selected randomly and presented unpredictably can result in a large number of trials being needed.

### 3.1.4 Situation Awareness Real-Time Performance Assessment

Real-time performance has been used as a potential indicator of SA (Durso et al., 1995a; Pritchett and Hansman, 2000; Vidulich and McMillan, 2000). The logic is based on the assumption that if an operator is aware of task demands and opportunities, he will

react appropriately to them in a timely manner. This approach, if successful, would be unintrusive to task performance, diagnostic of operator success or failure, and potentially useful for guiding automated aiding. Since the continuous stream of operator performance is assessed, real-time performance could help shed light on the SA processes involved.

The Global Implicit Measure (GIM; Vidulich and McMillan, 2000) is an example of this approach. The GIM is based on the assumption that the operator of a human–machine system is attempting to accomplish known goals at various priority levels. Therefore, it is possible to consider the momentary progress toward accomplishing these goals as a performance-based measure of SA. Development of the GIM was an attempt to develop a real-time SA measurement that could effectively guide automated pilot aiding (Brickman et al., 1995, 1999; Vidulich, 1995; Shaw et al., 2004). In this approach, a detailed task analysis was used to link measurable behaviors to the accomplishment of mission goals. The goals would vary depending on the mission phase. For example, during a combat air patrol, a pilot might be instructed to maintain a specific altitude and to use a specific mode of the on-board radar, but during an intercept the optimal altitude might be defined in relation to the aircraft being intercepted, and a different radar mode might be appropriate. For each phase, these measurable behaviors that logically should affect goal accomplishment were identified and scored. The scoring was based on the contribution to goal accomplishment. The proportion of mission-specific goals being accomplished successfully according to the GIM algorithms indicated how well the pilot was accomplishing the goals of that mission phase. More important, the behavioral components scored as failing should identify the portions of the task that the pilot was either unaware of or unable to perform at the moment. Thus, GIM scores could potentially provide a real-time indication of the person's SA as reflected by the quality of task performance and a diagnosis of the problem if task performance deviated from the ideal, as specified by the GIM task analysis and scoring algorithms.

Vidulich and McMillan (2000) tested the GIM metric in a simulated air-to-air combat task using two cockpit designs that were known from previous evaluations to produce different levels of mission performance, mental workload, and perceived SA. The subjects were seven U.S. military pilots or weapons systems officers. The real-time GIM scores distinguished successfully between the two cockpits and the different phases of the mission. No attempt was made to guide adaptation on the basis of the GIM scores, but the results suggested that such an approach has promise.

### 3.1.5 Situation Present Assessment Method

Another assessment technique is the Situation Present Assessment Method (SPAM) developed by Durso and his colleagues (e.g., Durso et al., 1995a; Durso and Dattel, 2004). SPAM is especially interesting because it not only combines some of the beneficial features of performance-based SA assessment with aspects of

the memory probe approach but also incorporates a performance-based assessment of mental workload.

SPAM utilizes probe questions as a tool for assessing SA, but unlike the more typical memory probe procedures, SPAM does not remove the participant from the situation to force reliance on the contents of working memory. In SPAM the participant continues to engage in the task and, if unable to reply to the SPAM query based on the current contents of working memory, is able to search the situation to determine the answer. The experimenter then assesses not only the correctness of the participant's response but also the latency to generate the response. The idea behind using the latency as a SA measure is that, if the participant possesses good SA, then the reply will either be based on the current contents of working memory or the participant will know exactly where to search for the needed information; in either case, the response should be relatively quick. On the other hand, if the participant's SA is low, the search for the relevant information will probably be inefficient and relatively slow. SPAM has been demonstrated to be sensitive to expertise differences in Chess (Durso et al., 1995b), automation failures in air traffic control (ATC) simulations (Durso et al., 2001), and individual differences in ATC trainee potential (Durso et al., 1998). In these tests, the SPAM approach not only proved to be sensitive but was also generally unintrusive to primary-task performance.

In addition to the SA measurement provided by SPAM, the technique also incorporates a simple performance-based measure of mental workload. This is done by starting each SPAM query with a warning signal. The participant must acknowledge the signal before the SPAM query is presented. If the participant's momentary workload is high, then it is expected that more time will elapse between the warning cue and the acknowledgment.

Thus, the elegant combination of two latency-based measures into the SPAM technique allows simultaneous assessment of mental workload and SA. This makes the SPAM approach especially attractive for the study of participants' strategies for managing workload, performance, and SA (Durso and Alexander, 2010).

### 3.2 Subjective Measures

Subjective measures consist primarily of using techniques that usually require subjects to quantify their experience of workload or SA. Many researchers are suspicious of subjective data, perhaps as a holdover from the behaviorists' rejection of introspection as a unscientific research method (Watson, 1913). However, Annett (2002a,b) argued that subjective ratings are maligned unfairly. In an in-depth discussion of the issues, he contended that the lack of precision associated with subjective measures was expected to prohibit their use in setting design standards. However, he also concluded that subjective ratings could be useful for evaluating the mechanism underlying performance or for the comparative evaluation of competing interface designs. Such a comparative process is how subjective ratings of workload and SA are typically used.

Vidulich and Tsang (1987) and Tsang and Vidulich (1994) identified three variables that were useful for categorizing subjective rating techniques: dimensionality, evaluation style, and immediacy. *Dimensionality* refers to whether the metric required the subjects to rate their experiences along a single dimension or multiple dimensions. *Evaluation style* refers to whether the subjects were asked to provide an absolute rating of an experience or a relative rating comparing one experience to another. *Immediacy* distinguishes between subjective metrics that were designed to be used as soon as possible after the to-be-rated experience and those that were used at the end of a session or even at the end of an experiment. Although it is theoretically possible to create a subjective technique that combines any level of the three variables, in practice two basic combinations have dominated. The most common techniques combine multidimensionality, the absolute evaluation style, and immediacy. The typical alternative to the multidimensional–absolute–immediate approach are techniques that are usually unidimensional, use a relative comparison evaluation style, and are collected retrospectively rather than immediately.

### 3.2.1 Multidimensional Absolute Immediate Ratings

According to Tsang and Vidulich (1994), the subject's immediate assessment after trial completion should benefit from the freshest memory for the experience of performing the trial while minimizing the potential damaging effects of the operator second guessing her evaluation. The absolute scale design should also encourage the operator to consider the workload of each trial condition individually rather than relatively to other conditions. The multidimensional aspect supports diagnosticity because the subjects can be more precise in describing how experimental conditions influence their experience.

***Workload Ratings***  Although numerous scales have been developed, two popular multidimensional, absolute, and immediate rating scales are the National Aeronautics and Space Administration's Task Load Index (NASA-TLX) (Hart and Staveland, 1988) and the Subjective Workload Assessment Technique (SWAT) (Reid and Nygren, 1988). NASA-TLX is based on six subscales (i.e., mental demand, physical demand, temporal demand, performance, effort, and frustration level), and the ratings on the six scales are weighted according to the subject's evaluation of their relative importance. SWAT is based on three subscales (i.e., time load, mental effort load, and psychological stress load). The weighting of the three subscales is determined by the subject's rankings of the workload inflicted by each combination of the various levels of workload (1–3) in each of the three workload scales. A conjoint analysis is then conducted to produce a look-up table that translates the ordinal rankings to ratings with interval-scale properties. Both NASA-TLX and SWAT ultimately produce a workload rating from 0 to 100 for each trial rated.

NASA-TLX and SWAT have been compared to each other and to a number of other rating scales a number of times (e.g., Battiste and Bortolussi, 1988; Hill et al., 1992; Tsang and Vidulich, 1994; Rubio et al., 2004). In reviewing the comparisons, Rubio et al. (2004) noted that SWAT and NASA-TLX both offer diagnosticity, due to their multiple scales, and have generally demonstrated good concurrent validity with performance. Rubio et al. (2004) also pointed out that both techniques have demonstrated sensitivity to difficulty manipulations, although some researchers have found NASA-TLX to be slightly more sensitive, especially for low levels of workload (e.g., Battiste and Bortolussi, 1988; Hill et al., 1992). On the other hand, Dey and Mann (2010) compared several variants of both NASA-TLX and SWAT to each other in assessing an agricultural sprayer task and found a simplified version of the SWAT to be the most sensitive subjective scale overall.

NASA-TLX and SWAT have also been compared in terms of their ease of use. Each technique consists of two major parts: individual weighting of the subscales and ratings that operators provide after each trial that are then weighted and converted into a final workload score. For the immediate ratings, the SWAT scale requires the operator to choose one of three possible levels for each of the three subscales. The NASA-TLX scale requires the operator to provide a rating between 0 and 100 for each of the six subscales. This makes the SWAT ratings a little easier to collect, especially in a prolonged task, such as flying, while the task performance actually continues. On the other hand, NASA-TLX's paired comparison of the subscales to generate weightings for the subscale is much easier than SWAT's card-sorting procedure for both the subject to complete and the researcher to process. The NASA-TLX procedure only requires the subject to make 15 forced choices of importance between the individual subscales. The raw count of the number of times that each subscale was considered more important than another is then used to weigh the individual subscale ratings provided by the subject. In contrast, the SWAT card sort requires each subject to consider and sort 27 cards (each representing a possible combination of the level of workload for each subscale). Subsequently, specialized software is used to convert the card sort data into an overall workload scale.

Some researchers have investigated simpler methods of generating weights for SWAT. The simple sum of the ratings from the three SWAT subscales has been shown to exhibit the same pattern of findings as SWAT ratings using the original procedure (Biers and Maseline, 1987; Biers and McInerney, 1988; Luximon and Goonetilleke, 2001). Additionally, Luximon and Goonetilleke (2001) found that SWAT sensitivity could be improved by using a continuous scale rather than a three-level discrete scale. As with SWAT, the weighting procedure of NASA-TLX has undergone testing. Both Nygren (1991) and Hendy et al. (1993) have argued that whether the weighting procedure adds any value to NASA-TLX's effectiveness is still ambiguous (e.g., see also Wiebe et al., 2010). Further, according to Nygren (1991), the criterion validity of the measures is not completely clear.

Lee and Liu (2003) provide an example of the use of NASA-TLX to assessing the workload of 10 China Airline pilots flying a Boeing 747 aircraft in a high-fidelity 747 simulator. They found that the overall

NASA-TLX ratings discriminated successfully among four flight segments. As expected, takeoff, landing, and approach were all rated to incur higher workload than cruise. Lee and Liu also used the multidimensional scales of NASA-TLX to diagnose the causes of the higher workload. For example, they found that temporal demand was an important contributor to the takeoff and approach segments, but effort was a more important contributor to landing. The authors concluded that training programs should be designed to help the pilot cope with the specific expected stresses of different flight segments.

***SA Ratings*** Multidimensional, absolute, and immediate ratings have also been a popular approach for assessing SA. Probably the most commonly used subjective rating tool for SA has been the Situation Awareness Rating Technique (SART) developed by Taylor (1990). The SART technique characterizes SA as having three main dimensions: attentional demands (D), attentional supply (S), and understanding (U). The ratings on each of the three dimensions are combined into a single SART value according to a formula (Selcon et al., 1992): SA = U − (D −S). Inasmuch as SART contains ratings of attentional supply and demand, it can be seen to incorporate elements of mental workload in its evaluation but also provides additional distinct information. For example, Vidulich et al. (1995) reported that increased difficulty of a PC-based flight simulation increased the Demand scale on the SART but not the Supply or Understanding scales. Providing additional information increased the Understanding scale but not the Demand or Supply scales. In another study, a direct comparison of NASA-TLX and SART revealed that, although both were sensitive to task demand level, SART was also sensitive to the experience level of the 12 Royal Air Force pilot subjects (Selcon et al., 1991).

### 3.2.2 Unidimensional Relative Retrospective Judgments

The unidimensional relative retrospective judgment approach is based on the assumption that the operator who has experienced all of the task conditions is considered a subject matter expert with knowledge about the subjective experience of performing the various task conditions under consideration. This approach attempts to extract and quantify the operator's opinions about the experiences associated with task performance.

***Workload Judgments*** The use of unidimensional relative retrospective judgments was strongly supported by the work of Gopher and Braune (1984). Inspired by Stevens's (1957, 1966) psychophysical measurement theory, Gopher and Braune adapted it to the measurement of subjective workload. The procedure used one task as a reference task with an arbitrarily assigned workload value. All of the other tasks' subjective workload values were evaluated relative to that of the reference task. The resulting ratings were found to be highly sensitive in a number of studies (e.g., Tsang and Vidulich, 1994; Tsang and Shaner, 1998). In addition, high reliability of these ratings was revealed by split-half correlations of repeated ratings of the task conditions.

Another approach to collecting unidimensional relative retrospective judgments was developed by a mathematician, Thomas Saaty (1980). Saaty's *Analytic Hierarchy Process* (AHP) technique was developed to aid decision making. When applied to workload assessment, the AHP requires operators to perform all pairwise comparisons of all task conditions. These comparisons fill a dominance matrix, which is then solved to provide the ratings for each task condition. Saaty's AHP was originally designed to evaluate all dimensions relevant to a decision and then combine the multiple dimensions to support selection of one option in a decision-making task. However, Lidderdale (1987) demonstrated that a unidimensional version of the AHP could be an effective workload assessment tool and inspired further investigations using the tool. Vidulich and Tsang (1987) compared the AHP to NASA-TLX and a unidimensional absolute immediate rating of overall workload in assessing the workload of selected laboratory tasks. The AHP was found to be both more sensitive and more reliable than the other techniques. Vidulich (1989) compared several methods for converting dominance matrices to the final ratings and used the results to create the Subjective Workload Dominance (SWORD) technique. In one application, Toms et al. (1997) used SWORD to evaluate a prototype decision aid for landing an aircraft. The participating pilots performed landings with and without the decision aid in both low- and high-task-load conditions. Task load was varied by changing the information available to the pilots. Overall, the results showed that the decision aid improved landing performance while lowering mental workload.

Vidulich and Tsang performed a series of studies to examine the various approaches to subjective assessment. Although specific instruments were compared in these studies, the goal was not to determine which instrument was superior. Rather, the objective was to determine the assessment approach that can elicit the most and accurate workload information. Tsang and Vidulich (1994) found that the unidimensional relative retrospective SWORD technique with highly redundant pairwise comparisons was superior to a procedure using relative comparisons to a single reference task. Tsang and Velazquez (1996) found that, compared to an immediate absolute instrument, a relative retrospective psychophysical scaling was more sensitive to task demand manipulation and had higher concurrent validity with performance. They also found that a subjective multidimensional retrospective technique, the Workload Profile, provided diagnostic workload information that could be subjected to quantitative analysis. Rubio et al. (2004) confirmed the diagnostic power of the Workload Profile technique. They found the Workload Profile to be more diagnostic than either NASA-TLX or SWAT. Collectively, these studies suggested a relative retrospective approach advantage.

***SA Judgments*** Unidimensional relative retrospective judgments have also been applied to SA assessment. For example, the SWORD workload technique was adapted to measure SA (SA-SWORD; Vidulich and

Hughes, 1991). Vidulich and Hughes used SA-SWORD to evaluate the effect of data-linked information in an advanced fighter aircraft simulation. The technique demonstrated good sensitivity to the experimental manipulation and good reliability. Toms et al. (1997) used SA-SWORD to assess SA along with SWORD to measure workload. Their results showed that a decision aid's benefits to landing performance and mental workload were also associated with improved SA.

## 3.3 Physiological Measures

A host of physiological measures have been used to assess mental workload with the assumption that there are physiological correlates to mental work. The most common measures include cardiovascular (e.g., heart rate and heart rate variability), ocular (e.g., pupil dilation, eye movement measures), and measures of brain activity. The present review focuses on the brain measures because (1) it would seem that brain activity could most directly reflect mental work; (2) in line with our framework that hypothesizes both an intensity aspect and a structural aspect to mental work, many of the brain measures have been demonstrated to be sensitive to parametric manipulation of task demands and to be diagnostic with regard to the types of cognitive demands involved in certain task performance; and (3) there appears to be much potential for applications in the burgeoning field of neuroergonomics (Parasuraman and Rizzo, 2007). While the present review focuses on the brain measures, readers are urged to consult many fine and instructive reviews of nonbrain measures (e.g., Beatty, 1982; Stern et al., 1984; Wilson and Eggemeier, 1991; Jorna, 1992; Mulder, 1992; Backs and Boucsein, 2000; Kramer and Weber, 2000; McCarley and Kramer, 2007).

### 3.3.1 Electroencephalographic Measures

Electroencephalographic (EEG) measures are recorded from surface electrodes placed directly on the scalp and have been shown to be sensitive to momentary changes in task demands in laboratory studies (e.g., Glass, 1966), simulated environments (e.g., Fournier et al., 1999; Gevins and Smith, 2003), and operational settings (e.g., Wilson, 2002b). Spectral power in two major frequency bands of the EEG have been identified as being sensitive to workload manipulations: the alpha (7–14-Hz) and theta (4–7-Hz) bands. Spectral power in the alpha band that arises in widespread cortical areas is inversely related to the attentional resources allocated to the task, whereas theta power recorded over the frontal cortex increases with increased task difficulty and higher memory load (Parasuraman and Caggiano, 2002). Sterman and Mann (1995) reported a series of EEG studies conducted in simulated and operational military flights. A systematic decrease in power in the alpha band of the EEG activity was observed with a degraded control responsiveness of a T4 aircraft. A graded decrease in the alpha band power was also observed as U.S. Air Force pilots flew more difficult in-flight refueling missions in a B2 aircraft simulator. In another study, Brookings et al. (1996) had Air Force air traffic controllers perform computer-based

air traffic control simulation (TRACON). Task difficulty was manipulated by varying the traffic volume (number of aircraft to be handled), traffic complexity (arriving to departing flight ratios, pilot skill, and aircraft types), and time pressure. Brookings et al. found the alpha power to decrease with increases in traffic complexity and the theta power to increases with traffic volume. Recently, Dussault et al. (2005) compared EEG and heart-based physiological workload measures for sensitivity to the demands of a simulated flight task with both expert and nonexpert pilots. They found the heart-based measures to be insensitive to the different simulated flight segment, although expert pilots generally had a lower heart rate than nonexpert pilots. In contrast, the EEG measures showed both a lower level of activation for the expert pilots and distinguished between different flight segments. In a study using laboratory tasks, Berka et al. (2007) collected EEG measures via a wireless headset while subjects performed a battery of vigilance, spatial, verbal, and memory tasks. Using different algorithms, Berka et al. derived an EEG metric for task engagement (that tracks demands for sensory processing and attention resources) and one for mental workload (that tracks demands of executive control). The results showed that the EEG engagement metric reliably reflects demands for information gathering, visual processing, and allocation of attention whereas the EEG workload metric reflects demands for working memory, information integration, problem solving, and reasoning, Further, the EEG measures were found to correlate with subjective ratings and performance measures that included accuracy and reaction time.

Kramer and Parasuraman (2007) point out that, in addition to being able to provide a somewhat precise temporal index of changes in alertness and attention, an important advantage of EEG measures is that they can be recorded in the absence of discrete stimuli or responses, which make them particularly useful in situations in which an operator is monitoring slowly changing displays that requires minimal intervention. One drawback of the EEG measures is their sensitivity to numerous artifacts such as head and body movements that pose special difficulties for extralaboratory applications. While Kramer and Weber (2000) were concerned with the diagnosticity of the EEG measures, recent research such as those of Berka et al. (2007) has made some progress in using EEG measures to distinguish the different types of demands incurred. More in-depth discussion can be found in Gevins et al. (1995) and Pizzagalli (2007).

### 3.3.2 Event-Related Potentials

Evoked potentials are embedded in the EEG background and are responsive to discrete environmental events. There are several different positive and negative voltage peaks and troughs that occur 100–600 ms following stimulus presentation. The P300 component of the event-related potential (ERP) has been extensively studied as a mental workload measure (e.g., Gopher and Donchin, 1986; Parasuraman, 1990; Wickens, 1990; Kramer and Parasuraman, 2007). The P300 is often examined in a dual-task condition with either the oddball

paradigm or the irrelevant probe paradigm. In the *oddball paradigm*, the P300 is elicited by the subject keeping track of an infrequent signal (e.g., counting infrequent tones among frequent tones). One drawback of this paradigm is that the additional processing of the oddball and having to respond to it could inflate the true workload of interest artifactually. As an alternative, additional stimuli are presented, but subjects are not required to keep track of them in the *irrelevant probe paradigm*. But the irrelevant probe paradigm would work only if the irrelevant probes were presented in a channel that would be monitored anyway (Kramer and Weber, 2000)

With the oddball paradigm, the amplitude of the P300 has been found to decrease with increased task difficulty manipulated in a variety of laboratory tasks (e.g., Hoffman et al., 1985; Strayer and Kramer, 1990; Backs, 1997). Importantly, P300 is found to be selectively sensitive to perceptual and central processing demands. For example, Isreal et al. (1980) found that the amplitude of P300 elicited by a series of counted tones was not sensitive to manipulations of the response-related demand of the concurrent tracking task but was affected by manipulations of display perceptual load of the concurrent monitoring task. Many of the laboratory-based findings have been replicated in simulator studies. For example, Kramer et al. (1987) had student pilots fly an instrument flight plan in a single-engine aircraft simulator. The P300s elicited by the secondary tone-counting task decreased in amplitude with increasing turbulence and subsystem failures (see also Fowler, 1994). Using the irrelevant probe paradigm, Sirevaag et al. (1993) had senior helicopter pilots fly low-level, high-speed flight in a high-fidelity helicopter simulator. The P300 amplitude was found to increase with increased difficulty in the primary tracking task. In addition, the P300 amplitude elicited by the secondary irrelevant probes decreased with increased communication load.

A recent modeling effort further validates the usefulness of P300 as a real-time mental workload metric. Wu et al. (2008) presented a queuing network modeling approach that was demonstrated to be able to account for changes in the P300 amplitude and latency as task demands varied. The researchers proposed that the queuing network model could be a candidate tool for supporting adaptive automation because of its ability to assess mental workload in real time as well as to predict workload both in the temporal dimension (as reflected by the P300 latency) and the intensity dimension (as reflected by the P300 amplitude).

In short, ERP measures have been found to be sensitive to, and diagnostic of, changes in the perceptual and central processing task demands. Because of the sensitivity of ERP measures to the temporal aspects of neuronal activity, they may be a particularly good candidate tool for assessing workload dynamically. As with the EEG measures, ERP measures are subject to problems with motion artifacts and electrical noise. Another potential drawback is the possibility of artifactually augmenting the true workload of interest if the ERP measures are elicited from a secondary task. As with the performance measures, it would be ideal if a secondary

task naturally embedded in the test environment could be used. But because ERP signals are relatively small and ensemble averaging across many stimuli is necessary for meaningful interpretation, there are not always sufficient stimuli available from the embedded secondary task or the primary task. See Fabiani et al. (2007) and Kramer and Parasuraman (2007) for a more in-depth discussion.

### 3.3.3 Cerebral Blood Flow Measures

Cerebral blood flow measures are based on the principle that neuronal activity associated with mental processing could be assessed by measuring the blood flow responses of the brain (Kramer and Parasuraman, 2007). Three measures are considered here: positron emission tomography (PET), functional magnetic resonance imaging (fMRI), and Transcranial Doppler Sonography (TCD). The PET and fMRI measures have been used to localize cortical regions associated with various cognitive processing (e.g., D'Exposito et al., 1999; Posner and DiGirolamo, 2000) and could provide diagnostic information with regard to the type of task demand incurred. Recent research has shown that all three measures are sensitive to systematic variations with parametric manipulation of task difficulty (Kramer and Parasuraman, 2007).

As an example, Corbetta et al. (1990) had subjects determine whether two stimuli presented in two frames separated by a blank display were the same or different. Between the two frames, the stimuli could vary in one of three dimensions: shape, color, and velocity. In the selective-attention condition, one of the dimensions would be designated as the relevant dimension, and zero, one, or two irrelevant dimensions could covary with the relevant dimension. In the divided-attention condition, any one of the dimensions could vary. The behavioral data (*d'*) indicated that the divided-attention condition was more difficult. In the selective-attention condition, increased blood flow as revealed by PET scans was observed in the region of the visual cortex known to be related to the processing of the relevant dimension designated. Corbetta et al. proposed that the increased neuronal activity in the specialized regions for the different dimensions was a result of top-down attentional control since the sensory information should be the same across selective- and divided-attention conditions.

Just et al. (2003) reviewed a series of PET studies and found lower brain metabolic rate to be associated with higher language proficiency (Parks et al., 1988) and increased practice with a spatial computer game (Haier et al., 1992). Just et al. interpreted these results to mean that high-ability, high-skill persons could process more efficiently, thereby requiring a smaller amount of their total amount of processing resources available. They effectively would have a larger supply of processing resources (for other processing). Just and Carpenter (1992) proposed that the computational work underlying thinking must be accompanied by resource utilization. In their 3CAPS (Capacity-Constrained Concurrent Activation-based Production System) model, a brain region is considered a resource pool. Computational activities are resource consuming in the sense that

they all operate by consuming an entity called activation. The intensity and volume of brain activation in a given cortical area are expected to increase in a graded fashion with increased computational load. Indeed, Just et al. (1996) found that with increasing sentence complexity the level of neuronal activation and the volume of neural tissue activated increased in four cortical areas associated with language processing (Wernicke's, Broca's, and their right-hemisphere homologues). With a spatial mental rotation task, Carpenter et al. (1999) found a monotonic increase in signal intensity and volume activation in the parietal region as a function of increased angular disparity between the two stimuli whose similarity was to be judged.

Reviewing the results from a number of studies that use an array of behavioral and neurophysiological measures (ERPs, PET, and fMRI), Just et al. (2003) came to a similar viewpoint adopted in the present chapter in that cognitive workload is hypothesized to be a function of resource consumption and availability. Several similarities between the 3CAPS model and Wickens's multiple-resource model are apparent. According to both models, (1) mental workload is a function of the supply and demand of processing resources, (2) resources can be modulated in a graded fashion, (3) specific resources are used for different types of cognitive processing (e.g., verbal and spatial task demands bring about activations in different cortical regions), and (4) supply or availability of resources can be modulated by individual differences in ability and skill or expertise.

The PET and fMRI studies described so far were all conducted in the laboratory. Although more studies conducted in the operational settings would certainly be desirable, the equipment required for their measurements makes it impractical. Notwithstanding, one simulated study on pilot performance can be presented. Pérès et al. (2000) had expert (with at least 3000 flight hours and flight instructor qualifications) and novice (with less than 50 flight hours) French Air Force pilots perform a continuous simulated flight control task at two speeds (100 and 200 knots) while fMRI measures were collected. The fMRI measures showed that neuronal activation was dominant in the right hemisphere, as would be expected for a visual spatial task. Further, novice pilots exhibited more intense and more extensive activation than expert pilots. In the high-speed condition, the expert pilots exhibited increased activation in the frontal and prefrontal cortical areas and reduced activity in visual and motor regions. This suggested that the expert pilots were better able to use their knowledge to focus their resources for the higher level functions in working memory, planning, attention, and decision making. In contrast, novice pilots' increased activation in the high-speed condition was more widespread and extended across the frontal, parietal, and occipital areas, suggesting that they were engaged in nonspecific perceptual processing. Interestingly, when the expert pilots were asked to track at an even higher speed (400 knots), their pattern of activation resembled that of the novice pilots tracking at 200 knots.

The TCD technique uses a transducer mounted on the subject's head that directs ultrasound waves toward an artery within the brain. The application of TCD in mental workload assessment is relatively recent. Still, a number of studies have shown systematic increase in blood flow in the middle cerebral arteries (MCAs) with increased task demands. For example, Serrati et al. (2000) observed increased blood flow with increased difficulty of a mental rotation task. Frauenfelder et al. (2004) observed increased blood flow with increased complexity of a planning task. Wilson et al. (2003) found that blood flow associated with performance of the Multi-Attribute Task Battery varied with task difficulty.

Notably, there is a paucity of physiological studies on SA included in this chapter. This is partly because, compared to the concept of mental workload, the concept of SA is relatively new (Pew, 1994; Wickens, 2001) and both its theoretical and methodological developments have not reached the level of maturity that the concept of mental workload has. It is also the case that the concept of SA is not associated with a specific process. Although complex performance generally entails multiple processes, it is often possible to identify many of the processes and hence the type of workload involved. In contrast, whereas SA is supported by many of the same processes, SA is an emergent property that has not been hypothesized to be associated with specific cortical regions or other physiological responses.

The neurophysiological measures discussed here have all demonstrated to show sensitivity to task demands. Some of them are able to provide a continuous measure for online assessment (e.g., EEG, ERP, TCD). Some could provide valuable diagnostic information with regard to the type of cognitive demands entailed (e.g., ERPs, PET, fMRI). Most of them are not cognitively intrusive, as in having to perform additional work in order to provide a measure. However, some are particularly vulnerable to interference from motion and electrical noise. The main drawback with many physiological measures is that they are equipment intensive and they require specialized expertise to collect, analyze, and interpret properly. This makes assessment of some of them (e.g., PET and fMRI) in the operational settings impractical. But it is not impossible. Even for the costly fMRI studies, attempts have been made to assess the mental workload of simulated flight performance (Pérès et al., 2000). One would expect that many of the physiological measures would become more feasible with the present rapid technological advances (e.g., see Gevins et al., 1995; Wilson, 2002a,b; Parasuraman, 2003; Kramer and Parasuraman, 2007). Readers are encouraged to consult additional sources for more technological details of the neurophysiological techniques (e.g., Kramer and Parasuraman, 2007; Tripp and Warm, 2007).

## 3.4 Multiple Measures of Workload and Situation Awareness

There are several facets to the undertaking of assessing workload and SA of a complex, dynamic human–machine system. First, there are a number of candidate measures to choose from, each with strengths and weaknesses. Measures that provide global information

about the mental workload of these tasks may fail to provide more specific information about the nature of the demand. Measures that could provide more diagnostic information may be intrusive or insensitive to other aspects of interest, and certain sensitive measures may be collected only under restrictive conditions. Still, many workload measures often associate. For example, many subjective measures have been found to correlate with performance (e.g., Tsang and Vidulich, 1994; Hockey et al., 2003; Rubio et al., 2004). Just et al. (2003) present a convincing account of how the associations among a number of behavioral and neurophysiological measures support the extant understanding of many cognitive concepts relevant to mental workload (see also Wickens, 1990; Fournier et al., 1999; Lee and Liu, 2003). Importantly, when the measures do not associate, they do not do so in haphazard ways. The association and dissociation patterns among measures should therefore be evaluated carefully rather than treated as unreliable randomness. Below we discuss in greater detail the dissociation between the subjective and performance workload measures and the relation between the workload and SA measures.

### 3.4.1  Dissociations among Workload Measures

When different types of workload measures suggest different trends for the same workload situation, the workload measures are said to *dissociate*. Given that mental workload is a multidimensional concept, and that various workload measures may be differentially sensitive to the different workload dimensions, some dissociations among workload measures are to be expected. Measures having qualities of general sensitivity (such as certain unidimensional subjective estimates) respond to a wide range of task manipulations but may not provide diagnostic information about the individual contributors to workload. Measures having selective sensitivity (such as secondary-task measures) respond only to specific manipulations. In fact, the nature of the dissociation should be particularly informative with regard to the characteristics of the workload incurred by the task under evaluation.

Several conditions for the dissociation of performance and subjective measures have been identified (Vidulich and Wickens, 1986; Vidulich, 1988; Yeh and Wickens, 1988):

1. Dissociation tends to occur under low-workload conditions (e.g., Eggemeier et al., 1982). Performance could already be optimal when the workload is low and thus would not change further with additional effort that would be reflected in the subjective measures.

2. Dissociation would occur when subjects are performing data-limited tasks (when performance is governed by the quality of the data rather than by the availability of resources). If subjects are already expending their maximum resources, increasing task demand would further degrade performance but would not affect the subjective ratings.

3. Greater effort would generally result in higher subjective ratings; however, greater effort could also improve performance (e.g., Vidulich and Wickens, 1986).

4. Subjective ratings are particularly sensitive to the number of tasks that subjects have to time-share. For example, performing an easy dual task (that results in good performance) tends to produce higher ratings than does performing a difficult single task (that results in poor performance) (e.g., Yeh and Wickens, 1988).

5. Performance measures are sensitive to the severity of the resource competition (or similarity of resource demand) between the time-shared tasks, but subjective measures are less so (Yeh and Wickens, 1988).

6. Given that subjects only have access to information available in their consciousness (Ericsson and Simon, 1993), subjective ratings are more sensitive to central processing demand (such as working memory demand) than to demands that are not represented well consciously, such as response execution processing demand. Dissociation would therefore tend to occur when the main task demands lie in response execution processing (Vidulich, 1988). McCoy et al. (1983) provided an excellent list of realistic examples of how performance and subjective ratings may dissociate in system evaluations and discussed how the dissociations can be interpreted in meaningful ways.

Hockey (1997) offers a more general conceptual account for the relations among performance, subjective, and physiological measures. Hockey proposes a compensatory control mechanism that allocates resources dynamically through an internal monitor very much like the one proposed by Kahneman (1973). Performance may be protected (as in primary-task performance) by strategic recruitment of further resources, at the risk of incurring increased subjective effort, physiological costs, or degraded secondary-task performance. Alternatively, performance goals may be lowered. Although performance may then be lower, no additional effort or physiological cost will be incurred. Hockey emphasizes that the efficacy of the control mechanism hinges on the accuracy of the perception of the situation. For example, Sperandio (1978) found air traffic controllers to switch strategy when the traffic load increased. Beyond a certain number of aircraft that the controllers handled, controllers would switch to a uniform strategy across aircraft as opposed to paying more individual attention to the various aircraft. Although this strategy should reduce the cognitive resources needed for dynamic planning, it would also probably produce less optimal scheduling. That is, the primary-task performance might have been preserved with the strategy switch, but some secondary goals would have suffered (Hockey, 1997).

A recent study by Horrey et al. (2009) further illustrates the importance of understanding the potential dissociations among workload measures. Drivers of

varying age and experience drove an instrumented vehicle on a closed-loop track while simultaneously performing an additional in-vehicle task. One of the in-vehicle tasks was a more engaging task that was game-like and involved back-and-forth information exchange with the experimenter. Another in-vehicle task was a running addition task presented at a fixed interval. Driving performance as well as the performance of both the in-vehicle tasks degraded in the dual-task condition from the single-task condition. Driving performance was degraded more with the more engaging task than with the addition task. However, neither subjective workload ratings nor subjective performance estimates differentiated the two conditions. These results suggested that the subjects were insufficiently sensitive to the competition for resources between driving and the more engaging task, posting important implications for road safety.

### 3.4.2 Relations of Workload and Situation Awareness Measures

Wickens (2001) pointed out that, due to the energetic properties of workload, many physiological and subjective rating measures are suited for capturing the quantitative aspects of workload. In contrast, physiological measures are likely to be poor candidates for assessing the quality or content of SA. Further, self-ratings of one's awareness are unlikely to be informative since one cannot be aware of what one is not aware. However, subjective SA ratings could still be useful if they are used for system evaluative purposes. As illustrated earlier, subjects often could indicate reliably which system design affords greater SA. Last, Wickens pointed out that explicit performance measures designed to examine what one is aware of (content of SA) have no parallel use for workload assessment. However, implicit performance measures such as those used to check for reaction to unexpected events can be used to assess both workload and SA.

As discussed earlier, the relationship between workload and SA is multifaceted. Although high SA and an acceptable level of workload are always desirable, workload and SA can correlate positively or negatively with each other, depending on a host of exogenous and endogenous factors. Three sample studies will be described to illustrate their potential relationships. Vidulich (2000) reviewed a set of studies that examined SA sensitivity to interface manipulations. Of the nine studies that manipulated the interface by providing additional information on the display, seven showed an increase in SA, four showed a concomitant reduction in workload, and three showed a concomitant increase in workload. In contrast, of another nine studies that manipulated the interface by reformatting the display, all nine showed an increase in SA, six showed a concomitant reduction in workload, and none showed an increase in workload. In short, although different patterns in the relationship between the workload and SA measures were observed, the various patterns were reasonably interpretable given the experimental manipulations. In another study, Alexander et al. (2000) examined the relationship between mental workload and SA in a simulated air-to-air combat task. Seven pilots flew simulated air intercepts against four bombers supported by two fighters. The main manipulations were two cockpit designs (the conventional cockpit with independent gauges and a virtually augmented cockpit designed by a subject matter expert) and four mission phases of various degrees of difficulty and complexity. A negative correlation between the workload and SA measures were observed for both the cockpit design and the mission complexity manipulations. The augmented cockpit improved SA and reduced workload, whereas increased mission complexity decreased SA and increased workload. Perry et al. (2008) performed an ambitious evaluation of the effects of increasing physical demands on the performance of a rule-based helicopter load planning task, a memory-probe SA measure, and NASA-TLX workload ratings. All measures were collected as subjects performed the task either standing, walking, or jogging on a treadmill. Although subjects were able to maintain performance on the cognitive helicopter loading task, increasing physical demands both decreased SA and increased workload ratings, showing the costs the physical demands were exacting from the subjects. The findings of these studies underscore the value of assessing both the mental workload and SA involved in any test and evaluation.

### 3.4.3 Need for Multiple Measures

There are several broad guiding principles that would be helpful in measures selection. Muckler and Seven (1992) held that "the distinction between 'objective' and 'subjective' measurement is neither meaningful nor useful in human performance studies" (p. 441). They contended that all measurements contain a subjective element as long as the human is part of the assessment. Not only is there subjectivity in the data obtained from the human subject, the human experimenter also imparts his or her subjectivity in the data collection, analysis, and interpretation. Thus, performance measures are not all objective, nor are subjective measures entirely subjective (see also, Annett, 2002a,b; Salvendy, 2002). Muckler and Seven advocated that the selection of a measure (or a set of measures) be guided by the information needs. Candidate measures can be evaluated by considering their relative strengths (such as diagnosticity) and weaknesses (such as intrusiveness). In addition, Kantowitz (1992) advocated using theory to select the measures. He made an analogy between theory and the blueprint of a building. Trying to interpret data without the guidance of a theory is like assembling bricks randomly when constructing a building. To elaborate, Kantowitz pointed out that an understanding of both the substantive theory of human information processing and the psychometric theory of the measurements is helpful. The former dictates what one should measure, and the latter suggests ways of measuring them. Another useful (if not required) strategy is to use multiple measures as much as feasible. As discussed above, even seemingly dissociated measures are informative (and sometimes especially so) if one is cognizant of the idiosyncratic properties of the different measures. In fact, Wickens (2001) pointed out that converging evidence from multiple measures

is needed to ensure an accurate assessment of the level of workload incurred and the quality of SA attained. This need is also reflected in the literature, that many, if not the majority, of the studies now do use multiple measures in their workload or SA assessment (e.g., Brookhuis and de Waard, 2010; Lehrer et al., 2010).

To emphasize the value of assessing multiple measures, Parasuraman (1990) reported a study that examined the effectiveness of safety monitoring devices in high-speed electric trains in Europe (Fruhstorfer et al., 1977). Drivers were required to perform a secondary task by responding to the occurrence of a target light in a cab within 2.5 s. If no response was made, a loud buzzer would be activated. If the buzzer was not responded to within an additional 2.5 s, the train's braking system was activated automatically. Over a number of train journeys, onset of the warning buzzer was rare, and the automatic brake was activated only once. However, the EEG spectra showed that the secondary-task performance could remain normal even when the drivers were transiently in stage 1 sleep.

## 4 DESIGN FOR MENTAL WORKLOAD AND SITUATION AWARENESS: INTEGRATED APPROACH TO OPTIMIZING SYSTEM PERFORMANCE

It is probably fair to say that after a decade of debate there is now a general agreement that mental workload and SA are distinct concepts and yet are intricately intertwined. Both can be affected by very many of the same exogenous and endogenous factors and have a significant impact on each other and on system performance. One implication is that fairly well understood psychological principles can be applied to both concepts. For example, in the framework presented above, both workload and SA are subject to attentional and memory limits, and both can be supported by expertise. There exists an established body of knowledge about the effects of these limits and the enabling power of expertise to allow fairly reliable performance predictions. But the fact that the two concepts are distinct also means that they each contribute uniquely to the functioning of a human–machine system. Below we review several research areas that could be exploited for developing support that would manage workload and SA cooperatively to optimize system performance.

### 4.1 Transportation

Although much of the early work in mental workload and SA was conducted in the aviation domain, recent developments in the light-jet air taxi and NextGen that we reported at the beginning of the chapter illustrates the continued need for workload and SA considerations in aviation system development and evaluation. This is also evident in the emerging field of uninhabited aerial vehicles (UAVs). Given the physical separation between the human operator and the actual system being controlled and the common practice of having a single operator controlling multiple UAVs simultaneously, the UAV operation is typically characterized by heavy reliance on automated systems and poses unique challenges for the

cognitive capabilities of the operator. Not surprisingly, mental workload and SA have figured prominently in UAV-related human factors discussions (e.g., Parasuraman etal., 2009; Prewett et al., 2010).

In ship control, Hockey et al. (2003) used a PC-based radar simulator to examine the cognitive demands of collision avoidance. Subjective workload ratings and secondary-task performance responded similarly to the relative demands of the different collision threats and traffic density, providing valuable information for design of collision avoidance systems. In a recent study, Gould et al. (2009) used a high-fidelity simulator to examine the effects of different navigation aids on high-speed ship navigation while fast patrol boat navigators underwent up to 60 h of sleep deprivation. The navigators were provided with either standard paper charts or a modern electronic chart display and information system (ECDIS). Performance, subjective, and physiological workload measures were collected. Interestingly, some of the workload metrics dissociated from each other and responded to the different experimental manipulations differently. Secondary-task performance was degraded by the sleep deprivation but unaffected by the chart manipulation. But despite improved ship-handling performance with the ECDIS, subjective ratings and heart rate variability indicated that it incurred more workload. Obviously, any single measure would have provided an incomplete picture of what was happening to the ship navigators in this experiment

If the public, special interests groups, and policymakers cannot be convinced of the potential danger of talking or texting on the cell phone while driving, it is not due to a lack of confirming experimental and epidemiological data. Collet et al. (2010a,b) compiled an extensive review of the phoning-while-driving literature. The experimental studies included an array of measures of primary-task measures such as lane-keeping performance in a simulated driving task, secondary-task reaction time to signal detection, subjective workload and SA ratings, and physiological measures. Disturbingly, dissociations between performance and subjective ratings were sometimes observed, suggesting that subjects failed to account for the added demands in the dual-task conditions (Horrey et al., 2009). Collet et al. concluded that concurrent use of cell phones while driving generally has a negative impact upon safety. However, they found that certain variables had larger effects than others. For, example, poor road conditions, demanding conversation content, youth, and inexperience all exacerbated the magnitude of interference between phoning and driving. Counterintuitively, hands-free phones offer no special benefits over hand-held phones. Just et al. (2008) explained that the deterioration in driving performance associated with phoning results from competition for central cognitive resources rather than for motor output. This assertion was further supported by their results from a dual-task study in which subjects drove a simulated car along curving roads while judging spoken sentences to be either true or false. The dual-task scenario degraded driving accuracy and fMRI data revealed that the parietal lobe activation associated with spatial processing decreased by 37% as compared

to the single-task scenario. These results showed that language comprehension (as would be required in a cell phone conversation) could draw mental resources away from driving even when it did not require holding or dialing a phone. Collet et al. (2010a) further pointed out that cell phone use is not the only driver distraction, given the proliferation of in-vehicle technologies, each device should be evaluated for potentially producing an unacceptable level of driver distraction.

In all three of the reviewed domains it can be seen that mental workload and SA considerations have strong potential contributions. The cognitive impacts of new technologies in emerging domains, such as UAVs, and even very familiar domains, such as driving, mandate that the human factors community develop and exploit a full battery of cognitive concepts and measurement tools to ensure maximum safety and effectiveness. Although considerable progress has been made in the areas of mental workload and SA theories and measurement over the years, there is still room for improvement. For example, the development of improved psychophysiological assessment tools would be a great boon for investigating mental workload and SA in the complex, real-world transportation environment.

## 4.2   Adaptive Automation

Automation is often introduced to alleviate the heavy demand on an operator or to augment system performance and to reduce error. Many modern complex systems simply cannot be operated by humans alone without some form of automation aids. However, it is now recognized that automation often redistributes, rather than reduces, the workload within a system (e.g., Wiener, 1988; Lee and Moray, 1994; Casner, 2009). Further, an increasing level of automation could distance the operator from the control system (e.g., Adams et al., 1991; Billings, 1997). The upshot of this is that, even if automation reduces mental workload successfully, it could reduce SA and diminish an operator's ability to recover from unusual events. The idea of adaptive automation was introduced as a means of achieving the delicate balance of a manageable workload level and an adequate SA level. This idea has been around for some time (e.g., Rouse, 1977, 1988) and is receiving much attention in recent research (e.g., Rothrock et al., 2002; Parasuraman and Bryne, 2003; Scerbo, 2007). Proponents of adaptive automation argue that static automation that entails predetermined fixed task allocation will not serve complex dynamic systems well. Workloads can change dynamically due to environmental and individual factors (e.g., skill level and effectiveness of strategies used). It has been proposed that a major environmental determinant of workload is rapid (Huey and Wickens, 1993) and unexpected (Hockey et al., 2003) changes in task load. So, ideally, more or fewer tasks should be delegated to automation dynamically. More automation would be introduced during moments of high workload, but as the level of workload eases, more tasks would be returned to the operator, thereby keeping the operator in the loop without overloading the person. The key issue is the development of an implementation algorithm that could efficaciously adapt the level of automation to the operator's state of workload and SA.

Parasuraman and Bryne (2003) described several adaptation techniques that rely on different inputs to trigger an increase or decrease in the extent of automation in the system. These techniques mostly use physiological or performance indices that afford fairly continual measures. One advantage of basing the adaptive automation algorithm on physiological measures is their noninvasive nature. A number of physiological measures have been evaluated for their potential to provide real-time assessment of workload. They include eye movement measures (Hilburn et al., 1997; McCarley and Kramer, 2007), heart rate variability (e.g., Jorna, 1999), EEG (e.g., Gevins et al., 1998; Prinzel et al., 2000; Gevins and Smith, 2003; Berka et al., 2007; Kohlmorgen, et al., 2007), ERP (Parasuraman, 1990; Kramer et al., 1996; Wu et al., 2008), and TCD (Tripp and Warm, 2007). While Coffey et al. (2010) concluded that the use of neurophysiological signals for direct system control is likely to be inherently limited by the information content and quality of the signals, they see greater promise in using these signals for real-time adaptive automation to aid operators at an appropriate level to keep workload manageable.

Although physiological-based measures are the dominant measures used in adaptive automation, recent studies have demonstrated the potential promise of subjective and performance measures as well. For example, Vidulich and McMillan (2000) propose that the Global Implicit Measure (GIM, described above) could be developed as a real-time SA measurement that could guide effective automated pilot aiding based on real-time scoring of both continuous and discrete tasks. In one performance study, Kaber and Riley (1999) used a secondary monitoring task along with a target acquisition task. Adaptive computer aiding based on secondary-task performance was found to enhance primary-task performance. In a more recent study, Kaber et al. (2006) again used a secondary monitory task and found the adaptive automation to be differentially effective in supporting SA performance of different information-processing tasks in a low-fidelity ATC-related simulation. Lower order functions such as information acquisition were found to benefit from the automation but not the higher functions such as information analysis. In a simulated reconnaissance mission that involved operators supervising multiple uninhabited air and ground vehicles, Parasuraman et al. (2009) used the changed detection performance as the trigger for adaptive automation. Beneficial effects on performance, workload, and SA were observed. Note that for the purpose of adaptive automation the performance measures used would need to afford fairly continual assessment, a property that not many performance-based measures possess. Also, as discussed above, for the secondary-task methodology to provide useful workload information, the time-shared tasks would need to be competing for some common resources, which of course could add to the workload, as Kaber et al. (2006) had observed.

Of note is that many studies now employ and find the use of multiple measures for deriving the automation algorithm to be superior to using a single measure

(e.g., Haarmann et al., 2009; Hockey et al., 2009; Ting et al., 2010).

## 4.3 Medicine

Another domain in which workload and SA assessment is making in-roads is in medicine, especially within the setting of surgical operations. The parallel between aviation and many branches of medicine is increasingly recognized. Both aviation and medicine make considerable use of advanced technology to meet the demands of their complex, dynamic, and safety-critical operations. To accomplish this, researchers in both aviation and medical human factors acknowledge the necessity to understand and optimize the mental workload and SA of the operators involved. For example, Leedal and Smith (2005) identified the workload construct as being applicable to the role of the anesthetist. In particular, they invoked a relationship between "spare capacity" as determined by workload metrics and a margin of safety in controlling the patient's physical state. They identified the potential use of workload metrics as a means for evaluating procedures and tools to ensure that there is adequate supply of spare capacity.

A number of recent studies have used various workload assessment techniques to evaluate the efficacy of medical instruments. For example, Spain and Bliss (2008) used the NASA-TLX subjective workload ratings for evaluating a novel sonification display for its potential utility in patient state monitoring during surgery. Sound parameters were examined for their ability to be informative without being distracting, or in other words to provide the needed information without consuming excess reserve capacity. Charabati et al. (2009) also used NASA-TLX ratings as a primary evaluation tool of a proposed novel integrated monitor of anesthesia (IMA) designed to integrate three essential data components for use by an anesthetist. Davis et al. (2009) used a secondary task for workload assessment of trainees in anesthesiology simulations. Their results clearly indicated an increased cognitive load as the trainees worked through the simulated emergencies. Davis et al. suggested that mental workload assessment should become a routine part of assessing training and equipment design to reduce the potential for future errors.

Perhaps due to the early recognition of the relevance of the SA concept to anesthesiology by Gaba and Lee (1990) and Gaba et al. (1995), workload and SA applications appear to be most prominent in this branch of medicine. Still, a recent review by Fioratou et al. (2010) on the application of the SA concept to anesthesiology found empirical research examining SA in anesthesia to be inadequate. In recognizing the utility of incorporating the consideration of SA in minimizing errors, Fioratou et al. (2010) called for more research and training on not only individual SA but also distributed SA among members of the surgical team. In other areas, Hazlehurst et al. (2007) found the SA construct to be very helpful in categorizing the communications among surgical teams during cardiac surgery. They argued that such work is important for understanding current practices and using them as a baseline for analyzing proposed procedural or equipment changes. In the emergency medical dispatch domain, Blandford

and Wong (2004) discussed the importance of the dispatchers having good SA for optimal ambulance control and the importance of understanding the task demands and the information needed for developing SA such that effective support could be provided.

Given all of these promising initial applications of mental workload and SA to understanding and improving medical human factors, it seems very likely that the use of these concepts in medicine will continue to increase for the foreseeable future.

## 4.4 Display Design

To the extent that excessive workload could reduce SA, any display that supports performance without incurring excessive workload would at least indirectly support SA as well (see, e.g., Previc, 2000). Wickens (1995) proposed that displays that do not overtax working memory and selective attention are particularly attractive because SA depends heavily on these processes. Wickens (1995, 2002, 2003) further discussed various display principles (e.g., proximity compatibility principle, visual momentum) that have been shown to support various types of performance (e.g., flight control as opposed to navigation) and display features (e.g., frame of reference) that would lend support to SA.

While display formats that facilitate information-processing support performance and thereby free up resources for SA maintenance, Wickens (2002, 2008b) shows that display formats could also affect the product (type) of SA. For example, a display with an egocentric frame of reference (an inside-out view with a fixed aircraft and a moving environment) provides better support for flight control, whereas an egocentric frame of reference (an outside-in view with a moving aircraft and fixed environment) provides better support for noticing hazards and general awareness of one's location. Wickens points out further that there are often trade-offs between alternative display formats. For example, whereas an integrated, ecological display generally provides better information about three-dimensional motion flow, a three-dimensional representation on a two-dimensional viewing surface tends to create ambiguity in locating objects in the environment. Such ambiguity is less of a problem in a two-dimensional display format. But it would take more than one two-dimensional display to present the same information in a three-dimensional display. It has been shown that it can be more cognitively demanding in trying to integrate information from two separate two-dimensional displays. The trade-off between promoting SA for objects in the environment and accomplishing other tasks at a lower workload level could only be resolved with regard to the specific goals or the priorities of competing goals of the system.

## 4.5 Training

Given the role that expertise plays in one's workload and SA, there is great potential in training to support SA and to permit tasks to be accomplished with less resources at a lower level of workload. The issue is: What does one train for? That expertise is based largely on a large body of domain-specific knowledge suggests that a thorough understanding of the workings of the

system would be helpful, particularly in nonroutine situations. Although expertise speeds up performance and experts generally perform at a high level under normal situations, their expertise is particularly useful in unexpected circumstances because of their ability to use their acquired knowledge to recognize and solve problems. Until there exist automated systems with total reliability that would never operate outside a perfectly orchestrated environment, the concern that operators trained on automated systems (which would be especially helpful for novices because of the presumed lower level of workload involved) might never acquire the needed knowledge and experience to build up their expertise is certainly a valid one. One possibility might be to provide some initial and refresher training in a nonautomated or less automated simulated system.

Although there exists in the literature a large body of training research that aims at accelerating the learning process and there is much evidence to support the advantages of not subjecting a trainee to an excessive level of workload, there are additional considerations when the goal is to build SA as well. First, it would be most useful to have some ideas about the knowledge structure that experts have so that the training program can build upon reinforcing this structure. After all, it is the structure and organization of information that support fast and accurate pattern recognition and information retrieval. Second, experts do not merely possess more knowledge, they are better at using it. This would suggest that training should extend to strategic training. Given the growing body of evidence to support that strategic task management (or executive control) is a higher level generalizable skill, much of the strategic training could be accomplished with low-cost low-physical-fidelity simulated systems such as a complex computer game (see Haier et al., 1992; Gopher, 1993). The strategic training can be at odds with the goal of keeping the level of workload down while the operators are in training. However, research has shown that the eventual benefits outweigh the initial cost in mental workload. As desirable as it is to train to develop automatic processing that is characterized as fast, accurate, and attention free, this training strategy may have only limited utility in training operators who have to function within a dynamic complex system. This is because there would be relatively few task components in these systems that would have an invariant stimulus–response mapping (a requirement for automatic processing to be developed and applied).

All five research areas underscore the interdependence of the concepts of workload and SA. The design of any efficacious technical support or training program would need to take into account the interplay of the two. Any evaluation of the effectiveness of these supports would need to assess both the operator's workload and SA in order to have a clear picture of their impact on system performance.

## 5   CONCLUSIONS

The years of research into mental workload and SA have been profitable. The research has developed a multitude of metric techniques, and although the results of different mental workload or SA assessment techniques sometimes show dissociations, they seem to fit within the theoretical constructs behind the measures. Workload is primarily a result of the limited attentional resources of humans, whereas SA is a cognitive phenomenon emerging from perception, memory, and expertise. The concepts of workload and SA have been studied extensively in the laboratory and have been transitioned successfully to real-world system evaluation. Indeed, workload and SA have been useful tools of system evaluators for years, and now they are providing vital guidance for shaping future automation, display, and training programs. In short, these concepts have been, and should continue to be, essential tools for human factors researchers and practitioners.

## REFERENCES

Adams, M. J., Tenny, Y. J., and Pew, R. W. (1991), *Stategic Workload and the Cognitive Management of Advanced Multi-Task System*, State of the Art Report No. CSERIAC 91-6, Crew System Ergonomics Information Evaluation Center, Wright-Patterson Air Force Base, OH.

Adams, M. J., Tenney, Y. J., and Pew, R. W. (1995), "Situation Awareness and the Cognitive Management of Complex Systems," *Human Factors*, Vol. 37, pp. 85–104.

Adams, R. J., and Ericsson, K. A. (1992), *Introduction to Cognitive Processes of Expert Pilots*, DOT/FAA/RD-92/12, U.S. Department of Transportation, Federal Aviation Administration, Washington, DC.

Alexander, A. L., Nygren, T. E., and Vidulich, M. A. (2000), *Examining the Relationship between Mental Workload and Situation Awareness in a Simulated Air Combat Task*, Technical Report No. AFRL-HE-WP-TR-2000-0094, Air Force Research Laboratory, Wright-Patterson Air Force Base, OH.

Allport, D. A., Antonis, B., and Reynolds, P. (1972), "On the Division of Attention: A Disproof of the Single Channel Hypothesis," *Quarterly Journal of Experimental Psychology*, Vol. 24, pp. 255–265.

Andre, A. D., Wickens, C. D., Moorman, L., and Boschelli, M. M. (1991), "Display Formatting Techniques for Improving Situation Awareness in the Aircraft Cockpit," *International Journal of Aviation Psychology*, Vol. 1, pp. 205–218.

Annett, J. (2002a), "Subjective Rating Scales: Science or Art?" *Ergonomics*, Vol. 45, pp. 966–987.

Annett, J. (2002b). "Subjective Rating Scales in Ergonomics: A Reply," *Ergonomics*, Vol. 45, pp. 1042–1046.

Backs, R. W. (1997), "Psychophysiological Aspects of Selective and Divided Attention during Continuous Manual Tracking," *Acta Psychologica*, Vol. 96, pp. 167–191.

Backs, R. W., and Boucsein, W. (2000), *Engineering Psychophysiology: Issues and Applications*, Lawrence Erlbaum Associates, Mahwah, NJ.

Baddeley, A. (1990), *Human Memory: Theory and Practice*, Allyn and Bacon, Boston, MA.

Bainbridge, L. (1987), "Ironies of Automation," in *New Technology and Human Error*, J. Rassmusen, K. Duncan, and J. Leplat, Eds., Wiley, New York, pp. 271–283.

Baker, S. C., Rogers, R. D., Owen, A. M., Frith, C. D., Dolan, R. J., Frackowiak, R. S., and Robbins, T. W. (1996), "Neural Systems Engaged by Planning: A PET Study of

the Tower of London Task," *Neuropsychologia*, Vol. 34, pp. 515–526.

Baldauf, D., Burgard, E., and Wittmann, M. (2009), "Time Perception as a Workload Measure in Simulated Car Driving," *Applied Ergonomics*, Vol. 40, pp. 929–935.

Battiste, V., and Bortolussi, M. (1988), "Transport Pilot Workload: A Comparison of Two Subjective Techniques," in *Proceedings of the Human Factors Society 32nd Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 150–154.

Beatty, J. (1982), "Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources," *Psychological Bulletin*, Vol. 91, pp. 276–292.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R. E., Tremonlet, P. D., and Craven, P. L. (2007), "EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks," *Aviation, Space, and Environmental Medicine*, Vol. 78, pp. B231–B244.

Bi, S., and Salvendy, G. (1994), "Analytical Modeling and Experiment Study of Human Workload in Scheduling of Advanced Manufacturing Systems," *International Journal of Human Factors in Manufacturing*, Vol. 4, pp. 205–235.

Biers, D. W., and Maseline, P. J. (1987), "Alternative Approaches to Analyzing SWAT Data," in *Proceedings of the Human Factors Society 31st Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 63–66.

Biers, D. W., and McInerney, P. (1988), "An Alternative to Measuring Subjective Workload: Use of SWAT without the Card Sort," in *Proceedings of the Human Factors Society 32nd Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 1136–1139.

Billings, C. E. (1997), *Aviation Automation: The Search for a Human-Centered Approach*, Lawrence Erlbaum Associates, Mahwah, NJ.

Blandford, A., and Wong, B. L. W. (2004), "Situation Awareness in Emergency Medical Dispatch," *International Journal of Human-Computer Studies*, Vol. 61, pp. 421–452.

Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., and Noll, D. C. (1997), "A Parametric Study of Prefrontal Cortex Involvement in Human Working Memory," *NeuroImage*, Vol. 5, pp. 49–62.

Brickman, B. J., Hettinger, L. J., Roe, M. M., Stautberg, D., Vidulich, M. A., Haas, M. W., and Shaw, R. L. (1995), "An Assessment of Situation Awareness in an Air Combat Task: The Global Implicit Measure Approach," in *Experimental Analysis and Measurement of Situation Awareness*, D. J. Garland and M. R. Endsley, Eds., Embry-Riddle Aeronautical University Press, Daytona Beach, FL, pp. 339–344.

Brickman, B. J., Hettinger, L. J., Stautberg, D., Haas, M.W., Vidulich, M. A., and Shaw, R. L. (1999), "The Global Implicit Measurement of Situation Awareness: Implications for Design and Adaptive Interface Technologies," in *Automation Technology and Human Performance: Current Research and Trends*, M.W. Scerbo and M. Mouloua, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 160–164.

Brookhuis, K. A., and de Waard, D. (2010), "Monitoring Drivers' Mental Workload in Driving Simulators Using Physiological Measures," *Accident Analysis and Prevention*, Vol. 42, pp. 898–903.

Brookings, J., Wilson, G.F., and Swain, C. (1996), "Psychophysiological Responses to Changes in Workload during Simulated Air Traffic Control," *Biological Psychology*, Vol. 42, pp. 361–378,

Bryne, E. A. (1995), "Role of Volitional Effort in the Application of Psychophysiological Measures to Situation Awareness," in *Experimental Analysis and Measurement of Situation Awareness*, D. J. Garland and M. R. Endsley, Eds., Embry-Riddle Aeronautical University Press, Daytona Beach, FL, pp. 147–153.

Carpenter, P. A., Just, M. A., Keller, T. A., Eddy, W. F., and Thulborn, K. R. (1999), "Graded Functional Activation in the Visuospatial System with the Amount of Task Demand," *Journal of Cognitive Neuroscience*, Vol. 11, pp. 9–24.

Casner, S. M. (2009), "Perceived vs. Measured Effects of Advanced Cockpit Systems on Pilot Workload and Error: Are Pilots' Beliefs Misaligned with Reality?" *Applied Ergonomics*, Vol. 40, pp. 448–456.

Charabati, S., Bracco, D., Mathieu, P. A., and Hemmerling, T. M. (2009), "Comparison of Four Different Display Designs of a Novel Anaesthetic Monitoring System, the Integrated Monitor of Anaesthesia (IMA^tm)," *British Journal of Anaesthesia*, Vol. 103, pp. 670–677.

Charness, N. (1989), "Age and Expertise: Responding to Talland's Challenge," in *Everyday Cognition in Adulthood and Late Life*, L. W. Poon, D. C. Rubin, and B. A. Wilson, Eds., Cambridge University Cambridge, England, pp. 437–456.

Charness, N. (1995), "Expert Performance and Situation Awareness," in *Experimental Analysis and Measurement of Situation Awareness* D. J. Garland and M. R. Endsley, Eds., Embry-Riddle Aeronautical University Press, Daytona Beach, FL, pp. 35–42.

Charness, N., and Tuffiash, M. (2008), "The Role of Expertise Research and Human Factors in Capturing, Explaining, and Producing Superior Performance," *Human Factors*, Vol. 50, pp. 427–432.

Chase, W. G., and Ericsson, K. A. (1982), "Skill and Working Memory," in *The Psychology of Learning and Motivation*, Vol. 16, G. H. Bower, Ed., Academic, New York.

Chi, M., Glaser, R., and Farr, M., Eds. (1988), *On the Nature of Expertise*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Coffey, E. B. J., Brouwer, A.-M., Wilschut, E. S., and van Erp, J. B. F. (2010), "Brain-machine Interfaces in Space: Using Spontaneous Rather Than Intentionally Generated Brain Signals," *Acta Astronautica*, Vol. 67, pp. 1–11.

Collet, C., Guillot, A., and Petit, C. (2010a), "Phoning While Driving I: A Review of Epidemiological, Psychological, Behavioural, and Physiological Studies," *Ergonomics*, Vol. 53, pp. 589–601.

Collet, C., Guillot, A., and Petit, C. (2010b), "Phoning While Driving II: A Review of Driving Conditions Influence," *Ergonomics*, Vol. 53, pp. 602–616.

Corbetta, M., Miezin, F. M., Dobmeyer, S., Shulman, G. L., and Petersen, S. E. (1990), "Attentional Modulation of Neural Processing of Shape, Color, and Velocity in Humans," *Science*, Vol. 248, pp. 1556–1559.

Damos, D. L., Ed. (1991), *Multiple-Task Performance*, Taylor & Francis, Washington, DC.

Davis, D. H. J., Oliver, M., and Byrne, A. J. (2009), "A Novel Method of Measuring the Mental Workload of Anesthetists during Simulated Practice," *British Journal of Anaesthesia*, Vol. 103, pp. 665–669.

Dekker, S., and Hollnagel, E. (2004), "Human Factors and Folk Models," *Cognition, Technology, and Work*, Vol. 6, pp. 79–86.

Dekker, S. W. A., and Woods, D. D. (2002), "MABA-MABA or Abracadabra? Progress on Human-Automation Coordination," *Cognition, Technology & Work*, Vol. 4, pp. 240–244.

D'Exposito, M. Zarahn, E., and Aguire, G. K. (1999), "Event-Related Functional MRI: Implications for Cognitive Psychology," *Psychological Bulletin*, Vol. 125, pp. 155–164.

Dey, A., and Mann, D. D. (2010), "Sensitivity and Diagnosticity of NASA-TLX and Simplified SWAT to Assess the Mental Workload Associated with Operating an Agricultural Sprayer," *Ergonomics*, Vol. 53, pp. 848–857.

Druckman, D., and Bjork, R. A., Eds. (1991), *In the Mind's Eye: Enhancing Human Performance* (Chap. 4), National Academy Press, Washington, DC.

Durso, F. T., and Alexander, A. L. (2010), "Managing Workload, Performance, and Situation Awareness in Aviation Systems," in *Human Factors in Aviation*, 2nd ed., E. Salas and D. Maurino, Eds., Elsevier, Burlington, MA, pp. 217–247.

Durso, F. T., and Dattel A. R. (2004), "SPAM: The Real-Time Assessment of SA," in *A Cognitive Approach to Situation Awareness: Theory and Application*, S. Banbury and S. Tremblay, Eds., Ashgate, Aldershot, United Kingdom, pp. 137–154.

Durso, F. T., and Gronlund, S. D. (1999), "Situation Awareness," in F. T. Durso, Ed., *Handbook of Applied Cognition*, Wiley, New York, pp. 283–314.

Durso, F. T., and Manning, C. A. (2008), "Air Traffic Control," in *Reviews of Human Factors and Ergonomics*, Vol. 4, M. Carswell, Ed., Human Factors and Ergonomics Society, Santa Monica, CA, pp. 199–204.

Durso, F. T., and Sethumadhavan, A. (2008), "Situation Awareness: Understanding Dynamic Environments," *Human Factors*, Vol. 50, pp. 442–448.

Durso, F. T., Truitt, T. R., Hackworth, C. A., Crutchfield, J. M., Nikolic, D., Moertl, P. M., Ohrt, D., and Manning, C. A. (1995a), "Expertise and Chess: A Pilot Study Comparing Situation Awareness Methodologies," in *Experimental Analysis and Measurement of Situation Awareness*, D. J. Garland and M. R. Endsley, Eds., Embry-Riddle Aeronautical University Press, Daytona Beach, FL, pp. 295–304.

Durso, F. T., Truitt, T. R., Hackworth, C. A., Ohrt, D., Hamic, J. M., Crutchfield, J. M., and Manning, C. A. (1995b), "Expertise and Chess: A Pilot Study Comparing Situation Awareness Methodologies," in *Experimental Analysis and Measurement of Situation Awareness*, D. J. Garland and M. R. Endsley, Eds., Embry-Riddle Aeronautical University Press, Daytona Beach, FL, pp. 189–195.

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., and Manning, C. A. (1998), "Situation Awareness as a Predictor of Performance for en Route Air Traffic Controllers," *Air Traffic Control Quarterly*, Vol. 6, pp. 1–20.

Durso, F. T., Crutchfield, J. M., and Batsakes, P. J. (2001), "Cognition in a Dynamic Environment," in *Usability Evaluation and Interface Design*, M. J. Smith, G. Salvendy, D. Harris, and R. J. Koubek, Eds., Erlbaum, Hillsdale, NJ, pp. 1373–1377.

Durso, F. T., Rawson, K. A., and Girotto, S. (2007). "Comprehension and Situation awareness," in *Handbook of Applied Cognition*, 2nd ed., F. T. Durso, R. S. Nickerson, S. T. Dumais, S. Lewandowsky, and T. Perfect, Eds., Erlbaum, Hoboken, NJ, pp. 164–193.

Dussault, C., Jouanin, J. C., Philippe, M., and Guezennec, C.-Y. (2005), "EEG and ECG Changes during Simulator Operation Reflect Mental Workload and Vigilance," *Aviation, Space, and Environmental Medicine*, Vol. 76, pp. 344–351.

Eggemeier, F. T., Crabtree, M. S., Zingg, J. J., Reid, G. B., and Shingledecker, C. A. (1982), "Subjective Workload Assessment in a Memory Update Task," in *Proceedings of the 26th Human Factors Society Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 643–647.

Endsley, M. R. (1988), "Situation Awareness Global Assessment Technique (SAGAT)," in *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference—NAECON 1988*, Vol. 3, Institute of Electrical and Electronics Engineers, New York, pp. 789–795.

Endsley, M. (1990), "A Methodology for the Objective Measurement of Pilot Situation Awareness," in *Situation Awareness in Aerospace Operations* (AGARD-CP-478), AGARD, Neuilly Sur Seine, France, pp. 1-1–1-9

Endsley, M. (1993), "Situation Awareness and Workload: Flip Sides of the Same Coin," in *Proceedings of the Seventh International Symposium on Aviation Psychology*, The Ohio State University, Columbus, OH, pp. 906–911.

Endsley, M. R. (1995), "Toward a Theory of Situation Awareness in Dynamic Systems," *Human Factors*, Vol. 37, pp. 32–64.

Endsley, M. R. (2006), "Situation Awareness," in *Handbook of Human Factors and Ergonomics*, 3rd. ed., G. Salvendy, Ed., Wiley, Hoboken, NJ, pp. 528–542.

Ericsson, K. A. (1996), "The Acquisition of Expert Performance: An Introduction to Some of the Issues," in *The Road to Excellence*, K. A. Ericsson, Ed., Erlbaum, Mahwah, NJ, pp. 1–50.

Ericsson, K. A., and Delaney, P. F. (1998), "Working Memory and Expert Performance," in *Working Memory and Thinking*, R. H. Logie and K. J. Gilhooly, Eds., Erlbaum, Hillsdale, NJ, pp. 93–114.

Ericsson, K. A., and Kintsch, W. (1995), "Long-Term Working Memory," *Psychological Review*, Vol. 105, pp. 211–245.

Ericsson, K. A., and Simon, H. A. (1993), *Protocol Analysis: Verbal Reports as Data*, rev. ed., MIT Press, Cambridge, MA.

Fabiani, M., Gratton, G., and Federmeier, K. D. (2007), "Event-Related Brain Potentials: Methods, Theory, and Applications," in *Handbook of Psychophysiology*, 3rd ed., J. T. Cacioppo, L. G., Tassinary, and G. G. Berntson, Eds., Cambridge University Press, New York, pp. 85–119.

Fallows, J. (2001a), *Free Flight: From Airline Hell to a New Age of Travel*, PublicAffairs, New York.

Fallows, J. (2001b, June), "Freedom of the Skies," *The Atlantic*, Vol. 287, No. 6, pp. 37–49.

Fallows, J. (2008a, May), "Taxis in the Sky," *The Atlantic*, available: http://www.theatlantic.com/magazine/archive/2008/05/taxis-in-the-sky/6775/.

Fallows, J. (2008b, May 5), "Air-Taxi Update, Propeller Plane Division," *The Atlantic*, available: http://www.theatlantic.com/science/archive/2008/05/air-taxi-update-propeller-plane-division/8018/.

Fallows, J. (2008c, September 20), "Air Taxi Chronicles: Bad News," *The Atlantic*, available: http://www.theatlantic.com/science/archive/2008/09/air-taxi-chronicles-bad-news/8840/.

Fioratou, E., Flin, R., Glavin, R., and Patey, R. (2010), "Beyond Monitoring: Distributed Situation Awareness in Anaesthesia," *British Journal of Anaesthesia, 105,* 83–90.

Fitts, P. M., and Jones, R. E. (1961a), "Analysis of Factors Contributing to 460 'Pilot-Error' Experiences in Operating Aircraft Controls," in *Selected Papers on Human Factors in the Design and Use of Control Systems*, H. W. Sinaiko, Ed., Dover, New York, pp. 332–358. Reprinted from *Analysis of Factors Contributing to 460 "Pilot-Error" Experiences in Operating Aircraft Controls*, Memorandum Report No. TSEAA-694-12, Aero Medical Laboratory, Air Materiel Command, Wright-Patterson Air Force Base, OH.

Fitts, P. M., and Jones, R. E. (1961b), "Psychological Aspects of Instrument Display," in *Selected Papers on Human Factors in the Design and Use of Control Systems*, H. W. Sinaiko, Ed., Dover, New York, pp. 359–396. Reprinted from *Psychological Aspects of Instrument Display*, Memorandum Report No. TSEAA-694-12A, Aero Medical Laboratory, Air Materiel Command, Wright-Patterson Air Force Base, OH.

Fournier, L. R., Wilson, G. F., and Swain, C. R. (1999), "Electrophysiological, Behavioral, and Subjective Indexes of Workload When Performing Multiple Tasks: Manipulations of Task Difficulty and Training," *International Journal of Psychophysiology*, Vol. 31, pp. 129–145.

Fowler, B. (1994), "P300 as a Measure of Workload during a Simulated Aircraft Landing Task," *Human Factors*, Vol. 36, pp. 670–683.

Frauenfelder, B. A., Schuepback, D., Baumgartner, R. W., and Hell, D. (2004), "Specific Alterations of Cerebral Hemohynamics during a Planning Task: A Transcranial Doppler Sonography Study," *Neuroimage*, Vol. 22, pp. 1223–1230.

Friedman, A., and Polson, M. C. (1981), "Hemispheres as Independent Resources Systems: Limited-Capacity Processing and Cerebral Specialization," *Journal of Experimental Psychology*, Vol. 7, pp. 1031–1058.

Fruhstorfer, H., Langanke, P., Meinzer, K., Peter, J., and Plaff, U. (1977), "Neurophysiological Vigilance Indicators and Operational Analysis of a Train Vigilance Monitoring Device: A Laboratoy and Field Study," in *Viligance: Theory, Physiological Correlates, and Operational Performance*, R. R. Mackie, Ed., Plenum, New York, pp. 147–162.

Gaba, D. M., and Lee, T., (1990), "Measuring the Workload of the Anesthesiologist," *Anasthesia and Analgesia*, Vol. 71, pp. 354–361.

Gaba, D. M., Howard, S. K., and Small, S. D. (1995), "Situation Awareness in Anesthesiology," *Human Factors*, Vol. 37, pp. 20–31.

Gawron, V. J. (2008), *Human Performance, Workload, and Situational Awareness Measures Handbook*, 2nd ed., CRC Press, Boca Raton, FL.

Gevins, A., and Smith, M. E. (2003), "Neurophysiological Measures of Cognitive Workload during Human-Computer Interaction," *Theoretical Issues in Ergonomics Science*, Vol. 4, pp. 113–131.

Gevins, A., Leong, H., Du, R., Smith, M. E., Le, J., DuRousseau, D., Zhang, J., and Libove, J. (1995), "Towards Measurement of Brain Function in Operational Environments," *Biological Psychology*, Vol. 40, pp. 1689–186.

Gevins, A., Smith, M. E., Leong, H., McEvoy, L., Whitfield, S., Du, R., and Rush, G. (1998), "Monitoring Working Memory during Computer-Based Tasks with EEG Pattern Recognition," *Human Factors*, Vol. 40, pp. 79–91.

Glaser, R. (1987), "Thoughts on Expertise," in *Cognitive Functioning and Social Structure over the Life Course*, C. Schooler and K. W. Schaie, Eds, Ablex, Norwood, NJ, pp. 81–94.

Glass, A. (1966), "Comparison of the Effect of Hard and Easy Mental Arithmetic upon Blocking of the Occipital Alpha Rhythm," *Quarterly Journal of Experimental Psychology*, Vol. 8, pp. 142–152.

Gopher, D. (1993), "The Skill of Attention Control: Acquisition and Execution of Attention Strategies," in *Attention and Performance*, Vol. 14, D. Meyer and S. Kornblum, Eds., Erlbaum, Hillsdale, NJ, pp. 299–322.

Gopher, D. (1994), "Analysis and Measurement of Mental Load," in *International Perspectives on Psychological Science*, Vol. 2, *The State of the Art*, G. d'Ydewalle, P. Eelen, and P. Bertelson, Eds., Erlbaum, Hove, United Kingdom, pp. 265–291.

Gopher, D., and Braune, R. (1984), "On the Psychophysics of Workload: Why Bother with Subjective Ratings?" *Human Factors*, Vol. 26, pp. 519–532.

Gopher, D., and Donchin, E. (1986), "Workload—An Examination of the Concept," in *Handbook of Perception and Human Performance*, Vol. II, *Cognitive Processes and Performance*, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., Wiley, New York, Chapter 41, pp. 41-1–41-49.

Gopher, D., Brickner, M., and Navon, D. (1982), "Different Difficulty Manipulation Interact Differently with Task Emphasis: Evidence for Multiple Resources," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 8, pp. 146–157.

Gould, K. S., Hirvonen, K., Koefoed, V. F., Røed, B. K., Sallinen, M. Holm, A., Bridger, R. S., and Moen, B. E. (2009), "Effects of 60 Hours of Total Sleep Deprivation on Two Methods of High-Speed Ship Navigation," *Ergonomics*, Vol. 52, pp. 1469–1486.

Haarmann, A., Boucsein, W., and Schaefer, F. (2009), "Combining Electrodermal Responses and Cardiovascular Measures for Probing Adaptive Automation during Simulated Flight," *Applied Ergonomics*, Vol. 40, pp. 1026–1040.

Haier, R. J., Siegel, B., Tang, C., Abel, L., and Buchsbaum, M. (1992), "Intelligence and Changes in Regional Cerebral Glucose Metabolic Rate Following Learning," *Intelligence*, Vol. 16, pp. 415–426.

Hardiman, T. D., Dudfield, H. J., Selcon, S. J., and Smith, F. J. (1996), "Designing Novel Head-Up Displays to Promote Situational Awareness," in *Situation Awareness: Limitations and Enhancement in the Aviation Environment*, AGARD-CP-575, Advisory Group for Aerospace Research and Development, Neuilly Sur Seine, France, pp. 15-1–15-7.

Hart, S. G., and Staveland, L. E. (1988), "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Human Mental Workload*, P. A. Hancock and N. Meshkati, Eds., Elsevier, Amsterdam, The Netherlands, pp. 139–183.

Hazlehurst, B., McMullen, C. K., and Gorman, P. N. (2007), "Distributed Cognition in the Heart Room: How Situation Awareness Arises from Coordinated Communications during Cardiac Surgery," *Journal of Biomedical Informatics*, Vol. 40, pp. 539–551.

Hendy, K. C. (1995), "Situation Awareness and Workload: Birds of a Feather?" in *AGARD Conference Proceedings 575: Situation Awareness: Limitations and Enhancements in the Aviation Environment*, Advisory Group for Aerospace Research & Development, Neuilly Sur Seine, France, pp. 21-1–21-7.

Hendy, K. C., Hamilton, K. M., and Landry, L. N. (1993), "Measuring Subjective Workload: When Is One Scale Better Than Many?" *Human Factors*, Vol. 35, pp. 579–601.

Hilburn, B., Joma, P. G., Byrne, E. A., and Parasuraman, R. (1997), "The Effect of Adaptive Air Traffic Control (ATC) Decision Aiding on Controller Mental Workload," in *Human–Automation Interaction*, M. Mouloua and J. Koonce, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 84–91.

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zakland, A. L., and Christ, R. E. (1992), "Comparison of Four Subjective Workload Rating Scales," *Human Factors*, Vol. 34, pp. 429–439.

Hockey, G. R. J. (1997), "Compensatory Control in the Regulation of Human Performance under Stress and High Workload: A Cognitive-Energetic Framework," *Biological Psychology*, Vol. 45, pp. 73–93.

Hockey, G. R. J. (2008), "From Theory to Practice: Commentary on Bartlett (1962)," *Ergonomics*, Vol. 51, 21–29.

Hockey, G. R. J., Gillard, A. W. K., and Coles, M. G. H. (1986), *Energetics and Human Processing*, M. Nijhoff, Boston, MA.

Hockey, R. J., Healey, A., Crawshaw, M., Wastell, D. G., and Sauer, J. (2003), "Cognitive Demands of Collision Avoidance in Simulated Ship Control," *Human Factors*, Vol. 45, pp. 252–265.

Hockey, G. R. J., Nickel, P., Roberts, A. C., and Roberts, M. H. (2009), "Sensitivity of Candidate Markers of Psychophysiological Strain to Cyclical Changes in Manual Control Load during Simulated Process Control," *Applied Ergonomics*, Vol. 40, pp. 1011–1018.

Hoffman, J., Houck, M., MacMilliam, F., Simons, R., and Oatman, L. (1985), "Event Related Potentials Elicited by Automatic Targets: A Dual-Task Analysis," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 11, pp. 50–61.

Horn, J. L., and Masunaga, H. (2000), "New Directions for Research into Aging and Intelligence: The Development of Expertise," in *Models of Cognitive Aging*, T. J. Perfect and E. A. Maylor, Eds., Oxford University Press, New York, pp. 125–159.

Horrey, W. J., Lesch, M. F., and Garabat, A. (2009), "Dissociation between Driving Performance and Drivers' Subjective Estimates of Performance and Workload in Dual-Task Conditions," *Journal of Safety Research*, Vol. 40, pp. 7–12.

Huey, B. M., and Wickens, C. D., Eds. (1993), *Workload Transition*, National Academy Press, Washington, DC.

Isreal, J. B., Chesney, G. L., Wickens, C. D., and Donchin, E. (1980), "P300 and Tracking Difficulty: Evidence for Multiple Resources in Dual Task Performance," *Psychobiology*, Vol. 17, pp. 57–70.

Joint Planning and Development Office (2007), *Concept of Operations for the Next Generation Air Transportation System Version 2.0*, Author, Washington, DC.

Jorna, P. G. A. M. (1992), "Spectral Analysis of Heart Rate and Psychological State: A Review of Its Validity as a Workload Index," *Biological Psychology*, Vol. 34, pp. 237–258.

Jorna, P. G. A. M. (1999), "Automation and Free Flight: Exploring the Unexpected," in *Automation Technology and Human Performance*, M. W. Scerbo and Mouloua, Eds., Erlbaum, Mahwah, NJ, pp. 107–111.

Just, M. A., and Carpenter, P. A. (1992), "A Capacity Theory of Comprehension: Individual Differences in Working Memory," *Psychological Review*, Vol. 99, pp. 122–149.

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., and Thulborn, K. R. (1996), "Brain Activation Modulated by Sentence Comprehension," *Science*, Vol. 274, pp. 114–116.

Just, M. A., Carpenter, P. A., and Miyake, A. (2003), "Neuroindices of Cognitive Workload: Neuroimaging, Pupillometric and Event-Related Potential Studies of Brain Work," *Theoretical Issues in Ergonomics Science*, Vol. 4, pp. 56–88.

Just, M. A., Keller, T. A., and Cynkar, J. (2008), "A Decrease in Brain Activation Associated with Driving When Listening to Someone Speak," *Brain Research*, Vol. 1205, pp. 70–80.

Kaber, D. B., and Riley, J. M. (1999), "Adaptive Automation of a Dynamic Control Task Based on Secondary Task Workload Measurement," *International Journal of Cognitive Ergonomics*, Vol. 3, pp. 169–187.

Kaber, D. B., Perry, C., Segall, N., McClernon, C. K., and Prinzel, L. J., III (2006), "Situation Awareness Implications of Adaptive Automation for Information Processing in an Air Traffic Control-Related Task," *International Journal of Industrial Ergonomics*, Vol. 36, pp. 3447–3462.

Kahneman, D. (1973), *Attention and Effort*, Prentice-Hall, Englewood Cliffs, NJ.

Kantowitz, B. H. (1992), "Selecting Measures for Human Factors Research," *Human Factors*, Vol. 34, pp. 387–398.

Kessel, C. J. and Wickens, C. D. (1982), "The Transfer of Failure Detection Skills between Monitoring and Controlling Dynamic Systems," *Human Factors*, Vol. 24, pp. 49–60.

Kinsbourne, M., and Hicks, R. (1978), "Functional Cerebral Space," in *Attention and Performance*, Vol. VII, J. Requin, Ed., Erlbaum, Hillsdale, NJ, pp. 345–362.

Kohlmorgen, J., Dornhege, G., Braun, M. L., Blankertz, B., Müller, K, Curio, G., Hogemann, K., Bruns, A., Schrauf, M., and Kincses, W. E. (2007), "Improving Human Performance in a Real Operating Environment through Real-Time Mental Workload Detection," in *Toward Brain-Computer Interfacing*, G. Dornhege, J. del R. Millán, T. Hinterberger, D. J., McFarland, and K. Müller, Eds., MIT Press, Cambridge, MA, pp. 409–422.

Kramer, A. F., and Parasuraman, R. (2007), "Neuroergonomics: Applications of Neuroscience to Human Factors," in *Handbook of Psychophysiology*, 3rd ed., J. T. Cacioppo, L. G., Tassinary, and G. G. Berntson, Eds., Cambridge University Press. New York, pp. 704–722.

Kramer, A. F., and Weber, T. (2000), "Applications of Psychophysiology to Human Factors," in *Handbook of Psychophysiology*, J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, Eds., Cambridge University Press, New York, 2nd ed., pp. 794–814.

Kramer, A. F., Sirevaag, E., and Braune, R. (1987), "A Psychophysiological Assessment of Operator Workload during Simulated Flight Missions," *Human Factors*, Vol. 29, pp. 145–160.

Kramer, A. F., Trejo, L. J., and Humphrey, D. G. (1996), "Psychophysiological Measures of Workload: Potential Applications to Adaptively Automated Systems," in *Automation and Human Performance: Theory and Application*, R. Parasuraman and M. Mouloua, Eds., Erlbaum, Mahway, NJ, pp. 137–161.

Langan-Fox, J., Sankey, M. J., and Canty, J. M. (2009), "Human Factors Measurement for Future Air Traffic Control Systems," *Human Factors*, Vol. 51, pp. 595–637.

Lee, J. D., and Moray, N. (1994), "Trust, Self-Confidence, and Operators' Adaptation to Automation," *International Journal of Human-Computer Studies*, Vol. 40, pp. 153–184.

Lee, Y., and Liu, B. (2003), "Inflight Workload Assessment: Comparison of Subjective and Physiological Measurements," *Aviation, Space, and Environmental Medicine*, Vol. 74, pp. 1078–1084.

Leedal, J. M., and Smith, A. F. (2005), "Methodological Approaches to Anaesthetists' Workload in the Operating Theatre," *British Journal of Anaesthesia*, Vol. 94, pp. 702–709.

Lehrer, P., Karavidas, M., Lu, S.-E., Vaschillo, B., and Cheng, A., (2010), "Cardiac Data Increase Association between Self-Report and Both Expert Ratings of Task Load and Task Performance in Flight Simulator Tasks: An Exploratory Study," *International Journal of Psychophysiology*, Vol. 76, pp. 80–87.

Leplat, J., and Welford, A. T., Eds. (1978), *Ergonomics*, Vol. 21, No. 3. Issue on Workload.

Leyman, E. L. C., Mirka, G. A., Kaber, D. B., and Sommerich, C. M., (2004), "Cervicobrachial Muscle Response to Cognitive Load in a Dual-Task Scenario," *Ergonomics*, Vol. 47, pp. 625–645.

Lidderdale, I. G. (1987), "Measurement of Aircrew Workload in Low-Level Flight," in *AGARDograph No. 282—The Practical Assessment of Pilot Workload*, A. Roscoe, Ed., Advisory Group for Aerospace Research and Development, Neuilly Sur Seine, France, pp. 67–77.

Low, K. A., Leaver, E. E., Kramer, A. F., Fabiani, M., and Gratton, G. (2009), "Share or Compete? Load-Dependent Recruitment of Prefrontal Cortex during Dual-Task Performance," *Psychophysiology*, Vol. 46, pp. 1069–1079.

Luximon, A., and Goonetilleke, R. S. (2001), "Simplified Subjective Workload Assessment Technique," *Ergonomics*, Vol. 44, pp. 229–243.

Lysaght, R. J., Hill, S. G., Dick, A. O., Plamondon, B. D., Linton, P. M., Wierwille, W. W., Zaklad, A. L., Bittner, A. C. Jr., and Wherry, R. J. (1989), "Operator Workload: Comprehensive Review and Evaluation of Workload Methodologies," ARI Technical Report 851, U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Ma, R., and Kaber, D. B. (2005), "Situation Awareness and Workload in Driving While Using Adaptive Cruise Control and a Cell Phone," *International Journal of Industrial Ergonomics*, Vol. 35, pp. 939–953.

McCarley, J. S., and Kramer, A. F. (2007), "Eye Movements as a Window on Perception and Cognition," in *Neuroergonomics: The Brain at Work*, R. Parasuraman and M. Rizzo, Eds., Oxford University Press, New York, pp. 95–112.

McCoy, T. M., Derrick, W. L., and Wickens, C. D. (1983), "Workload Assessment Metrics—What Happens When They Dissociate?" in *Proceedings of the Second Aerospace Behavioral Engineering Technology Conference, P-132*, Society of Automotive Engineers, Warrendale, PA, pp. 37–42.

Meade, M. L., Nokes, T. J., and Morrow, D. G. (2009), "Expertise Promotes Facilitation on a Collaborative Memory Task," *Memory*, Vol. 17, pp. 39–48.

Moray, N., Ed. (1979), *Mental Workload, Theory and Measurement*, Plenum, New York.

Moray, N. (1986), "Monitoring Behavior and Supervisory Control," in *Handbook of Perception and Human Performance*, Vol. II, *Cognitive Processes and Performance*, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., Wiley, New York, Chapter 40, pp. 40-1–40-51.

Moray, N. (1997), "Human Factors in Process Control," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York, pp. 1944–1971.

Muckler, F.A., and Seven, S.A. (1992), "Selecting Performance Measures: 'Objective' versus 'Subjective' Measurement," *Human Factors*, Vol. 34, pp. 441–456.

Mulder, L. J. M. (1992), "Measurement and Analysis Methods of Heart Rate and Respiration for Use in Applied Environments," *Biological Psychology*, Vol. 34, pp. 205–236.

Navon, D., and Gopher, D. (1979), "On the Economy of the Human-Processing System," *Psychology Review*, Vol. 86, pp. 214–255.

Norman, D. A., and Bobrow, D. G. (1975), "On Data-Limited and Resource-Limited Processes," *Cognitive Psychology*, Vol. 7, pp. 44–64.

Norman, G. R., Brooks, L. R., and Allen, S. W. (1989), "Recall by Expert Medical Practitioners and Novices as a Record of Processing Attention," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 15, pp. 1166–1174.

Nygren, T. E. (1991), "Psychometric Properties of Subjective Workload Techniques: Implications for Their Use in the Assessment of Perceived Mental Workload," *Human Factors*, Vol. 33, pp. 17–33.

O'Donnell, R., and Eggemeier, F. T. (1986), "Workload Assessment Methodology," in *Handbook of Perception and Human Performance*, Vol. II, *Cognitive Processes and Performance*, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., Wiley, New York, Chapter 42, pp. 42-1–42-49.

Ogden, G. D., Levine, J. M., and Eisner, E. J. (1970), "Measurement of Workload by Secondary Tasks," *Human Factors*, Vol. 21, pp. 529–548.

Parasuraman, R. (1990), "Event-Related Brain Potentials and Human Factors Research," in *Event-Related Brain Potentials: Basic Issues and Applications*, J. W. Rohrbaugh, R. Parasuaman, and R. Johnson, Jr., Eds., Oxford University Press, New York, pp. 279–299.

Parasuraman, R. (2003), "Neuroergonomics: Research and Practice," *Theoretical Issues in Ergonomics Science*, Vol. 4, pp. 5–20.

Parasuraman, R., and Bryne, E. A. (2003), "Automation and Human Performance in Aviation," in *Principles and Practice of Aviation Psychology*, P. S. Tsang and M. A. Vidulich, Eds., Erlbaum, Mahwah, NJ, pp. 311–356.

Parasuraman, R., and Caggiano, D. (2002), "Mental Workload," in *Encyclopedia of the Human Vrain*, Vol. 3, V. S. Ramachandran, Ed., Academic, San Diego, pp. 17–27.

Parasuraman, R., and Rizzo, M., Eds. (2007), *Neuroergonomics: The Brain at Work*, Oxford University Press, New York.

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2008), "Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs," *Journal of Cognitive Engineering and Decision Making*, Vol. 2, pp. 140–160.

Parasuraman, R., Cosenzo, K. A., and De Visser, E. (2009), "Adaptive Automation for Human Supervision of Multiple Uninhabited Vehicles: Effects on Change Detection, Situation Awareness, and Mental Workload," *Military Psychology*, Vol. 21, pp. 270–297.

Parks, R. W., Lowenstein, D. A., Dodrill, K. L., Barker, W. W., Yoshi, F., Chang, J. Y., Emran, A., Apicella, A., Sheramata, W. A., and Duara, R. (1988), "Cerebral Metabolic Effects of a Verbal Fluency Test: A PET Scan Study," *Journal of Clinical and Experimental Neurophysiology*, Vol. 10, pp. 565–575.

Patel, V. L., and Groen, G. J. (1991), "The General and Specific Nature of Medical Expertise: A Critical Look," in *Toward a General Theory of Expertise*, K. A. Ericsson and J. Smith, Eds., Cambridge University Press, Cambridge, MA, pp. 93–125.

Pérès, M., Van De Moortele, P. F., Pierard, C., Lehericy, S., Satabin, P., Le Bihan, D., and Guezennec, C. Y. (2000), "Functional Magnetic Resonance Imaging of Mental Strategy in a Simulated Aviation Performance Task," *Aviation, Space, and Environmental Medicine*, Vol. 71, pp. 1218–1231.

Perry, C. M., Sheik-Nainar, M. A., Segall, N., Ma, R., and Kaber, D. B. (2008), "Effects of Physical Workload on Cognitive Task Performance and Situation Awareness," *Theoretical Issues in Ergonomics Science*, Vol. 9, pp. 95–113.

Pew, R. A. (1994), "An Introduction to the Concept of Situation Awareness," in *Situational Awareness in Complex Systems*, R. D. Gilson, D. J. Garland, and J. M. Koonce, Eds., Embry-Riddle Aeronautical University Press, Daytona Beach, FL, pp. 17–23.

Pizzagalli, D. A. (2007), "Electroencephalography and High-Density Electophysiological Source Localization," in *Handbook of Psychophysiology*, 3rd ed., J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, Eds., Cambridge University Press, New York, pp. 56–84.

Posner, M. I., and DiGirolamo, G. J. (2000), "Cognitive Neuroscience: Origin and Promise," *Psychological Bulletin*, Vol. 126, pp. 873–889.

Previc, F. H. (2000), "Neuropsychological Guidelines for Aircraft Control Stations," *IEEE Engineering in Medicine and Biology*, March/April, pp. 81–88.

Prewett, M. S., Johnson, R. C., Saboe, K. N., Elliot, L. R., and Coovert, M. D. (2010), "Managing Workload in Human-Robot Interaction: A Review of Empirical Studies," *Computers in Human Behavior*, Vol. 26, pp. 840–856.

Prinzel, L. J., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., and Pope, A. T. (2000), "A Closed-Loop System for Examining Psychophysiological Measures for Adaptive Automation," *International Journal of Aviation Psychology*, Vol. 10, pp. 393–410.

Pritchett, A. R., and Hansman, R. J. (2000), "Use of Testable Responses for Performance-Based Measurement of Situation Awareness," in *Situation Awareness Analysis and Measurement*, M. R. Endsley and D. J. Garland, Eds., Erlbaum, Mahwah, NJ, pp. 189–209.

Proctor, R. W., and Vu, K.-P. (2010), "Cumulative Knowledge and Progress in Human Factors," *Annual Review of Psychology*, Vol. 61, pp. 623–651.

Reid, G. B., and Nygren, T. E. (1988), "The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload," in *Human Mental Workload*, P. A. Hancock and N. Meshkati, Eds., Elsevier, Amsterdam, The Netherlands, pp. 185–218.

Reingold, E. M., Charness, N., Pomplun, M., and Stampe, D. M. (2001), "Visual Span in Expert Chess Players: Evidence from Eye Movements," *Psychological Science*, Vol. 12, pp. 48–55.

Rothrock, L., Koubek, R., Fuchs, F., Haas, M., and Salvendy, G. (2002), "Review and Reappraisal of Adaptive Interfaces: Toward Biologically Inspired Paradigms," *Theoretical Issues in Ergonomics Science*, Vol. 3, pp. 47–84.

Rouse, W. B. (1977), "Human-Computer Interaction in Multitask Situations," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 7, pp. 293–300.

Rouse, W. B. (1988), "Adaptive Aiding for Human/Computer Control," *Human Factors*, Vol. 30, pp. 431–443.

Rubio, S., Díaz, E., Martín, J., and Puente, J. M. (2004), "Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile methods," *Applied Psychology: An International Review*, Vol. 53, pp. 61–86.

Saaty, T. L. (1980), *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation,* McGraw-Hill, New York.

Salmon, P., Stanton, N., Walker, G., and Green, D. (2006), "Situation Awareness Measurement: A Review of Applicability for C4i Environments," *Applied Ergonomics*, Vol. 37, pp. 225–238.

Salvendy, G. (2002), "Use of Subjective Rating Scores in Ergonomics Research and Practice," *Ergonomics*, Vol. 45, pp. 1005–1007.

Scerbo, M. W. (2007), "Adaptive Automation," in *Neuroergonomics: The Brain at Work*, R. Parasuraman and M. Rizzo, Eds., Oxford University Press, New York, pp. 239–252.

Selcon, S. J., Taylor, R. M., and Koritsas, E. (1991), "Workload or Situational Awareness? TLX vs. SART for Aerospace System Design Evaluation," in *Proceedings of the Human Factors Society 35th Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 62–66.

Selcon, S. J., Taylor, R. M., and Shadrake, R. M. (1992), "Multi-Modal Cockpit Warnings: Pictures, Words, or Both?" in *Proceedings of the Human Factors Society 36th Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 57–61.

Selcon, S. J., Hardiman, T. D., Croft, D. G., and Endsley, M. R. (1996), "A Test-Battery Approach to Cognitive Engineering: To Meta-Measure or Not to Meta-Measure, That Is the Question!" in *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting*, Vol. 1, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 228–232.

Serrati, C., Finocci, C., Calautti, C., Bruzzone, G. L., Colucci, M. D S., Gandolfo, C., Lantieri, P. B., and Farale, E. (2000), "Absence of Hemispheric Dominance for Mental Rotation Ability: A Transcranial Doppler Study," *Cortex*, Vol. 36, pp. 415–425.

Shallice, T. (1988), *From Neuropsychology to Mental Structure*, Cambridge University Press, New York.

Shaw, R. L., Brickman, B. J., and Hettinger, L. J. (2004), "The Global Implicit Measure (GIM): Concept and Experience," in *Human Performance, Situation Awareness and Automation*, D. A. Vincenzi, M. Mouloua, and P. A. Hancock, Eds., Erlbaum, Mahwah, NJ, pp. 97–102.

Sheridan, T. B. (2009), "Human Factors Research Needs for NextGen-airportal Safety," Technical Report NASA/CR-2009-215372, National Aeronautics and Space Administration, Moffett Field, CA.

Shingledecker, C.A. (1984), "A Task Battery for Applied Human Performance Assessment Research", Technical Report AFAMRL-TR-84-071, Air Force Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, OH.

Sirevaag, E. J., Kramer, A. F., Coles, M. G. H., and Donchin, E. (1989), "Resource Reciprocity: An Event-Related Brain Potentials Analysis," *Acta Psychologica*, Vol. 70, pp. 77–97.

Sirevaag, E. J., Kramer, A. F., Wickens, C. D., Reisweber, M., Strayer, D. L., and Grenell, J. (1993), "Assessment of Pilot Performance and Mental Workload in Rotary Wing Aircraft," *Ergonomics*, Vol. 36, pp. 1121–1140.

Spain, R. D., and Bliss, J. P. (2008), "The Effect of Sonification Display Pulse Rate and Reliability on Operator Trust and Perceived Workload during a Simulated Patient Monitoring Task," *Ergonomics*, Vol. 51, pp. 1320–1337.

Sperandio, A. (1978), "The Regulation of Working Methods as a Function of Workload among Air Traffic Controllers," *Ergonomics*, Vol. 21, pp. 367–390.

Staszewski, J. J. (1988), "Skilled Memory and Expert Mental Calculation," in *The Nature of Expertise*, M. T. H. Chi and R. Glaser, Eds., Erlbaum, Hillsdale, NJ, pp. 71–128.

Sterman, B., and Mann, C. (1995), "Concepts and Applications of EEG Analysis in Aviation Performance Evaluation," *Biological Psychology*, Vol. 40, pp. 115–130.

Stern, J. A., Walrath, L. C., and Goldstein, R. (1984), "The Endogenous Eyeblink," *Psychophysiology*, Vol. 21, pp. 22–33.

Stevens, S. S. (1957), "On the Psychophysical Law," *Psychological Review*, Vol. 64, pp. 153–181.

Stevens, S. S. (1966), "On the Operation Known as Judgment," *American Scientist*, Vol. 54, pp. 385–401.

Strater, L. D., Endsley, M. R., Pleban, R. J., and Matthews, M. D., (2001), "Measures of Platoon Leader Situation Awareness in Virtual Decision-Making Exercises," Technical Report No. 1770, U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Strayer, D., and Kramer, A. F. (1990), "Attentional Requirements of Automatic and Controlled Processes," *Journal of Experimental Psychology: Learning, Memory and Cognition*, Vol. 16, pp. 67–82.

Taylor, R. M. (1990), "Situational Awareness Rating Technique (SART): The Development of a Tool for Aircrew Systems Design," in *AGARD-CP-478, Situational Awareness in Aerospace Operations*, Advisory Group for Aerospace Research and Development, Neuilly Sur Seine, France, pp. 3-1–3-17.

Tenny, Y. J., and Pew, R. W. (2006), "Situation Awareness Catches On. What? So What? Now What?" in *Reviews of Human Factors and Ergonomics*, Vol. 2, R. C. Williges, Ed., Human Factors and Ergonomics Society, Santa Monica, CA, pp. 89–129.

Tenny, Y. J., Adams, M. J., Pew, R. W. Huggins, A. W. F., and Rodgers, W. H. (1992), "A Principled Approach to the Measurement of Situation Awareness in Commercial Aviation," NASA Contractor Report 4451, Langley Research Center, National Aeronautics and Space Administration, Hampton, VA.

Ting, C-H., Mahfouf, M., Nassef, A., Linkens, D. A., Panoutsos, G., Nickel, P., Roberts, A. C., and Hockey, G. R. J. (2010), "Real-Time Adaptive Automation System Based on Identification of Operator Functional State in Simulated Process Control Operations," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, Vol. 40, pp. 251–262.

Toms, M. L., Cavallo, J. J., Cone, S. M., Moore, F. W., and Gonzalez-Garcia, A. (1997), "Pilot-in-the-Loop Evaluation of the Approach Procedures Expert System (APES)," Technical Report No. WL-TR-97-3102, Wright Laboratory, Wright-Patterson Air Force Base, OH.

Tripp, L. D., and Warm, J. S. (2007), "Transcranial Doppler Sonography," in *Neuroergonomics: The Brain at Work*, R. Parasuraman and M. Rizzo, Eds., Oxford University Press, New York, pp. 82–94.

Tsang, P. S. (2007), "The Dynamics of Attention and Aging in Aviation," in *Applied Attention: From Theory to Practice*, A. F. Kramer, A. Kirlik, and D. Weigmann, Eds., Oxford University Press, New York, pp. 170–184.

Tsang, P. S., and Shaner, T. L. (1998), "Age, Attention, Expertise, and Time-Sharing Performance," *Psychology and Aging*, Vol. 13, pp. 323–347.

Tsang, P. S., and Velazquez, V.L. (1996), "Diagnosticity and Multidimensional Workload Ratings," *Ergonomics*, Vol. 39, pp. 358–281.

Tsang, P. S., and Vidulich, M. A. (1989), "Cognitive Demands of Automation in Aviation," in *Aviation Psychology*, R. S. Jensen, Ed., Gower, Aldershot, United Kingdom, pp. 66–95.

Tsang, P. S., and Vidulich, M. A. (1994), "The Roles of Immediacy and Redundancy in Relative Subjective Workload Assessment," *Human Factors*, Vol. 36, pp. 503–513.

Tsang, P. S., and Vidulich, M. A. (2006), "Mental Workload and Situation Awareness," in *Handbook of Human Factors and Ergonomics*, 3rd ed., G. Salvendy, Ed., Wiley, Hoboken, NJ, pp. 243–268.

Tsang, P. S., and Wilson, G. (1997), "Mental Workload," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York, pp. 417–449.

Tsang, P. S., Velazquez, V. L., and Vidulich, M. A. (1996), "The Viability of Resource Theories in Explaining Time-Sharing Performance," *Acta Psychologica*, Vol. 91, pp. 175–206.

Ververs, P. M., and Wickens, C. D. (2000), "Designing Head-Up Displays (HUDs) to Support Flight Path Guidance While Minimizing Effects of Cognitive Tunneling," in *Proceedings of the IEA 2000/HFES 2000 Congress*, Vol. 3, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 3-45–3-48.

Vidulich, M. A. (1988), "The Cognitive Psychology of Subjective Mental Workload," in *Human Mental Workload*, P. Hancock and N. Meshkati, Eds., North Holland. Amsterdam, The Netherlands, pp. 219–229.

Vidulich, M. A. (1989), "The Use of Judgment Matrices in Subjective Workload Assessment: The Subjective WORkload Dominance (SWORD) Technique," in *Proceedings of the Human Factors Society 33rd Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1406–1410.

Vidulich, M. A. (1995), "The Role of Scope as a Feature of Situation Awareness Metrics," in *Experimental Analysis and Measurement of Situation Awareness*, D. J. Garland and M. R. Endsley, Eds., Embry-Riddle Aeronautical University Press, Daytona Beach, FL, pp. 69–74.

Vidulich, M. A. (2000), "Testing the Sensitivity of Situation Awareness Metrics in Interface Evaluations," in *Situation Awareness Analysis and Measurement*, M. R. Endsley and D. J. Garland, Eds., Erlbaum, Mahwah, NJ, pp. 227–246.

Vidulich, M. A. (2003), "Mental Workload and Situation Awareness: Essential Concepts for Aviation Psychology Practice," in *Principles and Practice of Aviation Psychology*, P. S. Tsang and M. A. Vidulich, Eds., Erlbaum, Mahwah, NJ, pp. 115–146.

Vidulich, M. A., and Bortolussi, M. R. (1988), "Speech Recognition in Advanced Rotorcraft: Using Speech Controls to Reduce Manual Control Overload," in *Proceedings of the American Helicopter Society National Specialist's Meeting: Automation Applications for Rotorcraft*, American Helicopter Society, Atlanta Southeast Region. Atlanta, GA, pp. 1–10.

Vidulich, M. A., and Hughes, E. R. (1991), "Testing a Subjective Metric of Situation Awareness," in *Proceedings of the Human Factors Society 35th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1307–1311.

Vidulich, M. A., and McMillan, G. (2000), "The Global Implicit Measure: Evaluation of Metrics for Cockpit Adaptation," in *Contemporary Ergonomics 2000*, P. T. McCabe, M. A. Hanson, and S. A. Robertson, Eds., Taylor and Francis, London, pp. 75–80.

Vidulich, M. A., and Tsang, P. S. (1987), "Absolute Magnitude Estimation and Relative Judgment Approaches to Subjective Workload Assessment," in *Proceedings of the Human Factors Society 31st Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1057–1061.

Vidulich, M. A., and Wickens, C. D. (1986), "Causes of Dissociation between Subjective Workload Measures and Performance: Caveats for the Use of Subjective Assessments," *Applied Ergonomics*, Vol. 17, pp. 291–296.

Vidulich, M. A., Dominguez, C., Vogel, E., and McMillan, G. (1994a), "Situation Awareness: Papers and Annotated Bibliography", Technical Report No. AL/CF-TR-1994-0085, Armstrong Laboratory, Wright-Patterson Air Force Base, OH.

Vidulich, M. A., Stratton, M., Crabtree, M., and Wilson, G. (1994b), "Performanced-Based and Physiological Measures of Situational Awareness," *Aviation, Space, and Environmental Medicine*, Vol. 65(s,), pp. A7–A12.

Vidulich, M. A., McCoy, A. L., and Crabtree, M. S. (1995), "The Effect of a Situation Display on Memory Probe and Subjective Situational Awareness Metrics," in *Proceedings of the Eighth International Symposium on Aviation Psychology*, The Ohio State University, Columbus, OH, pp. 765–768.

Vidulich, M. A., Wickens, C. D., Tsang, P. S., and Flach, J. M. (2010), "Information Processing in Aviation," in *Human Factors in Aviation*, 2nd ed., E. Salas and D. Maurino, Eds., Elsevier, Burlington, MA, pp. 175–215.

Wästlund, E., Norlander, T., and Archer, T. (2008), "The Effect of Page Layout on Mental Workload: A Dual-Task Experiment," *Computers in Human Behavior*, Vol. 24, pp. 1229–1245.

Watson, J. B. (1913), "Psychology as the Behaviorist Views It," *Psychological Review*, Vol. 20, pp. 158–177.

Welford, A. T. (1978), "Mental Workload as a Function of Demand, Capacity, Strategy and Skill," *Ergonomics*, Vol. 21, pp. 151–167.

Wickens, C. D. (1980), "The Structure of Attentional Resources," in *Attention and Performance*, Vol. VIII, R. S. Nickerson, Ed., Erlbaum, Hillsdale, NJ, pp. 239–257.

Wickens, C. D. (1984), "Processing Resources in Attention," in *Varieties of Attention*, R. Parasuraman and D. R. Davies, Eds., Academic, San Diego, CA, pp. 63–102.

Wickens, C. D. (1987), "Attention," in *Human Mental Workload*, P. A. Hancock and N. Meshkati, Eds., North-Holland, New York, pp. 29–80.

Wickens, C. D. (1990), "Applications of Event-Related Potential Research to Problems in Human Factors," in *Event-Related Brain Potentials: Basic Issues and Applications*, J. W. Rohrbaugh, R. Parasuaman, and R. Johnson, Jr., Eds., Oxford University Press, New York, pp. 301–309.

Wickens, C. D. (1995), "Situation Awareness: Impact of Automation and Display Technology," in *AGARD Conference Proceedings 575: Situation Awareness: Limitations and Enhancements in the Aviation Environment*, Advisory Group for Aerospace Research & Development, Neuilly, Sur Seine, France, pp. K2-1–K2-13.

Wickens, C. D. (2001), "Workload and Situation Awareness," in *Stress, Workload, and Fatigue*, P. A. Hancock and P. A. Desmond, Eds., Erlbaum, Mahwah, NJ, pp. 443–450.

Wickens, C. D. (2002), "Multiple Resources and Performance Prediction," *Theoretical Issues in Ergonomics Science*, Vol. 3, pp. 159–177.

Wickens, C. D. (2003), "Aviation Displays," in *Principles and Practice of Aviation Psychology*, P. S. Tsang and M. A. Vidulich, Eds., Erlbaum, Mahwah, NJ, pp. 147–200.

Wickens, C. D. (2008a), "Multiple Resources and Mental Workload," *Human Factors*, Vol. 50, pp. 449–455.

Wickens, C. D. (2008b), "Situation Awareness: Review of Mica Endsley's 1995 Articles on Situation Awareness Theory and Measurement," *Human Factors*, Vol. 50, pp. 397–403.

Wickens, C. D., and Hollands, J. G. (2000), *Engineering Psychology and Human Performance*, 3rd ed., Charles E. Merrill Publishing Company, Columbus, OH:

Wickens, C. D., Kramer, A. F., Vanasse, L., and Donchin, E. (1983), "The Performance of Concurrent Tasks: A Psychophysiological Analysis of the Reciprocity of Information Processing Resources," *Science*, Vol. 221, pp. 1080–1082.

Wickens, C. D., McCarley, J. S., Alexander, A. L., Thomas, L. C., Ambinder, M., and Zheng, S. (2008), "Attention-Situation Awareness Model of Pilot Error," in *Human Performance Modeling in Aviation*, D. Foyle and B. Hooey, Eds., Taylor and Francis, Boca Raton, FL, pp. 213–239.

Wiebe, E. N., Roberts, R., and Behrend, T. S. (2010), "An Examination of Two Mental Workload Measurement Approaches to Understanding Multimedia Learning," *Computers in Human Behavior*, Vol. 26, pp. 474–481.

Wiener, E. L. (1988), "Cockpit Automation," in *Human Factors in Aviation*, E. L. Wiener and D. C. Nagel, Eds., Academic, San Diego, CA, pp. 433–461.

Wiener, E. L., and Curry, R. E. (1980), "Flight-Deck Automation: Promises and Problems," *Ergonomics*, Vol. 23, pp. 955–1011.

Wilson, G., Finomore, Jr., V., and Estepp, J. (2003), "Transcranial Doppler Oximetry as a Potential Measure of

Cognitive Demand," in *Proceedings of the 12th International Symposium on Aviation Psychology*, Dayton, OH, pp. 1246–1249.

Wilson, G. F. (2000), "Strategies for Psychophysiological Assessment of Situation Awareness," in *Situation Awareness: Analysis and Measurement*, M. R. Endsley and D. J. Garland, Eds., Erlbaum, Mahwah, NJ, pp. 175–188.

Wilson, G. F. (2002a), "A Comparison of Three Cardiac Ambulatory Recorders Using Flight Data," *International Journal of Aviation Psychology*, Vol. 12, pp. 111–119.

Wilson, G. F. (2002b), "An Analysis of Mental Workload in Pilots during Flight Using Multiple Psychophysiological Measures," *International Journal of Aviation Psychology*, Vol. 12, pp. 3–18.

Wilson, G. F., and Eggemeier, F. T. (1991), "Physiological Measures of Workload in Multi-Task Environments," in *Multiple-Task Performance*, D. Damos, Ed., Taylor & Francis, London, pp. 329–360.

Wu, C., Liu, Y., and Quinn-Walsh, C. M. (2008), "Queuing Network Modeling of a Real-Time Psychophysiological Index of Mental Workload—P300 in Event-Related Potential (ERP)," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, Vol. 38, pp. 1068–1084.

Xie, B., and Salvendy, G. (2000a), "Prediction of Mental Workload in Single and Multiple Tasks Environments," *International Journal of Cognitive Ergonomics*, Vol. 4, pp. 213–242.

Xie, B., and Salvendy, G. (2000b), "Review and Reappraisal of Modelling and Predicting Mental Workload in Single- and Multi-Task Environments," *Work and Stress*, Vol. 14, pp. 74–99.

Yeh, Y.Y., and Wickens, C.D. (1988), "Dissociation of Performance and Subjective Measures of Workload," *Human Factors*, Vol. 30, pp. 111–120.

# CHAPTER 9

# SOCIAL AND ORGANIZATIONAL FOUNDATIONS OF ERGONOMICS

**Alvaro D. Taveira**
University of Wisconsin–Whitewater
Whitewater, Wisconsin

**Michael J. Smith**
University of Wisconsin–Madison
Madison, Wisconsin

## 1  INTRODUCTION

The importance and influence of social context have been debated recurrently in attempts to define the ergonomics discipline. Wilson (2000, p. 560) defines *ergonomics* "as the theoretical and fundamental understanding of human behavior and performance in purposeful interacting socio-technical systems, and the application of that understanding to the design of *interactions* in the context of real settings." The International Ergonomics Association (IEA) defines ergonomics (or human factors) as the scientific discipline concerned with the understanding of *interactions* among humans and other elements of a system and the profession that applies theory, principles, data, and methods to design in order to optimize human well-being and overall system performance (IEA, 2000).

Ergonomics focuses on *interactive* behavior and the central role of human behavior in complex interacting systems. In this chapter we argue that these behaviors are deeply immersed in and cannot be separated from their social context.

When one examines the assumptions held by researchers and practitioners about the fundamentals of ergonomics, it is no surprise that the role of social

context remains a focus of discussion. Daniellou (2001) described some of these diverging tacit conceptualizations on *human nature, health*, and *work*, which reflect on distinct research models and practices and on different consequences to the public. He said that when referring to the *human nature* ergonomists may have in mind a biomechanical entity, an information-processing system, a subjective person with unique psychological traits, or a social creature member of groups that influence his or her behaviors and values. Similarly, the concept of *health* may be thought of as the absence of recognized pathologies, which would exclude any notion of discomfort, fatigue, or poverty, or could be defined in more comprehensive fashion as a general state of well-being (i.e., physical, mental, and social) or even interpreted as a process in a homeostatic state. Daniellou (2001) observed that, although many ergonomists today embrace a definition of *work* that includes both its physical and cognitive aspects, other important distinctions remain. Work is often considered as the task and work environment requirements, a specific quantifiable definition of what is demanded from all workers to accomplish a given target. Another perspective on work is given by examination of the workload from the perspective of each worker, a vantage point that

emphasizes the individual and collective strategies to manage the work situation in a dynamic fashion. Finally, some ergonomists focus on the ethical aspects of work, on its role on the shaping of individual and societal character. Daniellou reminded us of the challenge of producing work system models which are necessarily reductions of the actual world while ensuring that the reduction process does not eliminate the essential nature of the larger context.

Discussion of the role of social context in ergonomics can be possibly described within a continuum between two opposite perspectives. On one side there is a view of ergonomics as an applied science, a branch of engineering, a practice, even an art, which is deeply embedded in and influenced by a larger social context (Perrow, 1983; Oborne et al., 1993; Moray, 1994; Shackel, 1996; Badham, 2001). On the other side, we have authors who see ergonomics as applying the same standards of the natural and physical sciences and as a *science* is able to produce overarching laws and principles that are able to predict human performance. It is a laboratory-based discipline in which contextual factors have minimal influence (e.g., Meister, 2001). We maintain that the latter position poses significant limitations for the relevance of ergonomics as a practice because it severely limits its usefulness and impact in the real world.

Ergonomics is a discipline with a set of fundamentals derived from the scientific method application, but these cannot be isolated from a vital applied (situational) component. We understand that the development of ergonomics is strongly affected by the social environment, and its permanence as a workable modern discipline depends on its sensitivity to economic, legal, technological, and organizational contexts.

Ergonomics is an integral part of a social enterprise to optimize work systems by improving productivity, quality, safety, and working conditions. It contributes to the complete and uniform utilization of human and technical resources and to the reliability of production systems. It is central to the quest for better work conditions by reducing exposure to physical hazards and the occurrence of fatigue. Ergonomic efforts play a fundamental role in the maintenance of workforce health by reducing the prevalence of musculoskeletal disorders and controlling occupational stress. It answers to market needs of higher product value, useful product features, and increased sales appeal. Ergonomics is ultimately instrumental to the advance and welfare of individuals, organizations, and the broader society.

One could draw an analogy between the role of social forces in technological development and of these social forces in the evolution of the ergonomics discipline. Technological development has been seen as an important source of social change. In the traditional view, technology was considered to be the progressive application of science to solve immediate technical problems, and their release in the marketplace in turn had social impacts. In summary, "technological determinism" saw technology as a natural consequence of basic science progress and, in many circumstances, a cause of changes in society. Conversely, a more recent view

of technological development sees it as part of a social system with specific needs and structure and not driven exclusively by basic science (Hughes, 1991). The latter perspective is also germane to the development of ergonomics (technology), in which social demands and structure marked its evolution. For example, the introduction of an (ergonomic) technological improvement such as stirrups allowed horse riders to hold themselves steady and freed both hands for work or warfare (Pacey, 1991). From a deterministic view this innovation accounted for some of the success of nomadic groups in India and China in the thirteenth century and for the permanence of feudalism in Western Europe. From a systems perspective, on the other hand, this innovation can be conceived as part of a social system and one of the by-products of the critical reliance on horses by these nomadic groups, which later fit the needs of feudalism in Western Europe (White, 1962).

In the next section we examine the path of ergonomics to become an established discipline and how this course was shaped by social context.

## 2 HISTORICAL PERSPECTIVE

The development over time of the ergonomics discipline can be clearly traced to evolving societal needs. Ergonomics as a practice started at the moment the first human groups selected or shaped pieces of rock, wood, or bone to perform specific tasks necessary to their survival (Smith, 1965). Tools as extensions of their hands allowed these early human beings to act on their environment. Group survival relied on this ability, and the fit between hands and tools played an important role in the success of the human enterprise. The dawning of the discipline is therefore marked by our early ancestors' attempts to improve the fit between their hands and their rudimentary tools and to find shapes that increased efficiency, dexterity, and the capacity to perform their immediate tasks effectively.

The practice of ergonomics has been apparent as different civilizations across history used ergonomics methods in their projects. Historically, body motions have been a primary source of mechanical power, and its optimal utilization was paramount to the feasibility of many individual and collective projects. Handles, harnesses, and other fixtures that matched human anatomy and task requirements allowed the use of human power for innumerous endeavors. Tools and later simple machines were created and gradually improved by enhancing their fit to users' and tasks' characteristics (Smith, 1965).

Although human work was recognized as the source of economic growth, workplaces have always been plagued with risks for workers. The groundbreaking work of Ramazzinni (1700) pointed out the connections between work and the development of illnesses and injuries. In contrast to the medical practitioners of his time who focused exclusively on the patients' symptoms, Ramazzini argued for the analysis of work and its environment as well as the workers themselves and their clinical symptoms. In his observations he noted that some common (*musculoskeletal*) disorders seem to follow from prolonged, strenuous, and unnatural

physical motions and protracted stationary postures of the worker's body. Ramazzini's contributions to the development of modern ergonomics were enormous, in particular his argument for a systematic approach to work analysis.

Concomitantly with the increasing effectiveness of human work obtained through improved tools and machines, a growing specialization and intensification of work were also realized. The advent of the factory, a landmark in the advancement of production systems, was a social rather than a purely technical phenomenon. By housing a number of previously autonomous artisans under the same roof it was possible to obtain gains in productivity by dividing the work process into smaller, simpler tasks (Smith, 1776; Babbage, 1832). This process of narrowing down the types of tasks (and motions) performed by each worker while increasing the pace of the work activities had tremendous economic impact but was gradually associated with health and social welfare issues for the workforce. This became more noticeable with the beginning of the Industrial Revolution, when significant changes in technology and work organization occurred and larger portions of the population engaged in factory work.

Industrialization brought along an accelerated urbanization process and an increased access to manufactured goods. In the United States, in particular, this endeavor allowed for the tremendous growth in the supply and demand for capital and consumer goods. Some concern with the incidence of injuries and illnesses related to the industrialization process and its influence on the national economy (i.e., damage to national human resources) led to some early attempts to improve safety in factories in the nineteenth century (Owen, 1816).

During the Industrial Revolution almost exclusive attention was paid to production output increase. This *productivity enhancement* thrust crystallized in the work of Taylor (1912) and his *scientific management* approach. Taylor compiled and put in practice one of the first systematic methods for work analysis and design, focusing on improving efficiency. His approach emphasized the analysis of the work situation and the identification of the "one best way" to perform a task. This technique called for the fragmentation of the work into small, simple tasks that could be performed by most people. It required analysis and design of tasks, development of specialized tools, determination of lead times, establishment of work pace, specification of breaks, and work schedules.

This effort to better utilize technology and personnel had profound effects on the interactions between humans and their work. It could be argued that expansion of output through increased efficiency was socially preferable to long work hours (Stanney et al., 1997, 2001), but one should not lose sight of the fact that the higher economic gains could be derived from the former, which provided a more compelling justification for work intensification.

Taylor's work represented a crucial moment for the ergonomics discipline, although this development had a lopsided emphasis on the shop-floor efficiency aspects. Frank and Lillian Gilbreth (1917) further elaborated the

study of human motions (e.g., micromotions or Therbligs) and provided the basis for time-and-motion methods. Concurrently with the focus on the worker and workstation efficiency, the need to consider the larger organizational structure to optimize organizational output was also recognized. Published at almost the same time as Taylor's *The Principles of Scientific Management*, Fayol's *General and Industrial Management* (1916) defined the basics of work organization. Fayol's principles were in fact very complementary to Taylor's, as the latter focused on production aspects whereas the former addressed organizational issues.

Although preceded by Ransom Olds by almost 10 years, it is ultimately Henry Ford's assembly line that best epitomized this early twentieth-century work rationalization drive. In Ford's factories, tasks were narrowly defined, work cycles drastically shortened, and the pace of work accelerated. Time required per unit produced dropped from 13 to 1 h, and costs of production were sharply reduced, making the Ford Model T an affordable and highly popular product (Konz and Johnson, 2004). Despite the high hourly wage rates offered, about double the prevalent wage rate of the time, turnover at Ford's facilities was very high in its early years. Concern with the workers' health and well-being at this juncture was limited to preventing acute injury and some initial concern with fatigue (Gilbreth and Gilbreth, 1920).

The institution of ergonomics as a modern discipline has often been associated with the World War II period, when human–complex systems interactions revealed themselves as a serious vulnerability (Christensen, 1987; Sanders and McCormick, 1993; Chapanis, 1999). The operation of increasingly sophisticated military equipment was being compromised repeatedly by the lack of consideration to the human–machine interface (Chapanis, 1999). Endeavors to address these issues during World War II and in the early postwar period led to the development and utilization of applied anthropometrics, biomechanics, and the study of human perception in the context of display and control designs, among other issues. The ensuing Cold War and Space Race period from 1950 to 1980 provided a powerful impetus for rapid expansion of ergonomics in the defense and aerospace arenas, with a gradual transfer of that knowledge to civil applications.

Increasingly during the twentieth century, the scientific management approach became the dominant paradigm, with several nations reaping large economic benefits from its widespread adoption. In this period, ergonomics practice experienced a tremendous growth of its knowledge base at the task and human–machine interface levels. Over time, however, this *work rationalization* led to a prevalence of narrowly defined jobs, with typical cycle times reduced to a few seconds. Repetition rates soared, and work paces became very intense. This increased specialization also resulted in the need for large support staffs and in a general underutilization of abilities and skills of the workforce. In fact, many jobs became mostly devoid of any meaningful content, characterized by fast and endless repetition of a small number of motions and by overwhelming monotony. This "dehumanization" of work was associated with

low worker morale, increasing labor–management conflicts, as well as the growing problem of musculoskeletal disorders.

It has been argued that much of the success of the scientific management approach was due to its suitability to social conditions in the early twentieth century, when large masses of uneducated and economically deprived workers were available (Lawler, 1986). Decades of industrialization and economic development changed this reality significantly, creating a much more educated and sophisticated workforce dealing with a much more complex work environment. These workers had higher expectations for their work content and environment than Taylor's approach could provide. In addition, there was growing evidence that this rigid work organization was preventing the adoption and effective utilization of emerging technologies (Trist, 1981). Some researchers and practitioners started to realize that this mechanistic work organization provided little opportunity for workers to learn and to contribute to the improvement of organizational performance (McGregor, 1960; Argyris, 1964; Smith, 1965). One could infer that during that period a micro-optimization of work (i.e., at the task and workstation level) was actually going against broader systems performance since it underutilized technological and human resources, exposed the workforce to physical and psychological stressors, burdened national health systems, and ultimately jeopardized national economic development (Hendrick, 1991).

## 3 BRINGING SOCIAL CONTEXT TO THE FOREFRONT OF WORK SYSTEMS

The social consequences of managerial decisions in the workplace were addressed systematically for the first time by Lewin starting in the 1920s (Marrow, 1969). In his research he emphasized the human(e) aspects of management, the need to reconcile scientific thinking and democratic values, and the possibility of actual labor–management cooperation. He revolutionized management (consulting) practice by advocating an *intervention orientation* to work analysis. Lewin believed that work situations should be studied in ways that make participants ready and committed to act (Weisbord, 2004). He was concerned with how workers find meaning in their work. This search for meaning led him to argue for the use of ethnographic methods in the study of work. He was a pioneer in proposing worker involvement in the analysis and (re)design of work and in defining job satisfaction as a central outcome for work systems. Lewin laid the seeds for the concept of interactive systems and highlighted the need for strategies to reduce resistance to change in organizations. Lewin could rightly be named as a precursor of the study of psychosocial factors in the workplace.

Lewin's research set the ground for the development of macroergonomics decades later. In particular, Lewin's work was a forerunner to participatory ergonomics, as he defined workers as legitimate knowledge producers and agents of change. Lewin understood that workers could "learn how to learn" and had a genuine interest in improving working conditions and, by consequence, human–systems interactions. He saw work improvement not simply as a matter of shortening the workday but as one of increasing the human value of work (Weisbord, 2004).

Introduction of general systems theory by Bertalanffy (1950, 1968) and its subsequent application to several branches of science had profound implications for ergonomics theory and practice. The seminal work by researchers at the Tavistock Institute (Emery and Trist, 1960; Trist, 1981) and establishment of sociotechnical systems theory, especially, made clear the importance of the social context in work systems optimization.

The proponents of sociotechnical systems considered the consequences of organizational choices on technical, social, and environmental aspects. Their studies indicated that the prevailing work organization (i.e., scientific management) failed to fully utilize workers' skills and was associated with high absenteeism and turnover, low productivity, and poor worker morale (Emery and Trist, 1960). They saw Taylorism as creating an imbalance between the social and technical components of the work system and leading to diminishing returns on technical investments. Sociotechnical researchers advocated the reversal of what they saw as extreme job specialization, which often led to underutilized work crews and equipment and to high economic and social costs. They argued for organizational structures based on flexible, multiskilled workers (i.e., knowledgeable in multiple aspects of the work system) operating within self-regulated or semiautonomous groups.

According to sociotechnical principles, high performance of the technical component at the expense of the social component would lead to dehumanization of work, possibly to a situation where some segments of society would enjoy the economic benefits of work and another (larger) portion would bear its costs (i.e., work itself!). The opposite situation, where the social component becomes preponderant, would be equally troubling because it would lead to system output reduction, with negative effects on organizational and national economies. In summary, a total system output decline could be expected in both cases of suboptimization.

Sociotechnical systems (STSs) emphasized the match between social needs and technology, or, more specifically, improvement in the interaction between work systems' technical and social components. Sociotechnical systems focused on the choice of technologies suitable to the social and psychological needs of humans. The path to improved integration was pursued, with the maximization of worker well-being as its primary system optimization criteria. An alternative path to the same end was proposed decades later in the macroergonomic approach, which posited the quick and effective adoption of new technology as the crux of organizational survival and the primary optimization criteria (Hendrick, 1991, 1997). The macroergonomic approach urged organizations to implement job and organization designs that increased the chances of successful technology implementation by utilizing its human resources fully. Although providing different rationales, the two propositions had similar visions (i.e., effective and healthy work systems), achieved through analogous work organization designs.

A major contribution of sociotechnical systems to an understanding of work systems (organizations) was that they were intrinsically open systems which needed to exchange information, energy, or materials with their environment to survive (Bertalanffy, 1950). A second major contribution was the recognition of fast and unpredictable changes in the environmental contexts themselves. Emery and Trist (1965) describe four different types of causal texture, which refer to different possible dispositions of interconnected events producing conditions that call for very diverse organizational responses. The *placid, randomized environment* was characterized by relatively stable and randomly distributed *rewards and penalties*. This implied a strategy that was indistinguishable from tactics, where the entity or organization attempted to do its best in a purely local basis. These organizations survived adaptively and tended to remain small and independent in this environment. The *placid, clustered environment* implied some type of knowable (i.e., nonrandom) distribution of rewards and penalties. This environment justified the adoption of a strategy distinct from tactics and rewarded organizations that anticipated these clusters. Successful organizations in this context tended to expand and to become hierarchical. The *disturbed-reactive environment* indicated the existence of multiple organizations competing for the same resources and trying to move, in the long term, to the same place in the environment. Competitive advantage was the dominant strategy in this situation. Finally, *turbulent fields* were marked by accelerating changes, increasing uncertainty, and unpredictable connections of environmental components (Weisbord, 2004). In Emery and Trist's (1965) words: "The ground is in motion," which creates a situation of *relevant uncertainty*. This circumstance demanded the development of new forms of data collection, problem solving, and planning from the organization.

A significant offshoot of the sociotechnical theory was the *industrial democracy* movement, which had considerable influence in organizational and governmental policies and labor–management relations in several Scandinavian and Western European countries in the 1970s and 1980s. Industrial democracy proponents argued that workers should be entitled to a significant voice in decisions affecting the companies in which they worked (Gardell, 1982; Johnson and Johansson, 1991). The term has several connotations, but essentially it advocated union rights to representation on the boards of directors of large companies. It has also been used to describe various forms of consultation, employee involvement, and participation (Broedling, 1977).

A related concept, *codetermination*, achieved prominent status in Germany and Sweden in the 1970s, when legislation expanded the right of workers to participate in management decisions. Codetermination typically implied having an organization's board of directors with its composition of up to 50% of employee representatives. In the German experience, codetermination applied to both the plant and industry levels, and the proportion of worker representation in executive boards varied depending on the business size and type. Whereas industrial democracy urged increased workforce participation across the entire organization, codetermination was concerned primarily with worker influence at the top of the organization.

Sociotechnical theory deeply influenced another important societal and organizational initiative, the *quality of work life* (QWL) movement. The QWL concept has its roots in the somewhat widespread dissatisfaction with the organization of work in the late 1960s and early 1970s. Taylorism's inflexibility, extreme task specialization, and the resulting low morale were identified as the primary sources of these issues. A number of models of job and organization design attempting to improve utilization of worker initiative and to reduce job dissatisfaction were developed and put in practice, particularly in the Scandinavian countries. QWL initiatives called for organizational changes involving increased task variety and responsibility, making the case for increasing worker participation as a potent intrinsic motivator.

Of particular relevance was the experimental assembly plant established in 1989 by Volvo in Uddevalla, Sweden. The project was an attempt to apply sociotechnical principles to a mass production facility on a wide scale. In the context of a very tight job market, this initiative tried to make jobs and the work environment more attractive to workers by increasing autonomy and group cohesiveness and by recovering some of the meaning lost by extreme task fragmentation. Despite tremendous expectations from company, labor, and sociotechnical researchers, the plant was closed in 1993 because of inferior performance. Although some debate remains about the causes of the plant closure, one could infer that increased pressure for short-term financial returns, larger labor availability, globalization, and fierce international competition were primary contributing factors. Most of all, the end of the Uddevalla plant highlighted the difficulties of achieving acceptable compromises or, preferably, converging strategies between enhanced quality of work life and organizational competitiveness (Huzzard, 2003). The challenge of competitiveness remained an important one for the viability of QWL. In Huzzard's words (p. 93): "Despite the evidence that firms can reap considerable performance advantages through attempts at increasing the quality of working life through greater job enlargement, job enrichment, competence development and participation, there is also considerable evidence that some firms are actually eschewing such approaches in deference to short-run pressure for immediate results on the 'bottom-line' of the profit and loss account and rapid increases in stock market valuation."

The sociotechnical theory helped to consolidate the objectives of ergonomics by virtue of its *joint-optimization principle*, which established the possibility and the need for work systems to achieve concurrently high social and technical performance. In other words, it advocated that through a work organization focused on human physical and social needs (i.e., free of hazards, egalitarian, team based, semiautonomous) it would be possible to attain high productivity, quality, and reliability while reaching high levels of job satisfaction, organizational cohesiveness, and mental and physical well-being. The affinities between ergonomics and the

sociotechnical theory are quite evident, as the former is a keystone to the feasibility of joint optimization. In fact, contemporary ergonomics interventions are deeply influenced by sociotechnical theory calling for worker involvement, delegation of operational decisions, and a decentralized management structure (Cohen et al., 1997; Chengalur et al., 2004; Konz and Johnson, 2004).

The sociotechnical approach inspired a number of successful work systems performance improvement initiatives over recent decades. These initiatives often accentuated the need for decision-making decentralization and the enlargement of workers' understanding of the entire system and of the technical and economic effects of their actions on it. These approaches advocated jobs with increased amounts of variety, control, feedback, and opportunity for growth (Emery and Trist, 1960; Hackman and Oldham, 1980). Workers needed to recover their perception of the whole and to be encouraged to regulate their own affairs in collaboration with their peers and supervisors. These STS-inspired initiatives aimed at eliminating the inefficiencies and bottlenecks created by decades of scientific management practice, which shaped an inflexible and highly compartmentalized workforce.

Not all STS-inspired programs were equally successful, and many of them were withdrawn after some years. Proposals that relied heavily on worker participation but did not provide adequate support and guidance were found to be ineffective and stressful to workers. Fittingly, Kanter (1983) observed that many of the participatory management failures were caused by too much emphasis on *participation* and too little on *management*. Similarly, in some cases workers found themselves overwhelmed by excessive job variety. Studies have also shown that job control is not always sufficient to attenuate workload effects (Jackson, 1989). Finally, experiments with the concept of industrial democracy in Sweden, Norway, and Germany, where labor representatives participated in executive boards, produced little in the way of increasing rank-and-file workers' involvement and improving the social meaning of their jobs and did not result in competitive advantages in productivity or quality, at least not in the short term.

## 4 ERGONOMICS AND THE ORGANIZATION

In the 1980s a number of practitioners and researchers started to realize that some ergonomic solutions considered to be superior at the workstation level failed to produce relevant outcomes at the organizational level (Dray, 1985; Carlopio, 1986; Brown, 1990; Hendrick, 1991, 1997). These failures were attributed to a narrow scope of analysis typical of these ergonomics interventions, which neglected to consider the overall organizational structure (DeGreene, 1986). A new area within the ergonomics discipline was conceived, under the term *macroergonomics* (Hendrick, 1986). According to Hendrick (1991, 1997), macroergonomics emphasized the interface between organizational design and technology with the purpose of optimizing work systems performance. Hendrick saw macroergonomics as a top-down sociotechnical approach to the design of work

systems. Imada (1991) stated that this approach recognized that organizational, political, social, and psychological factors of work could have the same influence on the adoption of new concepts as the merits of the concepts themselves. Finally, Brown (1991) conceived macroergonomics as addressing the interaction between the organizational and psychosocial contexts of a work system with, and emphasis on, the fit between organizational design and technology. Effective absorption of technology is at the forefront of macroergonomic endeavors.

Although having many of its concepts derived from sociotechnical theory, macroergonomics diverged from the former in some significant aspects. Whereas macroergonomics was a top-down approach, sociotechnical systems embraced a bottom-up approach, where the workstation is the building block for organizational design. Macroergonomics saw macrolevel decisions as a prerequisite to microlevel decisions (Brown, 1991), in sharp contrast with the STS. The STS affirmed that joint optimization must first be constructed into the primary work system or it would not become a property of the organization as a whole (Trist, 1981).

Hendrick (1991) defined organizational design around three concepts: complexity, formalization, and centralization. *Complexity* referred to an organization's degree of internal differentiation and extent of use of integration and coordination mechanisms. Internal differentiation was further elaborated into horizontal differentiation, which referred to job specialization and departmentalization; vertical differentiation, which related to the number of hierarchical levels in the organization; and spatial dispersion. *Formalization* conveyed the reliance on written rules and procedures. *Centralization* referred to the degree of dispersion of decision-making authority. Although some given combinations of these structural elements seemed to be suitable to some types of organizations pursuing specific goals, a number of possible interactions may produce results that were difficult to predict.

Hendrick (1997) pointed out three work system design practices that characteristically undermined ergonomic efforts. The first problematic practice related to a situation where technology (hardware or software) was taken as a given and user considerations came as an afterthought. These were efforts that typically overlooked motivational and psychosocial aspects of users. The second practice was also related to the overriding attention to the technical component, where technical feasibility was the only criterion for function allocation. In this situation, optimization of the technical component forced the leftover functions on the social (human) element (DeGreene, 1986). In other words, users were forced to accommodate the remaining tasks, frequently resulting in work situations void of factors that produced satisfaction and identity. Finally, there was a failure to consider adequately the four elements (subsystems) of sociotechnical systems: personnel, technology, organizational structure, and the external environment.

Methods employed in macroergonomics were numerous and were typically embedded in a four-step analysis/assessment, design, implementation, and evaluation

(Brown, 1991). For a review of methods in macro-ergonomics, see Hendrick and Kleiner (2001) and Hendrick (2002).

# 5   PARTICIPATORY ERGONOMICS

As discussed previously, sociotechnical theory provided several arguments and examples supporting the benefits of worker participation, particularly through teamwork, as an effective work system design strategy. The germinal study, which gave the essential evidence for development of the sociotechnical field, was a work group experiment observed in the English mining industry during the 1950s. In this new form of work organization a set of relatively autonomous work groups performed a complete collection of tasks, interchanging roles and shifts and regulating their affairs with a minimum of supervision. This experience was considered a way of recovering group cohesion and self-regulation concomitantly with a higher level of mechanization (Trist, 1981). The group had the power to participate in decisions concerning work arrangements, and these changes resulted in increased cooperation between task groups, personal commitment from participants, reduction in absenteeism, and fewer accidents. From the sociotechnical perspective, participation permitted the ordering and utilization of worker-accumulated experience; it validated and legitimized this experiential knowledge.

Another strong defense of employee participation was made by Sashkin (1984, p. 11), arguing that "participatory management has positive effects on performance, productivity, and employee satisfaction because it fulfills three basic human needs: increased autonomy, increased meaningfulness, and decreased isolation." Perhaps the most original and polemic of the Sashkin's contributions to the subject was his statement that participation was an "ethical imperative." His reasoning was that basic human needs were met by participation, and denial of the process produced psychological and physical harm to the workers. Research on worker participation is extensive and started as early as the 1940s (Coch and French, 1948) and is punctuated by controversial and often ideological debate (Locke and Schweiger, 1979; Sashkin, 1984; Cotton et al., 1988; Leana et al., 1990). For an extensive review on participation and teamwork, refer to Medsker and Campion (2001) and Sainfort et al. (2001).

Aspects of employee participation and participatory ergonomics can be seen in total quality management (TQM) approaches. TQM relied on teamwork for problem solving and change implementation related to quality and production issues (Dean and Bowen, 1994). Over time, some of those teams also started to focus on working conditions. The term *participatory ergonomics* (PE) originated from discussions among Noro, Kogi, and Imada in the 1980s (Noro, 1999). It assumed that ergonomics was bounded by the degree to which people were involved in conducting its practice. According to Imada (1991), PE required users (the real beneficiaries of ergonomics) to be involved directly in developing and implementing ergonomics. Wilson (1995, p. 1071) defined PE as "the involvement of people in planning and controlling a significant amount of their own work activities, with sufficient knowledge and power to influence both processes and outcomes in order to achieve desirable goals." PE has been identified with the use of participative techniques in the analysis and implementation of ergonomics solutions (Wilson and Haines, 1997) and recognized as an approach to disseminate ergonomics knowledge throughout the organization (Noro, 1991).

Imada (1991) pointed out three major arguments in support of worker involvement in ergonomics. First, because ergonomics was an intuitive science, which in many cases simply organized knowledge that workers were already using, it could validate workers' accumulated experience. Second, people were more likely to support and adopt solutions for which they felt responsible. Involving users and workers in the ergonomic process had the potential to transform them into makers and supporters of the process rather than passive recipients. Finally, developing and implementing technology enabled workers to modify and correct problems continuously.

Participatory ergonomics saw end users' contributions as indispensable elements of its scientific methodology. It stressed the validity of simple tools and workers' experience in problem solution and denied that these characteristics resulted in nonscientific outcomes (Imada, 1991; Taveira and Hajnal, 1997). In most situations, it was believed that employees or end users were in the best position to identify the strengths and weaknesses of work situations. Their involvement in the analysis and redesign of their workplace could lead to better designs as well as to increase theirs and the company's knowledge of the process.

Participatory ergonomics was also conceived as an approach to enhance the human–work systems fit. Work environments have become highly complex, often beyond the capacity of individual workers. This mismatch between workers' capabilities and work systems requirements was shown to be an important factor in organizational failures (Weick, 1987; Reason, 1990, 1997; Reason and Hobbs, 2003). A possible strategy to address this imbalance was to pool workers' abilities through group teamwork, making them collectively more sophisticated (Imada, 1991). Other beneficial outcomes of PE included increased commitment to change (Lawler, 1986; Imada and Robertson, 1987), increased learning experiences (i.e., reduced training costs) and improved performance (Wilson and Grey Taylor, 1995), and increased job control and skills (Karasek and Theorell, 1990).

# 6   ERGONOMICS AND QUALITY IMPROVEMENT EFFORTS

A growing amount of attention among ergonomics scholars and practitioners concerns whether and how ergonomics is affected by organizational transformations. In particular, the integration of ergonomics and quality management programs has been discussed extensively (Drury, 1997, 1999; Eklund, 1997, 1999; Axelsson et al., 1999; Taveira et al., 2003). It seems clear, at

least in principle, that the two are related and interact in a variety of applications, such as inspection, process control, safety, and environmental design (e.g., Drury, 1978; Eklund, 1995; Rahimi, 1995; Stuebbe and Houshmand, 1995; Warrack and Sinha, 1999; Axelsson, 2000). There is some consensus that the form of this relationship is one in which "good ergonomics" (e.g., appropriate workstation, job, and organization design) leads to improved human performance and reduced risk of injury, which in turn leads to improved product and process quality. Eklund (1995), for example, found that the odds of having quality deficiencies among ergonomically demanding tasks at a Swedish car assembly plant were 2.95 times more likely than for other tasks.

TQM practice seems to enlarge the employees' role by increasing their control over their activities and by providing them with information and skills and the opportunity to apply them. On the other hand, this positive evaluation of TQM is not shared unanimously. Some believe that TQM is essentially a new package for the old Taylorism, with the difference that *work rationalization* is being made by the employees themselves (Parker and Slaughter, 1994). Lawler et al. (1992), commenting on work simplification in the context of TQM, noticed that even though this emphasis seems to work at cross purposes with job enrichment, most TQM concepts have not referred methodically to the issue of how jobs should be designed. Lawler and his colleagues proposed that work process simplification can otherwise occur concurrently with the creation of jobs that have motivating characteristics (i.e., meaning, autonomy, feedback, etc.).

Implementation by the manufacturing industry of flexible production systems (e.g., just-in-time, lean manufacturing) has created a renewed demand for ergonomics. Flexible production systems are based on the assumption that regarding workers as repetitive mechanical devices will not provide any competitive advantage and that only skilled and motivated workers are able to add value to production. This approach embraced the concept of continuous improvement, although without totally rejecting Taylor's *one best way* (Reeves and Bednar, 1994). These systems saw line workers as capable of performing most functions better than specialists, allowing for a lean organization stripped of most personnel redundancies. These systems required that each step of the fabrication process be conducted perfectly every time, thus reducing the need for buffer stocks while also producing a higher quality end product.

Information technology has been connected to changes in internal organization because it facilitates the work of multidisciplinary teams whose members work together from the start of a job to its completion. Information systems make it efficient to push decision making down in the organization—to the teams that perform an organization's work. Efficient operations in the modern workplace call for a more equal distribution of knowledge, authority, and responsibility.

Social context, along with characteristics of the production process (i.e., type of industry), has a central role in the production philosophy embraced by organizations.

The choice of a particular production philosophy has in turn a substantial impact on the work organization, technology, human resources practices and policies, organizational efficiency, and the general quality of work life. Eklund and Berggren (2001) identified four generic production philosophies: Taylorism, sociotechnical systems, flexible production systems, and modern craft.

Most organizations use some mix of these approaches and rarely does one find an organization strictly employing a specific philosophy. Taylorism, Fordism, or scientific management is characterized by high levels of standardization, job and equipment specialization, close control over worker activities, machine-paced work, and short cycle times. This production philosophy allowed for tremendous efficiency gains but produced poor work conditions, marked by fatigue, monotony, and poor utilization of workers capabilities. The second approach, sociotechnical systems (STS), addressed some of the Taylorism shortcomings by expanding job content in terms of its variety and meaningfulness, by reducing the rigidity of the production flow by adopting alternative production layouts and using buffers, and by transferring some decision-making power to workers. Although lessening some of the ergonomics issues associated with Taylorism, STS did not effectively address some of workload-related issues, had difficulty in maintaining high levels of output, and the use of buffers came at the cost of additional work in process. Flexible production systems were associated with the quality movement discussed above and were characterized by flexible equipment with quick setup times, a multiskilled workforce able to perform a variety of tasks, teamwork, and continuous improvement. Finally, modern craft was typically limited to large, specialized capital equipment production involving highly skilled workers and high-precision tools in very long work cycles. In many respects it was very representative of pre–Industrial Revolution work organization.

## 7 SOCIALLY CENTERED DESIGN

As defined by Stanney et al. (1997, p. 638), socially centered design was an approach concerned with system design variables that "reflect socially constructed and maintained world views which both drive and constrain how people can and will react to and interact with a system or its elements." It is a strategy aimed at filling the gap between system-centered approaches and user-centered approaches to job and organization design. In other words, it was conceived of as a bridge between a macro- and a microorientation to ergonomics.

Although conceding that the macroergonomic approach attempted to alleviate some of the issues associated with widespread adoption of scientific management, these authors argued that it did not sufficiently address group and intergroup interactions. In fact, the authors tend to group Taylorism and macroergonomics at one end of the technical versus social spectrum, as both consider the worker or user as a resource to be optimized (Stanney et al., 1997). User-centered design was positioned at the other end of the spectrum, as it regarded the users' abilities, limitations, and preferences as the

key design objectives (i.e., enabling user control as the primary goal). User-centered methodologies emphasized human error avoidance but generally ignored the social context of work by focusing instead on the individual and his or her workstation.

Socially centered design focuses on real-time interactions among people in the context of their work practices. In the Lewin tradition, the use of naturalistic methods was advocated to identify informal skills employed in the work process. It was assumed that the identification of these "unofficial" organizational and social factors added value to system design (Grudin, 1990). It saw workers' roles and responsibilities as characterized dynamically by ever-changing local interactions. Similarly, the concept of optimization was local and contingent, being defined by the system context. Socially centered design looked at artifacts (i.e., objects and systems) as solutions to problems. Considering that these problems were situated in a social context, one could expect that different groups in different situations may use the same artifacts in dissimilar ways. It led to the conclusion that artifacts must be examined from both a physical and a social perspective.

The methodologies for data acquisition advocated by this approach focused on group processes related to artifact use. These methods took into consideration the influence of the social work environment on system design effectiveness. Socially centered design saw everyday work interactions as relevant to effective system design and these factors must be extracted through contextual inquiries (Stanney et al., 1997). The most common techniques included group studies and ethnographic studies (Jirotka and Goguen, 1994). These methods were time and resource consuming, and transference of acquired knowledge to other situations required further refinement since these were locally generated explanations.

So far we have examined some of the critical aspects of the social foundations of ergonomics. The views described were rooted primarily in the seminal work of Kurt Lewin and of researchers at the Tavistock Institute, particularly Eric Trist and Fred Emery. Next we review approaches that, although related to the same tradition, put a stronger emphasis on how the social environment affects workers' mental and physical health.

## 8   PSYCHOSOCIAL FOUNDATIONS

For more than a century, medical practitioners have known that social, psychological, and stress factors influence the course of recovery from disease. During World Wars I and II, much attention was placed on how social, physical, and psychological stress affected soldier, sailor, and pilot performance, motivation, and health. Then, beginning in the 1950s and continuing until today, much attention has been paid to the relationship between job stress and employee ill health (Caplan et al., 1975; Cooper and Marshall, 1976; Smith, 1987a; National Institute for Occupational Safety and Health (NIOSH), 1992; Smith et al., 1992; Kalimo et al., 1997; Kivimaki and Lindstron, 2006). What has emerged is an understanding that "psychosocial"

attributes of the environment can influence human behavior, motivation, performance, and health and psychosocial factors have implications for human factors design considerations. Several conceptualizations have been proposed over the years to explain the human factors aspects of psychosocial factors and ways to deal with them. Next we will discuss some foundational considerations for social influences on human factors.

## 9   JOB STRESS AND THE PSYCHOSOCIAL ENVIRONMENT

Selye (1956) defined stress as a biological process created by social influences. The environment (physical and psychosocial) produces stressors that lead to adaptive bodily reactions by mobilizing energy, disease fighting, and survival responses. The individual's reactions to the environment are automatic survival responses (autonomic nervous system) and can be mediated by cognitive processes that are built on social learning. In Selye's concept, an organism undergoes three stages leading to illness. In the first, the state of *alarm*, the body mobilizes biological defenses to resist the assault of an environmental demand. This stage is characterized by high levels of hormone production, energy release, muscle tension, and increased heart rate. In the second stage, *adaptation*, the body's biological processes return to normal, as it seems that the environmental threat has been defeated successfully. In this second stage, the body is taking compensatory actions to maintain its homeostatic balance. These compensatory actions often carry a heavy physiological cost, which ultimately leads to the third stage. In the third and final phase, *exhaustion*, the physiological integrity of the organism is in danger. In this stage several biological systems begin to fail from the overwork of trying to adapt. These biological system failures can result in serious illness or death.

Much research has demonstrated that, when stress occurs, there are changes in body chemistry that may increase the risk of illness. Changes in body chemistry include higher blood pressure, increases in corticosteroids and peripheral neurotransmitters in the blood, increased muscle tension, and increased immune system responses (Selye, 1956; Levi, 1972; Frankenhaeuser and Gardell, 1976; Frankenhaeuser, 1986; Karasek et al., 1988; Shirom et al, 2008; Hansson et al., 2008; Asberg et al., 2009). Selye's pioneering research defined the importance of environmental stressors and the medical consequences of stress on the immune system, the gastrointestinal system, and the adrenal glands. However, Selye emphasized the physiological consequences of stress and paid little attention to the psychological aspects of the process or the psychological outcomes of stress.

Lazarus (1974, 1977, 1993, 1998, 1999, 2001) proposed that physiological changes caused by stressors came from a need for action resulting from emotions in response to the stressors (environment). The quality and intensity of the emotional reactions that lead to physiological changes depend on cognitive appraisal of the "threat" posed by the environment to personal security and safety. From Lazarus's perspective the

threat appraisal process determined the quality and intensity of the emotional reaction and defined coping activities that affected the emotional reactions. The extent of the emotional reaction influenced the number of physiological reactions.

Levi (1972) tied together the psychological and physiological aspects of stress. He recognized the importance of psychological factors as primary determinants of the stress sources (perception of stressors). He proposed a model that linked psychosocial stimuli with disease. In this model any psychosocial stimulus, that is, any event that takes place in the social environment, can act as a stressor. In accordance with a *psychobiological program*, psychosocial stimuli may evoke physiological responses similar to those described by Selye (1956). In turn, when these physiological responses occur chronically, they can lead to disease. Several intervening variables (individual susceptibility, personality, coping strategies, or social support) can moderate the link between psychosocial stimuli and disease. Levi emphasized that the sequence of events described by the model is not a one-way process but a *cybernetic system with continuous feedback*, a significant application of a human factors concept (Smith, 1966). The physiological responses to stress and disease can influence the psychosocial stimuli as well as the individual's psychobiological program. These feedback loops are important to an understanding of how stress reactions and disease states themselves can, in turn, act as additional stressors or mediate a person's response to environmental stressors.

Frankenhaeuser and her colleagues emphasized a psychobiological model of stress that defined the specific environmental factors most likely to induce increased levels of cortisol and catecholamines (Lundberg and Frankenhaeuser, 1980; Frankenhaeuser and Johansson, 1986). There are two different neuroendocrine reactions in response to a psychosocial environment: (1) secretion of catecholamines via the sympathetic–adrenal medullary system and (2) secretion of corticosteroids via the pituitary–adrenal-cortical system. Frankenhaeuser and her colleagues observed that different patterns of neuroendocrine stress responses occurred depending on the particular characteristics of the environment. They considered the most important environmental factors to be effort and the individual factors to be distress. The effort factor "involves elements of interest, engagements, and determination"; the distress factor "involves elements of dissatisfaction, boredom, uncertainty, and anxiety" (Frankenhaeuser and Johansson, 1986). Effort with distress is accompanied by increases in both catecholamine and cortisol secretion.

What Frankenhaeuser and her associates found was that effort without distress was characterized by increased catecholamine secretion but no change in cortisol secretion. Distress without effort was generally accompanied by increased cortisol secretion, with a slight elevation of catecholamines. Their approach emphasized the role of personal control in mediating the biological responses to stress. A lack of personal control over the stressors was almost always related to distress, whereas having personal control tended to stimulate greater effort. Studies performed by Frankenhaeuser and her colleagues showed that work overload led to increased catecholamine secretion but not to increased cortisol secretion when the employee had a high degree of control over the environment (Frankenhaeuser and Johansson, 1986).

Several studies have shown a link between elevated blood pressure and job stressors, especially workload, work pressure, and lack of job control. Rose et al. (1978) found that workload was associated with increased systolic and diastolic blood pressure. Van Ameringen et al. (1988) found that intrinsic pressures related to job content were related to increased standing diastolic blood pressure. The index of intrinsic pressures included a measure of quantitative workload (demands) and a measure of job participation (job control). Matthews et al. (1987) found that having few opportunities for participating in decisions at work was related to increased diastolic blood pressure. Longitudinal studies of job stress and blood pressure show that blood pressure increases were related to the introduction of new technologies at work and to work complexity (Kawakami et al., 1989). Schnall et al. (1990) demonstrated the link between hypertension and job stress and between emotions (e.g., anger and anxiety). James et al. (1986) demonstrated a relationship between emotions and increased blood pressure.

French (1963) and Caplan et al. (1975) have proposed a stress model that defined the interaction between the environment and the individual, including coping processes for controlling the external stressors. In their approach the development of stress is an outcome of an imbalanced interaction between the environmental resources that are available and the person's needs for resources. If the environmental demands are greater than a person's capacities and/or if the person's expectations are greater than the environmental supplies, stress will occur (Caplan et al., 1975; Cooper and Payne, 1988; Kalimo, 1990; Ganster and Schaubroeck, 1991; Johnson and Johansson, 1991; Cox and Ferguson, 1994). They proposed that social support from family and colleagues is a coping process that can mitigate the stress effects on health.

## 10 WORK ORGANIZATION AND PSYCHOSOCIAL INFLUENCES

The characteristics of a work organization are often sources of occupational stress that can lead to health consequences. Cooper and Marshall (1976) categorized these job stress factors into groups as those intrinsic to the job, the role in the organization, career development, the relationships at work, and the organizational structure and climate. Factors intrinsic to the job were similar to those studied and by Frankenhaeuser's group and French and Caplan. They included (1) physical working conditions; (2) workload, both quantitative and qualitative, and time pressure; (3) responsibilities (for lives, economic values, safety of other persons); (4) job content; (5) decision making; and (6) perceived control over the job. Smith and Carayon-Sainfort (1989) believed that the work organization level of psychosocial factors

dictated the extent of environmental exposure in terms of workload, work pace, work schedule, work–rest cycle, design of equipment and workstations, product and materials design, and environmental design. In addition, the psychosocial work environment affected a person's motivation to work safely, the attitude toward personal health and safety, and the willingness to seek health care. They postulated that the work organization defined the job stressors through the task demands, personal skill requirements, extent and nature of personal training, and supervisory methods that influenced the work methods used by employees (Carayon and Smith, 2000).

Many work organization factors have been linked to short- and long-term stress reactions. Short-term stress reactions include increased blood pressure, adverse mood states, and job dissatisfaction. Studies have shown a link between overload, lack of control and work pressure, and increased blood pressure (Matthews et al., 1987; Van Ameringen et al., 1988; Schnall et al., 1990). Other studies have found a link between job future uncertainty, lack of social support and lack of job control, and adverse mood states and job dissatisfaction (Karasek, 1979; Sainfort, 1989; Smith et al., 1992). Long-term stress reactions include cardiovascular disease and depression. Studies have shown that job stressors are related to increased risk for cardiovascular disease (Karasek, 1979; Karasek et al., 1988; Johnson, 1989). Carayon et al. (1999) have proposed that work organization factors can define or influence ergonomic risk factors for musculoskeletal: for example, the extent of repetition, force, and posture. Work organization policies and practices define the nature, strength, and exposure time to ergonomic risk factors.

According to Smith and Carayon-Sainfort (1989), stress results from an imbalance between various elements of the work system. This imbalance produces a load on the human response mechanisms that can produce adverse reactions, both psychological and physiological. The human response mechanisms, which include behavior, physiological reactions, and perception/cognition, act to exert control over the environmental factors that are creating the imbalance. These efforts to bring about balance, coupled with an inability to achieve a proper balance, produce overloading of

the response mechanisms that leads to mental and physical fatigue. Chronic exposure to these "fatigues" leads to stress, strain, and disease. This model emphasizes the effects of the environment (stressors), which can be manipulated to produce proper balance in the work system. These stressors can be categorized into one of the following elements of the work system: (1) the task, (2) the organizational context, (3) technology, (4) physical and social environment, and (5) the individual (see Figure 1).

The organizational context in which work is done often can influence worker stress and health (Landy, 1992). Career considerations such as over- and under-promotion, status incongruence, job future ambiguity, and lack of job security have been linked to worker stress (Cooper and Marshall, 1976; Cobb and Kasl, 1977; Jackson and Schuler, 1985; Sainfort, 1989 Israel, et al., 1996). In particular, companies that have the potential for reductions in the labor force (layoff or job loss) may be more susceptible to employees reporting more problems and more serious problems as an economic defense. These conditions may create a working climate of distrust, fear, and confusion that could lead employees to perceive a higher level of aches and pains.

Other organizational considerations that act as environmental stressors with social elements are work schedule and overtime. Shift work has been shown to have negative social, mental, and physical health consequences (Tasto et al., 1978; Monk and Tepas, 1985). In particular, night and rotating shift regimens affect worker sleeping and eating patterns, family and social life satisfaction, and injury incidence (Rutenfranz et al., 1977; Smith et al., 1982). Caplan et al. (1975) found that unwanted overtime was a far greater problem than simply the amount of overtime. Overtime may also have an indirect effect on worker stress and health because it reduces the amount of rest and recovery time, takes time away from relaxation with family and friends, and reduces time with these sources of social support that can buffer stress.

Technology is an environmental influence that can produce stress: for example, physical and mental requirements that do not match employee competencies, poorly designed software, and poor system performance, such as crashes and breakdowns (Turner and Karasek,
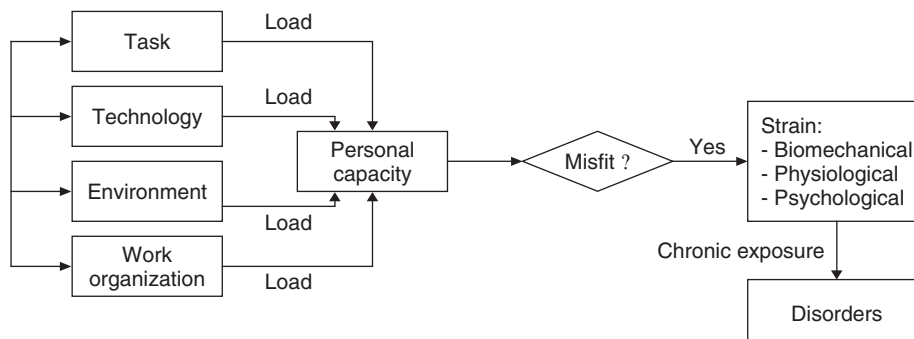


**Figure 1**   Balance model (Smith and Carayon-Sainfort, 1989) and stressor effects (Smith et al., 1999).

1984; Carayon-Sainfort, 1992). There is some evidence showing that computers can be a source of physical and mental stress [Smith et al., 1981, 1992; National Academy of Sciences (NAS), 1983]. New technology may exacerbate worker fears of job loss due to increased efficiency of technology (Ostberg and Nilsson, 1985; Smith 1987b). The way the technology is introduced may also influence worker stress and health (Smith and Carayon, 1995). For instance, when workers are not given enough time to get accustomed to the technology, they may develop unhealthy work methods.

## 11 ERGONOMIC WORK ANALYSIS AND ANTHROPOTECHNOLOGY

The development of ergonomics in France and French-speaking Belgium has been characterized by a distinctive effort to create a unified approach to work analysis. Those endeavors have produced a practice-oriented method aiming at improving work conditions through a multistep analysis and intervention process. The method places a strong emphasis on social aspects of the activities and on the operators' (i.e., workers) perspectives and experiences within the situation. The ergonomic work analysis (EWA) approach resulted chiefly from efforts of researchers at the Conservatoire National des Arts et Métiers (CNAM) with a significant role played by Alain Wisner and several others (e.g., De Keyser, De Montmollin, Daniellou, Faverge, Falzon, Laville, Leplat, and Ombredane). These researchers were convinced of the importance of having a common inquiry structure (i.e., a single ergonomic method) that would allow for a truly *interdisciplinary* approach to workplace ergonomics, as opposed to a *multidisciplinary* one. This unified method would serve as the framework enabling the integration of the different disciplines that provide the foundation for ergonomics.

The EWA method underscores the differences between the *task*, meaning work as prescribed or specified by management, and the *activity,* meaning work as actually performed and experienced by workers. The method considers that to be a critical gap and assumes that *tasks* as designed and described by engineers and managers tend to oversimplify the work situation missing critical complexities that constitute the essential reality of work and the core of ergonomic issues (Daniellou, 2005, De Keyser, 1991). While the EWA attempts to evaluate work situations from multiple viewpoints, it clearly emphasizes the workers' perspectives and the description of their actions. Significant attention is paid to the workers' efforts to meet production and procedural goals placed on them. It is posited that the approach places the operator at the center of the work situation and, therefore, of its (re)design (Laville, 2007).

As outlined by De Keyser (1991), the EWA procedure begins with an initial investigation of the request (i.e., the demand for the investigation), where the expectations and perspectives of stakeholders (e.g., corporate management, area supervisors, and employees) regarding the project are identified and considered. In this stage, as the demands and expectations for the study are spelled out, the analyst (ergonomist) would have

a clearer idea of what a successful intervention would need to accomplish (Wisner, 1995a). This step is followed by an analytical description of the physical, technical, economic, and social aspects of the situation and a detailed analysis of the work activities. A diagnostic is then established and recommendations for improvement defined. Depending on the nature of the activity, recommendations are expected to be pilot tested prior to full implementation. Follow-up on the performance of implemented solutions is advocated. Proponents of the EWA recognize that its application can be time consuming and unwieldy and concede that in most cases the process should be abbreviated (De Keyser, 1991).

According to Wisner (1995b, p. 1549), a main objective of the EWA is to learn how operators "constitute the problems of their work (situation and action) in a stable or variable way and, to a lesser extent, how they solve them." This process is termed "problem building" and is seen as critical in understanding work situations and in helping operators to improve their work conditions. Wisner (1995a) proposes that the operator's conceptualization of work problems (i.e., how they cognitively construct the issues) provide a better explanation for errors and accidents than the conditions under which the problem itself is solved. The author also posits that a vital aspect of the analysis includes what he calls "*self-confrontation*" or "*autoconfrontation*." The process involves a face-to-face discussion with workers focusing on potential discrepancies between their reported actions and what was observed by the analyst (Wisner, 1995a,b).

The EWA proponents argue for an inquiry that is primarily based on field research and call for the primacy of knowledge derived from natural/local settings over laboratory findings (De Keyser, 1992). Due to its emphasis on obtaining a rich or, in ethnographic parlance, "*thick*" description of the work situation from the operators' viewpoint, the EWA often makes use of qualitative, naturalistic, or ethnographic techniques. Wisner (1995b) remarks that the ergonomics discipline has clear connections with anthropology, and he defines "situated activity" as the focus of its analysis. The author admits, however, that when compared to actual anthropological studies the use of ethnographic methods in EWA is clearly restrained by the need to find an acceptable solution to a problem within a time frame.

As it evolved the nexus of the "Francophonic" ergonomics shifted toward the analysis of verbal communication among workers, with growing attention being paid to operators' cognition and the creation of meaning (De Keyser, 1991). Fittingly, Wisner (1995b) defines his viewpoint as one of an ergonomist or a cognitive psychologist who endeavors to understand cognitive phenomena but who also includes in his observation contextual aspects of the situation such as the physical environment, work organization, and the workers' prior experience, knowledge, and relationships. He sees cognitive anthropology as close to the principles of EWA as it attempts to understand the operator's own cognition as opposed to impose the observer's or management's one.

Wisner coined the term *anthropotechnology* to describe an approach that, similar to ergonomics, attempts

to adapt (or transfer) technology to a population by using knowledge from the social sciences in order to improve the design of productive systems (Gueslin, 2005). Different from most ergonomic applications, which typically address an individual or a small group of persons in a work situation, anthropotechnology focuses on collective, organizational, and even national human–technology fit issues.

It has been noted that efforts to transfer technologies to industrially developing countries (IDCs) have often been marred by health, economic, and environmental issues, including high prevalence of workplace injuries and illnesses, poor quality and productivity, and pollution (Moray, 2000; Shahnavaz, 2000; Wisner, 1985). Wisner's research compares the functioning of similar industrial facilities located in industrialized and developing countries and highlights geographical, historical, and cultural factors as key factors shaping the workers' activities and the overall system's performance. He sees the interface between ergonomics and anthropology as propitious to conduct in-depth analyses of the factors defining success or failure in technology transfers.

Anthropotechnology focuses on the differences between countries selling and buying technologies in terms of their social and industrial contexts as well as in regards to other anthropological domains such as physical anthropology (i.e., anthropometry), cultural anthropology (i.e., norms and values), and cognitive anthropology (i.e., role of prior knowledge on work situations including education and training) (Gueslin, 2005). It is an approach to ergonomics that attempts to link the reality of the individual human work to large-scale factors which affect the functioning of entire societies.

Antropotechnology emphasizes the importance of the (pluridisciplinary) work team in the successful technology transfer. It has a clear sociotechnical foundation, embracing the joint-optimization principle discussed elsewhere in this chapter. It shares some of its perspectives with macroergonomics, but it is clearly distinct in terms of its objectives, priorities, and methods.

## 12   COMMUNITY ERGONOMICS

Recently, human factors has looked beyond work systems to examine more complex environments where multiple systems interact with each other. Community ergonomics (CE) is an approach to applying human factors principles to the interaction of multiple systems in a community setting for the improvement of its quality of life (Smith et al., 1994; Cohen and Smith, 1994; Smith et al., 1996; Smith et al., 2002). CE evolved from two parallel directions, one being theories and principles in human factors and ergonomics, the other the evaluation of specific improvements in communities that led to theories and principles at the societal level of analysis. The CE approach focuses on distressed community settings characterized by poverty, social isolation, dependency, and low levels of self-regulation (and control). Deteriorating inner city areas in the United States are examples of such communities, as are underdeveloped countries and countries devastated by war and poverty. The practice of CE seeks to identify and implement

interventions that provide the disadvantaged residents of a community with the resources within their social environment that resolve their problems systematically in a holistic way. In macroergonomic terms this means achieving a good fit among people, the community, and the social environment. McCormick (1970) explained the difficulty in transforming the human factors discipline to deal with communities. He stated that the human factors aspects of this environment required a significant jump from the conventional human factors context of airplanes, industrial machines, and automobiles. In the first place, the systems of concern (the community) are amorphous and less well defined than are pieces of hardware. Lodge and Glass (1982) recognized the need for a systems approach to dealing with issues at a societal level. For improvement to occur, they advocated a cooperative, holistic approach with multiple reinforcing links from several directions. They pointed out correctly that providing jobs is an ineffective way to introduce change if training, day care, and other support systems are unavailable to the disadvantaged residents who are employed.

Several concepts of CE are extrapolated from behavioral cybernetic principles proposed by Smith (1966) regarding how a person interacts with her or his environment. CE proposes that community residents be able to track competently their environment and other community members within it. In addition, it is important that people be able to exert control over their lives within the environment and in social interactions. Residents need to be taught how to develop an awareness of the impact of their own actions on the environment and on others in their community. This understanding and control over their interactions with the environment and people enables residents to build self-regulating mechanisms for learning, social tracking, and feedback control of their lives. An effective self-regulating process is one that helps community residents identify situations of misfit between community residents and their environment, as well as allowing for the generation and implementation of solutions to improve fit and to deal with emerging challenges in a continuously changing and turbulent environment.

To put the situation in distressed communities in context, it can be likened to the cumulative trauma injuries and stress observed at the workplace. Many residents in a distressed inner city suffer from what can be termed *cumulative social trauma* (CST). Like work-related musculoskeletal disorders, CST results from long-term chronic exposure to detrimental circumstances in this case, societal conditions that create a cycle of dependency, social isolation, and learned helplessness. CST results from repeated exposure to poorly designed environments and/or long-term social isolation that leads to ineffective individual performance or coping abilities. Harrington (1962) observed a personality of poverty, a type of human being produced by the grinding, wearing life of the slums. Cumulative social trauma is the repetition of an activity or combinations of activities, environmental interactions, and daily life routines that develop gradually over a period of time and produce wear and tear on motivation, skill, and

emotions, leading to social and behavioral disruption, psychosomatic disorders, and mental distress.

The obstacles encountered on the path to progress in the declining areas of inner cities can be defined in terms of human errors by community residents and by the public institution employees who serve them that lead to a lack of system effectiveness and reliability. Human errors are decisions, behaviors, and actions that do not result in desired and expected outcomes: for example, failing to keep a job, relying on public assistance, failing to pay bills, being involved in illegal activities for the residents, and a lack of understanding, compassion, and effective services for the public employees. Social (and economic) system effectiveness and reliability are measured by the ability of public (and private) institutions to achieve positive community performance outcomes (Smith and Smith, 1994).

Social and economic achievements are enhanced by the level of the goodness of the fit between the characteristics and needs of community residents and those of the community environment. In the case of substantial misfit between people and the environment, poor urban residents are likely to make repeated "errors" in life which will result in poor economic and societal outcomes. Institutions with low system effectiveness and reliability do not provide adequate feedback (performance information or direction) and/or services to enable residents to correct their errors. One of the fundamental purposes of community ergonomics is to improve the goodness of fit between environmental conditions and residents' behaviors to reduce residents' errors. To be effective, interventions have to deal with the total system, including the multifaceted elements of the environment and community residents. The theory and practice of community ergonomics is based on the assumptions that individuals or community groups must attain and maintain some level of self-regulation and individuals or groups need to have control over their lives and their environments in order to succeed. This can be achieved by having residents participate actively in their own self-improvement and in improvement of their environment.

The resident–community–environment system has multilateral and continuous interactions among residents, groups, living conditions, public institutions, stores, and workplaces. These are linked through interactions and feedback from the interactions. Other communities and external public and private institutions that may have very different beliefs, values, and modes of behavior surround the community. Similarly, the community is surrounded by communication systems, architecture, transportation, energy systems, and other technology that influence and act on the community. These affect the life quality of individuals and groups in the community. This resident–community–environment system includes institutions for education, financial transactions, government and politics, commerce and business, law enforcement, transportation, and housing. The organizational complexity of this system affects the ways by which individuals and groups try to control the environment through their behavior.

According to Smith and Kao (1971), a *social habit* is defined as self-governed learning in the context of the control of the social environment. The self-governance of learning becomes patterned through sustained and persistent performance and reward, and these are critically dependent on time schedules. As these timed patterns become habituated, the individual can predict and anticipate social events, a critical characteristic in the management of social environments. Self-control and self-guidance in social situations further enhance the ability of the individual to adjust to various environments (old and new) by following or tracking the activities of other persons or groups. Social tracking patterns during habit cycles determine what is significant behavior for success or failure. Residents, groups, and communities develop and maintain their identity through the establishment and organization of social habits (i.e., accepted behaviors). In addition, the maintenance of group patterns and adherence to this process are achieved through social yoking (or mutual tracking), so that people can sense each other's social patterns and respond appropriately.

A resident–community–environment management process should build from the social habits of the community, coupled with a human-centered concept of community design. This seeks to achieve better community fit by making public and private institutions more responsive to resident capabilities, needs, and desires and residents more responsive to community norms. Smith and Kao (1971) indicated that cultural design could aid and promote development of individuals and communities if it is compliant with people's built-in or learned behavior. It can adversely affect behavior, learning, and development of individuals and communities if it is noncompliant with their needs and built-in makeup. Management of the integration of residents, the community, and the environment must be approached as a total system enterprise. The aim is to build proper compliances among the residents, the community, and the environment using social habits and individual control as key elements in compliance. This leads to the design of a resident–community–environment system that improves residents' perceptions, feedback, level of control, adherence to social habits, and performance through improved services and opportunities provided by public and private institutions.

The management process seeks to establish positive social tracking between the residents and public and private institutions. Residents can be viewed as being nested within the community. The guidance of a resident's behavior is determined by her or his ability to develop reciprocal control over the economic, social, and cultural institutions in the community using feedback from interactions with these institutions. The feedback concept is a significant aspect because it shows the resident the effectiveness of self-generated, self-controlled activity when interacting with the community and the environment. The quality of the feedback determines the course, rate, and degree of individual learning and behavior improvement in relation to interaction with the community and the environment.

The performance of the resident–community–environment system is dependent on critical timing considerations, such as work, school, and, public services

schedules. Successful participation in these activities can act as a developmental aspect of residents acquiring social habits for daily living. Some difficulties faced by the community are due to the misfit of critical timing factors between residents and institutions. There are many reasons for disruption: for example, residents not being motivated to adhere to a fixed schedule or the poor scheduling of the public services. An important element of a CE management process is to develop a tracking system to aid residents and public agencies in the sensing of critical aspects of timing and in the development of a "memory" for determining future events. The tracking system is aimed at building good habits of timing for the residents and the agencies. The absence of good timing habits results in delays in functioning, social disruption, organizational adjustment problems, emotional difficulty, and poor performance.

Smith et al. (2002) described seven principles that constituted the philosophy of the CE approach:

1. *Action Orientation.* Rather than trying to change community residents in order to cure them of unproductive behavior, CE believes in getting them actively involved in changing the environmental factors that lead to misfit. The CE approach strives to reach collective aims and perspectives on issues of concern and to meet specific goals and aspirations through specific actions developed by all involved parties. This process requires an organized and structured evaluation process and the formulation of plans for solutions and their implementation. The approach is based on new purposes, goals, and aspirations developed through community–environment reciprocal exchange rather than on existing community skills, needs assessments, or external resources. Other approaches have tended to become bogged down trying to fulfill institutional rules and requirements or institutional directives that pursue a good solution but to the wrong problems (Nadler and Hibino, 1994).

2. *Participation by Everyone.* Community improvements often fail because residents are not substantially involved in the process of selecting the aims, objectives, and goals. It is essential to have resident participation from start to finish. Such participation is a source of ideas, a means of motivation for the residents, and a way to educate residents to new ideas and modes of behavior. There are many mechanisms for participation, including individual involvement, action groups, and committees. Early involvement of strategic persons and institutions in the process brings the necessary concepts, technical expertise, and capital into the process and the solutions. Although participation by every resident of the community may not be possible at first, the goal is to get everyone involved at some point in some way. It is expected that reluctant and passive involvement will be minimized and will decrease continually throughout

a project or activity. Effective information transfer among individuals, organizations, and institutions is essential for success.

3. *Diversity and Conflict Management.* Communities are made up residents who may have differing perspectives, values, cultures, habits, and interests. Problems in a community are never neat and compartmentalized, and because the community system is complex, it is difficult to comprehend all of the perspectives. However, it is necessary to recognize the diversity in perspectives and opinions and find ways to work toward consensus. It is essential to formulate a process for managing diversity, conflict, and confusion that will occur in a community with many cultures and perspectives. Distressed communities typically have a low level of self-regulatory capability and high levels of diversity and dissention. It is important to spend the necessary time designing a process for handling diversity and conflict. Nadler and Hibino (1994) and Cohen (1997) have developed methodologies for working with diversity and conflict in designing solutions to problems.

4. *Encouraging Learning.* A well-designed process will allow residents and institutions to interact positively and effectively even in the conditions of a highly turbulent environment. It is expected that community residents and planners will learn from each other and from participating in the process. In addition, these learning effects will be transferred to subsequent related endeavors, with or without the presence of a formal community ergonomics process to facilitate the interaction. Thus, there is a transfer of "technology" (control, knowledge, skills) to the community in the form of the process and the learning experiences. Furthermore, participants will enhance their abilities in leadership, management, group activities, evaluation, and design learned while being involved in the CE process. Formal documentation of the system management process within the group setting as it occurs provides for better understanding and management of the community ergonomics process and provides historical documentation for future endeavors of a similar nature.

5. *Building Self-Regulation.* One aspiration of the community ergonomics process is to provide participants with an increased ability and capacity for self-regulation. *Self-regulation*, defined as the ability of a person or group to exert influence over the environmental context, is enhanced by creating specific tasks, actions, and learning opportunities that lead directly to the successful development of skill. When community members participate in a project that achieves specific goals as part of its evaluation, design, and implementation processes, they develop new abilities and skills to self-regulate themselves. This serves as motivation toward more community improvement activity.

6. *Feedback Triad.* Feedback is a critical aspect of the CE improvement process. Different levels of feedback provide opportunities for learning. Smith (1966) and Smith and Kao (1971) defined three levels of feedback for individual performances: reactive, instrumental, and operational feedback. *Reactive feedback* is an understanding of the response of the muscles in taking action, *instrumental feedback* is the feel of the tool being used to take the action, and *operational feedback* is the resulting change in the environment when action of the tool occurs. Smith (1966) stated that these are integrated into a feedback triad allowing for closed-loop control of the activity, social tracking, and self-regulation. Feedback that provides these levels of information at the community level is important in designing and implementing community environment improvements. Thus, mutatis mutandis, feedback on resident perspectives and actions provides reactive feedback, feedback on institutional perspectives and actions provides instrumental feedback, and feedback on the success of community improvements provides operational feedback. Reactive feedback is the personal sense that one's actions (or the group's actions) result in a perceived outcome on the environment. Instrumental feedback is sensed from subsequent movement of an institution or group in the form of milestones achieved and output produced. Operational feedback comes from the results of such activities as planning, designing, installing, and managing group intentions as well as the new policies, laws, buildings, and institutions resulting from such activities. These are examples of persisting results that can be sensed directly by community environment social tracking systems. Without the feedback triad, self-regulation by group participants would not be very effective and the system could quickly degenerate. Participants must sense that their personal actions, words, and participation have effects on themselves, on others, and on the environment.

7. *Continuous Improvement and Innovation.* The community ergonomics approach recognizes the need for continuous improvement, which can be achieved by continuous planning and monitoring of results of projects implemented. Private organizations can be encouraged to provide guidance and feedback on the purpose, goals, and management of improvement initiatives. Inputs from private organizations and governmental programs can be utilized to promote an entrepreneurial spirit that encourages effective community habits. These can be benchmarked against other communities and other programs. Valuable information can be elicited by studying the effects of a solution over a period of many years to prevent the redevelopment of a dysfunctional system and to give members of troubled communities opportunities for better lives.

This implies the need for ongoing monitoring to evaluate the operational requirements for implementation, measuring effectiveness, and use of feedback to alter programs already in existence. Consistent monitoring of citizen needs, desires, and values must be established to verify that programs and products are accessible, usable, useful, and helpful to community residents.

Community ergonomics is a long-overdue answer to the application of human factors engineering principles to address complex societal problems. Community ergonomics is a philosophy, a theory, a practice, a solution-finding approach, and a process, all in one. CE is a way to improve complex societal systems that are showing signs of CST (cumulative social trauma). CST is not to be taken lightly, as the costs are immense in every respect: financial, human, social, and developmental. When any group of people, a community, or a region is isolated, alienated, and blocked from access to resources needed to prosper, the consequences are long lasting and deeply detrimental.

Recently, concepts of CE have been developed specifically for corporations engaged in international development and trade (Smith et al., 2009). During the last half of the twentieth century, a struggle began for fairness, equality, freedom, and justice for people in many developing nations. This has occurred during a time of expansion of the global economy. Companies now operate in a complex world economy characterized by continuous change; a heterogeneous (often international) workforce at all job levels; increased spatial dispersion of their financial, physical, and human assets; increased diversification of products and markets; uneven distribution of resources; variable performance within and between locales; increased operational and safety standards; and differences in economic, social, political, and legal conditions. In this climate of increased international trade, important issues of social and cultural values need to be examined carefully. There is need for an understanding of how specific cultures and cultural values can affect corporate operations.

Companies have developed management practices in response to numerous obstacles encountered when they expand abroad or when they transfer processes abroad. Difficulties that they have encountered in their growth, expansion, internationalization, and globalization are due primarily to the following factors: (1) the lack of a process for effective transnational transfers; (2) the lack of knowledge of operational requirements and specifications in newly entered markets; (3) ignorance of cultural norms and values in different countries; (4) the lack of adaptability mechanisms; (5) a low tolerance for uncertainty, ambiguity, and change and diversity; and (6) a lack of acceptance by segments in the populations in which they are starting operations. One of the most important issues in acceptance by the local people is that a company's commitment to social responsibility has the same intensity as those striving for large profits and the use of cheap local natural resources and labor. In fact, some managers believe that social responsibility may conflict with corporate financial goals and tactics.

However, CE postulates that long-term corporate stability and success (profits) will occur only if there is local community support for the enterprise and its products. Social responsibility nurtures the mutual benefits for the enterprise and the community that lead to stability and success. CE proposes that multinational corporations must accept a corporate social responsibility that recognizes the universal rights of respect and fairness for all employees, neighbors, purchasers, and communities. Whether these rights are profitable or not in the short term or difficult to attain, such corporate social responsibility is a requirement if global ventures are to be successful in the long term. The survivability, acceptability, and long-term success of a corporation will depend not only on quick profits but also on social responsibility that builds long-term community acceptance and support of the enterprise. This will lead to greater accessibility to worldwide markets. Organizations must address multicultural design, a comfortable corporate culture, and principles of respect and fairness for employees, customers, and neighbors. This will improve the fit with the international cultures in which the corporation operates.

CE has developed principles for successful multinational organizational design (Smith et al. 2009). These principles focus on a goal of social responsibility, fairness, and social justice and do not threaten the prosperity of a company or organization. They are built on the premise that the society and communities in which a company operates should benefit from the presence of the company in the society. Thus, the corporation should contribute to the development, growth, and progress of local communities. These principles propose a reciprocal relationship between the hosting community or society and the outside corporation.

1. *Fit Principle*. Accommodations need to be made by companies for a diverse workforce. That is, companies need to design for cultural diversity. Corporations must understand and incorporate the norms, customs, beliefs, traditions, and modes of behavior of the local community into their everyday operations. In some communities multiple cultures will need to be included in this process. Often, there is a need to strike a balance among the various cultures, as they are not always compatible. The corporation's organizational structure and operational style need to be flexible to bridge the gap between the corporate and local cultures.

2. *Balance Principle*. The balance principle defines the need to find the proper relationship among components of a larger system. Based on this concept we believe that there is a need for balancing corporate financial goals and objectives with societal rights and corporate social responsibility. Companies are a part of a larger community, and through their employment, purchasing, civic, and charity activities they can influence community development and prosperity. As an integral part of this larger system, companies have a responsibility to promote positive balance for the benefit of the community and the corporation.

3. *Sharing Principle*. Traditionally, a corporation's success has been measured in terms of its financial growth, but there are other factors that will become more critical as social awareness becomes more prominent. For instance, customer loyalty, community support, and acceptability of products will be critically related to the corporation's long-term financial success. If a corporation chooses to invest some of its profits back into a community in ways that are significant to that community, the corporation may be viewed not only as a business but also as a community partner. In giving something back to the community the corporation is developing loyalty to its products and protecting its long-term profitability.

4. *Reciprocity Principle*. The reciprocity principle deals with the mutual commitment, loyalty, respect, and gain between producers and consumers. A bond results from the corporation giving something back to the community, which builds loyalty from the consumers to the company, and eventually leads to a genuine sense of loyalty from the organization back to the community. In this respect, what might have started as responsibility will over time become mutual loyalty and commitment. Within the corporate organization, the same phenomenon takes place when the organization shows responsibility toward its employees (producers), who in turn become loyal and committed partners with the corporation.

5. *Self-Regulation Principle*. Corporations should be viewed as catalysts of self-regulation and socioeconomic development in host communities. Communities and countries in disadvantaged economic conditions typically show symptoms of learned helplessness, dependency, isolation, and cumulative social trauma. Instead of perpetuating conditions that weaken people and institutions, an effort should be made to help people to self-regulate, grow, flourish, and become productive. In this effort, corporations are very important because they provide employment, training, and professional development opportunities that give people the tools to help themselves. Corporations can also invest in the community infrastructure, such as schools, clinics, and hospitals, which leads to stronger, healthier, and more independent communities in the future.

6. *Social Tracking Principle*. Awareness of the environment, institutional processes, and social interaction are necessary for people and corporations to navigate through their daily lives and for communities to fit into the broader world. Clear awareness helps to control the external world and leads to more robust, flexible, open system design. It is important for community members, employees, and corporations to be aware of their surroundings to be able to predict potential outcomes of actions taken. Similarly,

it is important for corporations to develop a certain level of awareness regarding the workforce, the community, and the social, economic, and political environment within which they operate. This includes the cultural values of the people affected by a corporation's presence in a particular community.

7. *Human Rights Principle*. The human rights principle underscores the belief that every person has the right to a reasonable quality of life, fair treatment, a safe environment, cultural identity, respect, and dignity. There is no reason for anyone not to be able to breathe fresh air, preserve their natural resources, achieve a comfortable standard of living, feel safe and dignified while working, and be productive. People should not be assigned a difference in worth based on class, gender, race, nationality, or age. The workplace is a good starting point to bring about fairness and justice in societies where these do not exist as a norm.

8. *Partnership Principle*. This principle proposes a partnership among the key players in a system in order to achieve the best possible solution: corporation, community, government, employees, and international links. By doing this, balance may be achieved between the interests of all parties involved and everyone is treated fairly. In addition, partnership assures commitment to common objectives and goals.

The essence of internationalization, globalization, and multiculturalism is in the culture and social climate that the corporation develops to be sensitive to the community and the diversity of the workforce. This includes respect, partnership, reciprocity, and social corporate responsibility toward employees, the community, and society as a whole. It requires seeking a balance between the corporate culture and that of the community where the business operates and the cultures brought into the company by the diverse employees. In the past, corporations have entered new markets all over the world, profiting from cheap labor and operating freely with little or no safety or environmental liability. However, the level of social awareness has increased all over the world, exposing sweat shops, inhumane working conditions, labor exploitation, and environmental violations across the board. The focus in the future will be on doing business with a social conscience. By doing this, corporations will become welcome in any part of the world they wish to enter.

## 13  CONCLUDING REMARKS

As ergonomics has matured as a science, the emphasis has broadened from looking primarily at the individual worker (user, consumer) and her or his interaction with tools and technology to encompass larger systems. A natural progression has led to an examination of how the social environment and processes affect an individual and groups using technology as well as how the behavior and uses of technology affect society.

We have described aspects of these reciprocal effects by emphasizing select theories and perspectives where social considerations have made a contribution to system design and operation.

We see ergonomics as an essential aspect of the continuous human effort to survive and prosper. It is central to our collective endeavor to improve work systems. Ergonomics answers to the social needs of effective utilization of human talent and skills and of respect and support for their different abilities. By reducing exposure to physical hazards, particularly those associated with the onset of musculoskeletal disorders, by controlling occupational stress and fatigue, ergonomics is essential for the improvement of work conditions and the overall health of the population. It is instrumental to the safety and functionality of consumer products and key to the reliability of systems on which we depend. In summary, ergonomics is a critical aspect for the well-being of individuals, for the effectiveness of organizations, and for the prosperity of national economies.

Ergonomics can be seen as a technology and part of a broader social context. As such, it responds to the needs, conditions, and structure of that society. Ergonomics technology evolution is shaped by ever-changing societal motivations. The relationships between ergonomics and societal demands can be understood as reciprocal: where social needs determine the direction of ergonomics development and ergonomic innovations once introduced in the environment allow the fruition and reinforcement of some aspects or drives of the social process.

Social needs for increased productivity, higher quality, better working conditions, and reliability changed over time as some of these drives become more prominent. As work systems become more complex and the workforce more educated and sophisticated, the consideration by the ergonomics discipline of broader social, political, and financial aspects has been heightened. These changes have been answered by ergonomics in different but ultimately interrelated approaches.

Earlier we highlighted the seminal contribution of Kurt Lewin to an understanding of the social aspects of work with emphasis on the humane side of the organization. Lewin focused on reconciling scientific thinking and democratic values in the workplace and on recovering the meaning of work. He was an early advocate of worker involvement and saw job satisfaction as an essential goal to be met by work systems. Lewin's ideas and later the application of general systems theory had deep implications for the ergonomics discipline. Work by researchers at the Tavistock Institute, also inspired by Lewin, led to establishment of the sociotechnical systems theory, which confirmed the importance of the social context in work systems optimization.

While Kurt Lewin was one of the leaders in defining the critical need for work to provide psychological and social benefits to the employees, others who followed him carried through with these ideas by turning them into reality at the workplace. At the heart of most of these ideas was the concept of participation by employees in the design and control of their own work and workplaces. French, Kahn, Katz, McGregor, K. U.

Smith, Emory, Trist, Davis, Drucker, Deming, Juran, Lawler, M. J. Smith, Hendrick, Noro, Imada, Karasek, Wilson, Carayon, and Sainfort have all promoted use of employee participation as a process or vehicle to engage employees more fully in their work as a source of motivation, as a means for providing more employee control over the work process and decision making, or as a mechanism to enlist employee expertise in improving work processes and products. The nature of employee participation has varied from something as simple as asking for employee suggestions for product improvements to something as complex as semiautonomous work groups, where the employees make critical decisions about production and resource issues. The critical feature of participation is the active engagement of employees in providing input into decisions about how things are done at the workplace. The social benefits of participation are in the development of company, group, and team cohesiveness and a cooperative spirit in fulfilling company and individual goals and needs. Individuals learn that they can contribute to something bigger than themselves and their jobs, learn how to interact with other employees and managers positively, and receive social recognition for their contributions. This "socialization" process leads to higher ego fulfillment, greater motivation, higher job satisfaction, higher performance, less absenteeism, and fewer labor grievances. Participatory ergonomics has become an essential aspect of work improvement.

The social side of work is more than just the positive motivation and ego enhancement that can occur with good workplace design; it is also the stress that can develop when there is poor workplace design. Work design theorists and practitioners have put substantial emphasis on the psychosocial aspects of work and how these aspects influence employee productivity and health (Smith, 1987a; Kalimo et al., 1997; Carayon and Smith, 2000; Kivimaki and Lindstrom, 2006). This tradition grew out of work democracy approaches in Scandinavia and Germany (Levi, 1972; Gardell, 1982) that emphasized the role of employee participation and codetermination in providing "satisfying" work. The absence of satisfying work and/or combinations of chronic exposure to high physical and psychological work demands were shown to lead to ill health and reduced motivation at work. Strong ties between poor working conditions, employee dissatisfaction, poor quality of working life, and negative outcomes for motivation, production, and health were documented Smith, 1987). Thus, social and psychological aspects of working conditions were shown to influence not only personal and group satisfaction, ego, motivation, and social and productive outcomes but also the health, safety, and welfare of the workforce. This tradition defined critical social and organizational design features that could be improved through organizational and job design strategies. Many of the strategies included considerations of employee involvement in the work process, social mechanisms such as laws and rules defining workplace democracy and codetermination, and health care approaches that encompass psychosocial aspects of work as a consideration.

Macroergonomics grew out of several traditions in organizational design, systems theory, employee participation, and psychosocial considerations in work design. Championed by Hendrick (1984, 1986, 2002) and Brown (1986, 2002), this tradition expanded the focus of work design from the individual employee and work group to a higher systems level that examined the interrelationship among various subsystems of an enterprise. By definition there are structures (organizational, operational, social) that define the nature of the interaction among the subsystems, and macroergonomics aims for the joint optimization of these subsystems. Hendrick (2002) describes how macroergonomics was a response to bad organizational design and management practices that minimized the importance and contributions of the human component of the work system. This is in line with and an elaboration of the prior traditions described above but with a primary focus on the systems nature of the work process and integration of the subsystems. At the heart of macroergonomics is the use of employee participation to achieve system integration and balance (Carayon and Smith, 2000; Brown, 2002).

Community ergonomics is a natural extension of macroergonomics to a higher level above the enterprise, in this case to the community. Like many of the preceding theories and concepts, an important aspect of community ergonomics is improvement in the quality of life for the people in the system. Such improvement concerns economic benefits, but there is also a strong emphasis on developing the social and psychosocial aspects of individual and community life that enhance overall well-being. Like macroergonomics, community ergonomics examines the subsystems in an enterprise and how to optimize them jointly to achieve benefits for the people. At a higher level, community ergonomics has provided advice to multinational enterprises in how to provide reciprocal benefits to the enterprise and the several communities in which the enterprise operates. Among the considerations is the critically important concept of the enterprise's sensitivity to, and accommodation of, social and cultural differences (Derjani-Bayeh and Smith, 2000; Smith et al., 2002).

## REFERENCES

Argyris, C. (1964), *Integrating the Individual and the Organization*, Wiley, New York.

Asberg, M., Nygren, A., Leopardi, R., Rylander, G., Peterson, U., Wilczek, L., Kallmen, H., Ekstedt, M., Akerstedt, T., Lekander, M., and Ekman, R. (2009), "Novel Biochemical Markers of Psychosocial Stress in Women," *PLos ONE*, Vol. 4, No. 1, e3590. doi:10.1371/journal.pone.0003590.

Axelsson, J. R. C. (2000), "Quality and Ergonomics: Towards a Successful Integration," Dissertation 616, Linköping Studies in Science and Technology, Linköping University, Linköping, Sweden.

Axelsson, J. R. C., Bergman, B., and Eklund, J., Eds. (1999), in *Proceedings of the International Conference on TQM and Human Factors*, Linköping, Sweden.

Babbage, C. (1832), *On the Economy of Machinery and Manufactures*, 4th ed., A. M. Kelley, New York; reprinted 1980.

Badham, R. J. (2001), "Human Factor, Politics and Power," in *International Encyclopedia of Ergonomics and Human Factors*, W. Karwowski, Ed., Taylor & Francis, New York, pp. 94–96.

Bertalanffy, L. von (1950), "The Theory of Open Systems in Physics and Biology," *Science*, Vol. 111, pp. 23–29.

Bertalanffy, L. von (1968), *General Systems Theory: Foundations, Development, Applications*, George Brazilier, New York; rev. ed. 1976.

Broedling, L. A. (1977), "Industrial Democracy and the Future Management of the United States Armed Forces," *Air University Review*, Vol. 28, No. 6, available: http://www.airpower.maxwell.af. mil/airchronicles/aureview/1977/sep-oct/broedling.html, accessed August 15, 2004.

Brown, O., Jr. (1986), "Participatory Ergonomics: Historical Perspectives, Trends, and Effectiveness of QWL Programs," in *Human Factors in Organizational Design and Management*, Vol. 2, O. Brown, Jr., and H. W. Hendrick, Eds., North-Holland, Amsterdam, pp. 433–437.

Brown, O., Jr. (1990), "Macroergonomics: A Review," in *Human Factors in Organizational Design and Management*, Vol. 3, K. Noro and O. Brown, Eds., North-Holland, Amsterdam, pp. 15–20.

Brown, O., Jr. (1991), "The Evolution and Development of Macroergonomics," in *Designing for Everyone: Proceeding of the 11th Congress of the International Ergonomics Association*, Y. Queinnec and F. Daniellou, Eds., Taylor & Francis, Paris, pp. 1175–1177.

Brown, O., Jr. (2002), "Macroergonomic Methods: Participation," in *Macroergonomics: Theory, Methods, and Applications*, H. W. Hendrick and B. M. Kleiner, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 25–44.

Caplan, R. D., Cobb, S., French, J. R., Jr., van Harrison, R., and Pinneau, S. R. (1975), *Job Demands and Worker Health*, U.S. Government Printing Office, Washington, DC.

Carayon, P., and Smith, M. J. (2000), "Work Organization and Ergonomics," *Applied Ergonomics*, Vol. 31, pp. 649–662.

Carayon, P., Smith, M.J. and Haims, M.C. (1999). Work organization, job stress, and work-related musculoskeletal disorders. Human factors, 41, 644–663.

Carayon-Sainfort, P. (1992), "The Use of Computer in Offices: Impact on Task Characteristics and Worker Stress," *International Journal of Human–Computer Interaction*, Vol. 4, No. 3, pp. 245–261.

Carlopio, J. (1986), "Macroergonomics: A New Approach to the Implementation of Advanced Technology," in *Human Factors in Organizational Design and Management*, Vol. 2, O. Brown and H. Hendrick, Eds., North-Holland, Amsterdam, pp. 581–591.

Chapanis, A. (1999), *The Chapanis Chronicles: 50 Years of Human Factors Research, Education, and Design*, Aegean, Santa Barbara, CA.

Chengalur, S. N., Rodgers, S. H., and Bernard, T. E. (2004), *Kodak's Ergonomic Design for People at Work*, 2nd ed., Wiley, New York.

Christensen, J. M. (1987), "The Human Factors Function," in *Handbook of Human Factors*, G. Salvendy, Ed., Wiley, New York, pp. 1–16.

Cobb, S., and Kasl, S. (1977), *Termination: The Consequences of Job Loss*, U.S. Government Printing Office, Washington, DC.

Coch, L., and French, J. R. (1948), "Overcoming Resistance to Change," *Human Relations*, Vol. 1, pp. 512–532.

Cohen, A. L., Gjessing, C. C., Fine, L. J., Bernard, B. P., and McGlothlin, J. D. (1997), *Elements of Ergonomics Programs: A Primer Based on Workplace Evaluations of Musculoskeletal Disorders*, National Institute for Occupational Safety and Health, Cincinnati, OH.

Cohen, W. J. (1997), *Community Ergonomics: Design Practice and Operational Requirements*, University of Wisconsin Library, Madison, WI.

Cohen, W. J., and Smith, J. H. (1994), "Community Ergonomics: Past Attempts and Future Prospects Toward America's Urban Crisis," in *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 734–738.

Cooper, C. L., and Marshall, J. (1976), "Occupational Sources of Stress: A Review of the Literature Relating to Coronary Heart Disease and Mental Ill Health," *Journal of Occupational Psychology*, Vol. 49, pp. 11–28.

Cooper, C. L., and Payne, R., Eds. (1988), *Causes, Coping and Consequences of Stress at Work*, Wiley, Chichester, West Sussey, England.

Cotton, J. L., Vollrath, D. A., Frogatt, K. L., Lengnick-Hall, M. L., and Jennings, K. R. (1988), "Employee Participation: Diverse Forms and Different Outcomes," *Academy of Management Review*, Vol. 13, pp. 8–22.

Cox, T., and Ferguson, E. (1994), "Measurement of the Subjective Work Environment," *Work and Stress*, Vol. 8, pp. 98–109.

Daniellou, F. (2001), "Epistemological Issues About Ergonomics and Human Factors," in *International Encyclopedia of Ergonomics and Human Factors*, W. Karwowski, Ed., Taylor & Francis, New York, pp. 43–46.

Daniellou, F. (2005), "The French-Speaking Ergonomists' Approach to Work Activity: Cross-Influences of Field Intervention and Conceptual Models," *Theoretical Issues in Ergonomics Science*, Vol. 6, No. 5, pp. 409–427.

Derjani-Bayeh, A. and Smith, M.J. (2000), "Application of Community Ergonomics Theory to International Corporations," in *Proceedings of the IEA/HFES 2000 Congress* Volume 2. Human Factors and Ergonomics Society, Santa Monica, CA, pp. 788–791.

De Keyser, V. (1991), "Work Analysis in French Language Ergonomics: Origins and Current Research Trends," *Ergonomics*, Vol. 34, No. 6, pp. 653–669.

Dean, J. W., and Bowen, D. E. (1994), "Management Theory and Total Quality: Improving Research and Practice through Theory Development," *Academy of Management Review*, Vol. 19, pp. 392–418.

DeGreene, K. B. (1986), "Systems Theory, Macroergonomics, and the Design of Adaptive Organizations," in *Human Factors in Organizational Design and Management*, Vol. 2, O. Brown and H. Hendrick, Eds., North-Holland, Amsterdam, pp. 479–491.

De Keyser, V. (1992), "Why Field Studies," in *Design for Manufacturability: A Systems Approach to Concurrent Engineering*, M. Helander and M. Nagamachi, Eds., Taylor & Francis, London, pp. 305–316.

Dray, S. M. (1985), "Macroergonomics in Organizations: An Introduction," *Ergonomics International*, Vol. 85, pp. 520–522.

Drury, C. G. (1978), "Integrating Human Factors Models into Statistical Quality Control," *Human Factors*, Vol. 20, No. 5, pp. 561–572.

Drury, C. G. (1997), "Ergonomics and the Quality Movement," *Ergonomics*, Vol. 40, No. 3, pp. 249–264.

Drury, C. G. (1999), "Human Factors and TQM," in *The Occupational Ergonomics Handbook*, W. Karwowski and W. Marras, Eds., CRC Press, Boca Raton, FL, pp. 1411–1419.

Eklund, J. (1995), "Relationships between Ergonomics and Quality in Assembly Work," *Applied Ergonomics*, Vol. 26, No. 1, pp. 15–20.

Eklund, J. (1997), "Ergonomics, Quality, and Continuous Improvement: Conceptual and Empirical Relationships in an Industrial Context," *Ergonomics*, Vol. 40, No. 10, pp. 982–1001.

Eklund, J. (1999), "Ergonomics and Quality Management: Humans in Interaction with Technology, Work Environment, and Organization," *International Journal of Occupational Safety and Ergonomics*, Vol. 5, No. 2, pp. 143–160.

Eklund, J., and Berggren, C. (2001), "Ergonomics and Production Philosophies," in *International Encyclopedia of Ergonomics and Human Factors*, W. Karwowski, Ed., Taylor & Francis, New York, pp. 1227–1229.

Emery, F. E., and Trist, E. L. (1960), "Sociotechnical Systems," in *Management Sciences: Models and Techniques*, C. W. Churchman et al., Eds., Pergamon, London.

Emery, F. E., and Trist, E. L. (1965), "The Causal Texture of Organizational Environments," *Human Relations*, Vol. 18, No. 1, pp. 21–32.

Fayol, H. (1916), *General and Industrial Management*, revised by Irwin Gray, 1987, David S. Lake Publications, Belmont, CA.

Frankenhaeuser, M. (1986), "A Psychobiological Framework for Research on Human Stress and Coping," in M. H. Appley and R. Trumbull, Eds., *Dynamics of Stress—Physiological, Psychological and Social Perspectives*. Plenum, New York, pp. 101–116.

Frankenhaeuser, M., and Gardell, B. (1976), "Underload and Overload in Working Life: Outline of a Multidisciplinary Approach," *Journal of Human Stress*, Vol. 2, No. 3, pp. 35–46.

Frankenhaeuser, M., and Johansson, G. (1986), "Stress at Work: Psychobiological and Psychosocial Aspects," *International Review of Applied Psychology*, Vol. 35, pp. 287–299.

French, J. R. P. (1963), "The Social Environment and Mental Health," *Journal of Social Issues*, Vol. 19, pp. 39–56.

French, J. R. P., and Caplan, R. D. (1973), "Organizational Stress and Individual Strain," in *The Failure of Success*, A. J. Marrow, Ed., AMACOM, New York, pp. 30–66.

Ganster, D. C., and Schaubroeck, J. (1991), "Work Stress and Employee Health," *Journal of Management*, Vol. 17, No. 2, pp. 235–271.

Gardell, B. (1982). "Scandinavian Research on Stress in Working Life," *International Journal of Health Services*, Vol. 12, pp. 31–41.

Gilbreth, F., and Gilbreth, L. (1917), *Applied Motion Study*, Sturgis & Walton, New York; reprinted, Hive Publishing, Easton, PA, 1973.

Gilbreth, F., and Gilbreth, L. (1920), *Fatigue Study*, Sturgis & Walton, New York; rev. ed., Macmillan, New York; reprinted, Hive Publishing, Easton, PA, 1973.

Grudin, J. (1990), "The Computer Reaches Out: The Historical Continuity of Interface Design," in *Proceedings of CHI '90: Empowering People*, J. C. Chew and J. Whiteside, Eds., Association for Computing Machinery, New York, pp. 261–268.

Gueslin, P. (2005), "The Development of Anthropotechnology in the Social and Human Sciences: Its Applications on Fieldworks", in *Human Factors in Organizational Design and Management*, Vol. 8, P. Carayon, M. Robertson, B. Kleiner, and P. Hoonaker, Eds., IEA Press,CD-ROM.

Hackman, R. J., and Oldham, G. R. (1980), *Work Redesign*, Addison-Wesley, Reading, MA.

Hansson, A-S, Vingard, E., Arnetz, B. B., and Anderzen, I. (2008), "Organizational Change, Health, and Sick Leave among Health Care Employees: A Longitudinal Study Measuring Stress Markers, Individual, and Work Site Factors," *Work & Stress*, Vol. 22, No. 1, pp. 69–80.

Harrington, M. (1962), *The Other America: Poverty in the United States*, Macmillan, New York.

Hendrick, H. W. (1984), "Wagging the Tail with the Dog: Organizational Design Considerations in Ergonomics," in *Proceedings of the Human Factors Society 28th Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 899–903.

Hendrick, H. W. (1986), "Macroergonomics: A Concept Whose Time Has Come," in *Human Factors in Organizational Design and Management*, Vol. 2, O. Brown, Jr., and H. W. Hendrick, Eds., North-Holland, Amsterdam, pp. 467–478.

Hendrick, H. W. (1991), "Human Factors in Organizational Design and Management," *Ergonomics*, Vol. 34, pp. 743–756.

Hendrick, H. W. (1997), "Organizational Design and Macroergonomics," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York.

Hendrick, H. W. (2002), "An Overview of Macroergonomics," in *Macroergonomics: Theory, Methods, and Applications*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1–23.

Hendrick, H. W., and Kleiner, B. M. (2001), *Macroergonomics: An Introduction to Work System Design*, Human Factors and Ergonomics Society, Santa Monica, CA.

Hughes, T. P. (1991), "From Deterministic Dynamos to Seamless-Web Systems," in *Engineering as a Social Enterprise*, H. E. Sladovich, Ed., National Academy Press, Washington, DC, pp. 7–25.

Huzzard, T. (2003), "The Convergence of Quality of Working Life and Competitiveness: A Current Swedish Literature Review," National Institute for Working Life, available: http://nile.lub.lu.se/arbarch/aio/2003/aio2003_09.pdf, accessed August 10, 2004.

Imada, A. S. (1991), "The Rationale for Participatory Ergonomics," in *Participatory Ergonomics*, K. Noro and A. Imada, Eds., Taylor & Francis, London.

Imada, A. S., and Robertson, M. M. (1987), "Cultural Perspectives in Participatory Ergonomics," in *Proceedings of the Human Factors Society 31st Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 1018–1022.

International Ergonomics Association, IEA (2000). Available: http://www-iea.me.tut.fi/, accessed August 10, 2004.

Israel, B.A., Baker, E.A., Goldenhar, L.M. and Heaney, C.A. (1996). Occupational stress, safety, and health: Conceptual framework and principlesfor effective prevention interventions. Journal of occupational health psychology, 1 (3), 261–286.

Jackson, S. E. (1989), "Does Job Control Control Job Stress?" in *Job Control and Worker Health*, S. L. Sauter, J. J. Hurrel, and C. L. Cooper, Eds., Wiley, Chichester, West Sussex, England, pp. 25–53.

Jackson, S. E., and Schuler, R. S. (1985), "A Meta-Analysis and Conceptual Critique of Research on Role Ambiguity and Role Conflict in Work Settings," *Organizational Behavior and Human Decision Processes*, Vol. 36, pp. 16–78.

James, G. D., Yee, L. S., Harshfield, G. A., Blank, S. G., and Pickering, T. G. (1986), "The Influence of Happiness, Anger, and Anxiety on the Blood Pressure of Borderline Hypertensives," *Psychosomatic Medicine*, Vol. 48, No. 7, pp. 502–508.

Jirotka, M., and Goguen, J. (1994), *Requirements Engineering: Social and Technical Issues*, Academic, London.

Johnson, J.V. and Hall, E.M. (1988). Job strain, workplace social support and cardiovascular disease: A cross-sectional study of a random sample of the Swedish working population. American Journal of Public health, 78 (10), 1336–1342.

Johnson, J.V., Hall, E.M. and Theorell, T. (1989). Combined effects of job strain and social isolation on cardiovascular disease morbity and mortality in a random sample of the Swedish male working population. Scandinavian Journal of Work, Environment & Health, 15 (4), 271–279.

Johnson, J. V., and Johansson, G. (1991), "Work Organisation, Occupational Health, and Social Change: The Legacy of Bertil Gardell," in *The Psychosocial Work Environment: Work Organization, Democratization and Health*, J. V. Johnson and G. Johansson, Eds., Baywood, Amityville, NY.

Kalimo, R. (1990), "Stress in Work," *Scandinavian Journal of Work, Environment and Health*, Vol. 6, Suppl. 3.

Kalimo, R., Lindstrom, K., and Smith, M. J. (1997), "Psychosocial Approach in Occupational Health," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York, pp. 1059–1084.

Kanter, R. M. (1983), *The Change Masters*, Simon and Schuster, New York.

Karasek, R., and Theorell, T. (1990), *Healthy Work*, Basic Books, New York.

Karasek, R. A. (1979), "Job Demands, Job Decision Latitude, and Mental Strain: Implications for Job Redesign," *Administrative Science Quarterly*, Vol. 4, pp. 285–308.

Karasek, R. A., Theorell, T., Schwartz, J. E., Schnall, P. L., Pieper, C. F., and Michela, J. L. (1988), "Job Characteristics in Relation to the Prevalence of Myocardial Infarction in the U.S. Health Examination Survey (HES) and the Health and Nutrition Examination Survey (HANES)," *American Journal of Public Health*, Vol. 78, pp. 910–918.

Kawakami, N., Haratani, T., Kaneko, T., and Araki, S. (1989), "Perceived Job-Stress and Blood Pressure Increase among Japanese Blue Collar Workers: One-Year Follow-up Study," *Industrial Health*, Vol. 27, No. 2, pp. 71–81.

Kivimaki, M., and Lindstrom, K. (2006), "Psychosocial Approach to Occupational Health," in *Handbook of Human Factors and Ergonomics*, 3rd ed., G. Salvendy, Ed., Wiley, Hoboken, NJ, pp. 801–817.

Konz, S., and Johnson, S. (2004), *Work Design: Occupational Ergonomics*, Holcomb Hethaway, Scottsdale, AZ.

Landy, F. J. (1992), "Work Design and Stress," in *Work and Well-Being*, G. P. Keita and S. L. Sauter, Eds., American Psychological Association, Washington, DC, pp. 119–158.

Laville, A. (2007), "Referências para uma história da ergonomia francófona," in *Ergonomia*, P. Falzon, Ed., Editora Blucher, São Paulo, pp. 21–32.

Lawler, E. E., III (1986), *High-Involvement Management*, Jossey-Bass, San Francisco.

Lawler, E. E., III., Morhman, S. A., and Ledford, G. E., Jr. (1992), *Employee Participation and Total Quality Management*, Jossey-Bass, San Francisco.

Lazarus, R. S. (1974), "Psychological Stress and Coping in Adaptation and Illness," *International Journal of Psychiatry in Medicine*, Vol. 5, pp. 321–333.

Lazarus, R. S. (1977), "Cognitive and Coping Processes in Emotion," in *Stress and Coping: An Anthology*, A. Monat and R. S. Lazarus, Eds., Columbia University Press, New York, pp. 145–158.

Lazarus, R. S. (1993), "From Psychological Stress to the Emotions: A History of Changing Outlooks," *Annual Reviews in Psychology*, Vol. 44, pp. 1–21.

Lazarus, R. S. (1998), *Fifty Years of Research and Theory by R. S. Lazarus: An Analysis of Historical and Perennial Issues*, Lawrence Erlbaum Associates, Mahwah, NJ.

Lazarus, R. S. (1999), *Stress and Emotion: A New Synthesis*, Springer, New York; reprinted 2006.

Lazarus, R. S. (2001), "Relational Meaning and Discrete Emotions," in *Appraisal Processes in Emotion: Theory, Methods, and Research*, K. R. Scherer, A. Schorr, and J. Johnstone, Eds., Oxford University Press, New York.

Leana, C. R., Locke, E. A., and Schweiger, D. M. (1990), "Fact and Fiction in Analyzing Research on Participative Decision Making: A Critique of Cotton, Vollrath, Frogatt, Lengnick-Hall, and Jennings," *Academy of Management Review*, Vol. 15, pp. 137–146.

Levi, L. (1972), "Stress and Distress in Response to Psychosocial Stimuli," *Acta Medica Scandinavia*, Vol. 191, Suppl. 528.

Locke, E. A., and Schweiger, D. M. (1979), "Participation in Decision Making: One More Look," *Research in Organizational Behavior*, Vol. 1, pp. 265–339.

Lodge, G. C., and Glass, W. R. (1982), "The Desperate Plight of the Underclass: What a Business–Government Partnership Can Do About Our Disintegrated Urban Communities," *Harvard Business Review*, Vol. 60, pp. 60–71.

Lundberg, U., and Frankenhaeuser, M. (1980), "Pituitary–Adrenal and Sympathetic–Adrenal Correlates of Distress and Effort," *Journal of Psychosomatic Research*, Vol. 24, pp. 125–130.

Marrow, A. F. (1969), *The Practical Theorist: The Life and Work of Kurt Lewin*, Basic Books, New York.

Matthews, K. A., Cottington, E. M., Talbott, E., Kuller, L. H., and Siegel, J. M. (1987), "Stressful Work Conditions and Diastolic Blood Pressure among Blue Collar Factory Workers," *American Journal of Epidemiology*, Vol. 126, No. 2, pp. 280–291.

McCormick, E. J. (1970), *Human Factors Engineering*, 3rd ed., McGraw-Hill, New York.

McGregor, D. (1960), *The Human Side of Enterprise*, McGraw-Hill, New York.

Medsker, G. J., and Campion, A. (2001), "Job and Team Design," in *Handbook of Industrial Engineering*, G. Salvendy, Ed., Wiley, New York.

Meister, D. (2001), "Fundamental Concepts of Human Factors," in *International Encyclopedia of Ergonomics and Human Factors*, W. Karwowski, Ed., Taylor & Francis, New York, pp. 68–70.

Monk, T. H., and Tepas, D. I. (1985), "Shift Work," in *Job Stress and Blue-Collar Work*, C. L. Cooper and M. J. Smith, Eds., Wiley, New York, pp. 65–84.

Moray, M. (2000), "Culture, Politics and Ergonomics," *Ergonomics*, Vol. 43, No. 7, pp. 858–868.

Moray, N. (1994), "De Maximus non Curat Lex" or "How Context Reduces Science to Art in the Practice of Human Factors," in *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 526–530.

Nadler, G., and Hibino, S. (1994), *Breakthrough Thinking: The Seven Principles of Creative Problem Solving*, 2nd ed., Prima Communications, Rocklin, CA.

National Academy of Sciences (NAS) (1983), *Video Displays, Work and Vision*, National Academy Press, NAS, Washington, DC.

National Institute for Occupational Safety and Health (NIOSH) (1992), *Health Hazard Evaluation Report: HETA 89-299-2230-US West Communications*, U.S. Department of Health and Human Services, Washington, DC.

Noro, K. (1991), "Concepts, Methods, and People," in *Participatory Ergonomics*, K. Noro and A. Imada, Eds., Taylor & Francis, London.

Noro, K. (1999), "Participatory Ergonomics," in *Occupational Ergonomics Handbook*, W. Karwowski and W. S. Marras, Eds., CRC Press, Boca Raton, FL, pp. 1421–1429.

Oborne, D. J., Branton, R., Leal, F., Shipley, P., and Stewart, T. (1993), *Person-Centered Ergonomics: A Brantonian View of Human Factors*, Taylor & Francis, London.

Ostberg, O., and Nilsson, C. (1985), "Emerging Technology and Stress," in *Job Stress and Blue-Collar Work*, C. L. Cooper and M. J. Smith, Eds., Wiley, New York, pp. 149–169.

Owen, R. (1816), *A New View of Society*; 2004 ed., Kessinger Publishing, Kila, MT.

Pacey, A. (1991), *Technology in World Civilization: A Thousand-Year History*, MIT Press, Cambridge, MA.

Parker, M., and Slaughter, J. (1994), *Working Smart: A Union Guide to Participation Programs and Reengineering*, Labor Notes, Detroit, MI.

Perrow, C. (1983), "The Organizational Context of Human Factors Engineering," *Administrative Science Quarterly*, Vol. 28, pp. 521–541.

Rahimi, M. (1995), "Merging Strategic Safety, Health, and Environment into Total Quality Management," *International Journal of Industrial Ergonomics*, Vol. 16, pp. 83–94.

Ramazzinni, B. (1700), *De Morbis Artificum Diatriba. Modena: Antonii Capponi* [Diseases of Workers], translation by W. C. Wright, 1940, University of Chicago Press, Chicago, IL.

Reason, J. (1990), *Human Error*, Cambridge University Press, New York.

Reason, J. (1997), *Managing the Risks of Organizational Accidents*, Ashgate, Brookfield, VT.

Reason, J., and Hobbs, A. (2003), *Managing Maintenance Error: A Practical Guide*, Ashgate, Brookfield, VT.

Reeves, C. A., and Bednar, D. A. (1994), "Defining Quality: Alternatives and Implications," *Academy of Management Review*, Vol. 19, pp. 419–445.

Rose, R. M., Jenkins, C. D., and Hurst, M. W. (1978), *Air Traffic Controller Health Change Study*, U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Medicine, Washington, DC.

Rutenfranz, J., Colquhoun, W. P., Knauth, P., and Ghata, J. N. (1977), "Biomedical and Psychosocial Aspects of Shift Work," *Scandinavian Journal of Work Environment and Health*, Vol. 3, pp. 165–182.

Sainfort, F., Taveira, A. D., Arora, N. K., and Smith, M. J. (2001), "Teams and Team Management and Leadership," in *Handbook of Industrial Engineering*, G. Salvendy, Ed., Wiley, New York.

Sainfort, P. C. (1989), "Job Design Predictors of Stress in Automated Offices," *Behaviour and Information Technology*, Vol. 9, No. 1, pp. 3–16.

Sanders, M. S., and McCormick, E. J. (1993), *Human Factors in Engineering and Design*, McGraw-Hill, New York.

Sashkin, M. (1984), "Participative Management Is an Ethical Imperative," *Organizational Dynamics*, Spring, pp. 5–22.

Schnall, P. L., Pieper, C., Schwartz, J. E., Karasek, R. A., Schlussel, Y., Devereux, R. B., Ganau, A., Alderman, M., Warren, K., and Pickering, T. G. (1990), "The Relationship between 'Job Strain,' Workplace Diastolic Blood Pressure, and Left Ventricular Mass Index," *Journal of the American Medical Association*, Vol. 263, pp. 1929–1935.

Selye, H. (1956), *The Stress of Life*, McGraw-Hill, New York.

Shackel, B. (1996), "Ergonomics: Scope, Contribution and Future Possibilities," *The Psychologist*, Vol. 9, No. 7, pp. 304–308.

Shahnavaz, H. (2000), "Role of Ergonomics in the Transfer of Technology to Industrially Developing Countries," *Ergonomics*, Vol. 43, No. 7, pp. 903–907.

Shirom, A., Toker, S., Berliner, S. and Shapira, I. (2008), "The Job Demand-Control-Support Model and Stress-Related Low-Grade Inflammatory Responses among Healthy Employees: A Longitudinal Study," *Work & Stress*, Vol. 22, No. 2, pp. 138–152.

Smith, A. (1776), *An Inquiry into the Nature and Causes of the Wealth of Nations*; 2003 ed., Bantam Classics, New York.

Smith, J. H., and Smith, M. J. (1994), "Community Ergonomics: An Emerging Theory and Engineering Practice," in *Proceedings of the Human Factors and Ergonomics Society, 38th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 729–733.

Smith, J. H., Cohen, W., Conway, F., and Smith, M. J. (1996), "Human Centered Community Ergonomic Design," in *Human Factors in Organizational Design and Management*, Vol. 5, J. O. Brown and H. W. Hendrick, Eds., Elsevier Science, Amsterdam, pp. 529–534.

Smith, J. H., Cohen, W. J., Conway, F. T., Carayon, P., Derjani-Bayeh, A., and Smith, M. J. (2002), "Community Ergonomics," in *Macroergonomics: Theory, Methods, and Applications*, H. W. Hendrick and B. M. Kleiner, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 289–309.

Smith, K. U. (1965), *Behavior Organization and Work*, rev. ed., College Printing and Typing Company, Madison, WI.

Smith, K. U. (1966), "Cybernetic Theory and Analysis of Learning," in *Acquisition of Skill*, Academic, New York.

Smith, K. U., and Kao, H. (1971), "Social Feedback: Determination of Social Learning," *Journal of Nervous and Mental Disease*, Vol. 152, No. 4, pp. 289–297.

Smith, M. J. (1987a), "Occupational Stress," in *Handbook of Human Factors*, G. Salvendy, Ed., Wiley, New York, pp. 844–860.

Smith, M. J. (1987b). "Mental and Physical Strain at Computer Workstations," *Behaviour and Information Technology*, Vol. 6, pp. 243–255.

Smith, M. J., and Carayon, P. C. (1995), "New Technology, Automation and Work Organization: Stress Problems and Improved Technology Implementation Strategies," *International Journal of Human Factors in Manufacturing*, Vol. 5, pp. 99–116.

Smith, M. J., and Carayon-Sainfort, P. (1989), "A Balance Theory of Job Design for Stress Reduction," *International Journal of Industrial Ergonomics*, Vol. 4, pp. 67–79.

Smith, M. J., Cohen, B. G. F., Stammenjohn, L. W., Jr., and Happ, A. (1981), "An Investigation of Health Complaints and Job Stress in Video Display Operations," *Human Factors*, Vol. 23, pp. 389–400.

Smith, M. J., Colligan, M. J., and Tasto, D. L. (1982), "Health and Safety Consequences of Shift Work in the Food Processing Industry," *Ergonomics*, Vol. 25, No. 2, pp. 133–144.

Smith, M. J., Carayon, P., Sanders, K. J., Lim, S. -Y., and LeGrande, D. (1992), "Electronic Performance Monitoring, Job Design and Worker Stress," *Applied Ergonomics*, Vol. 23, No. 1, pp. 17–27.

Smith, M. J., Carayon, P., Smith, J. H., Cohen, W., and Upton, J. (1994), "Community Ergonomics: A Theoretical Model for Rebuilding the Inner City," in *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 724–728.

Smith, M.J., Karsh, B-T. and Moro, F.B. (1999). A review of research on interventions to control musculoskeletal disorders. In *Work-Related Musculoskeletal Dirorders*. Washington, DC: National Research Council, National Academy Press, 200–229.

Smith, M. J., Derjani-Bayeh, A. and Carayon, P. (2009), "Community Ergonomics and Globalization: A Conceptual Model of Social Awareness," in *Industrial Engineering and Ergonomics*, C. M. Schlick, Ed., Springer, Berlin, pp. 57–66.

Stanney, K. M., Maxey, J., and Salvendy, G. (1997), "Socially Centered Design," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York, pp. 637–656.

Stanney, K. M., Maxey, J., and Salvendy, G. (2001), "Socially Centered Design," in *International Encyclopedia of Ergonomics and Human Factors*, W. Karwowski, Ed., Taylor & Francis, New York, pp. 1712–1714.

Stuebbe, P. A., and Houshmand, A. A. (1995), "Quality and Ergonomics," *Quality Management Journal*, Winter, pp. 52–64.

Tasto, D. L., Colligan, M. J., Skjei, E. W., and Polly, S. J. (1978), *Health Consequences of Shift Work*, U.S. Department of Health, Education and Welfare, Publication NIOSH-78-154, U.S. Government Printing Office, Washington, DC.

Taveira, A. D., and Hajnal, C. A. (1997), "The Bondage and Heritage of Common Sense for the Field of Ergonomics," in *Proceedings of the 13th Congress of the International Ergonomics Association*, Tampere, Finland.

Taveira, A. D., James, C. A., Karsh, B., and Sainfort, F. (2003), "Quality Management and the Work Environment: An Empirical Investigation in a Public Sector Organization," *Applied Ergonomics*, Vol. 34, pp. 281–291.

Taylor, F. W. (1912), *The Principles of Scientific Management*; 1998 reprint, Dover Publications, New York.

Trist, E. (1981), *The Evolution of Socio-Technical Systems*, Ontario Quality of Working Life Centre, Toronto, Ontario, Canada.

Turner, J., and Karasek, R. A. (1984), "Software Ergonomics: Effects of Computer Application Design Parameters on Operator Task Performance and Health," *Ergonomics*, Vol. 27, No. 6, pp. 663–690.

Van Ameringen, M. R., Arsenault, A., and Dolan, S. L. (1988), "Intrinsic Job Stress and Diastolic Blood Pressure among Female Hospital Workers," *Journal of Occupational Medicine*, Vol. 30, No. 2, pp. 93–97.

Warrack, B. J., and Sinha, M. N. (1999), "Integrating Safety and Quality: Building to Achieve Excellence in the Workplace," *Total Quality Management*, Vol. 10, Nos. 4–5, pp. S779–S785.

Weick, K. E. (1987), "Organizational Culture as a Source of High Reliability," *California Management Review*, Vol. 24, pp. 112–127.

Weisbord, M. R. (2004), *Productive Workplaces Revisited: Dignity, Meaning, and Community in the 21st Century*, Jossey-Bass, San Francisco.

White, L. (1962), *Medieval Technology and Social Change*, Clarendon, Oxford.

Wilson, J. R. (1995), "Ergonomics and Participation," in *Evaluation of Human Work: A Practical Ergonomics Methodology*, Taylor & Francis, London, pp. 1071–1096.

Wilson, J. R. (2000), "Fundamentals of Ergonomics in Theory and Practice," *Applied Ergonomics*, Vol. 31, pp. 557–567.

Wilson, J. R., and Grey Taylor, S. M. (1995), "Simultaneous Engineering for Self Directed Teams Implementation: A Case Study in the Electronics Industry," *International Journal of Industrial Ergonomics*, Vol. 16, pp. 353–366.

Wilson, J. R., and Haines, H. M. (1997), "Participatory Ergonomics," in *Handbook of Human Factors and Ergonomics*, G. Salvendy, Ed., Wiley-Interscience, New York, pp. 353–366.

Wisner, A. (1985), "Ergonomics in Industrially Developing Countries," *Ergonomics*, Vol. 28, No. 8, pp. 1213–1224.

Wisner, A. (1995a), "Understanding Problem Building: Ergonomic Work Analysis," *Ergonomics*, Vol. 38, No. 3, pp. 596–605.

Wisner, A. (1995b), "Situated Cognition and Action: Implications for Ergonomic Work Analysis and Anthropotechnology" *Ergonomics*, Vol. 38, No. 8, pp. 1542–1557.

# CHAPTER 10

# HUMAN FACTORS AND ERGONOMIC METHODS

**Julie A. Jacko**
**University of Minnesota**
**Minneapolis, Minnesota**


**Ji Soo Yi**
**Purdue University**
**West Lafayette, Indiana**


**François Sainfort**
**University of Minnesota**
**Minneapolis, Minnesota**


**Molly McClellan**
**University of Minnesota**
**Minneapolis, Minnesota**

## 1 INTRODUCTION

Methods are a core component in the successful practice of human factors and ergonomics (HF/E). Methods are necessary to (1) collect data about people, (2) develop new and improved systems, (3) evaluate system performance, (4) evaluate the demands and effects of work on people, (5) understand why things fail, and (6) develop programs to manage HF/E. A primary concern for the disciplines of HF/E resides in the ability to make generalizations and predictions about human interactions for improved productivity, safety, and overall user satisfaction. Accordingly, HF/E methods play a critical role in the corroboration of these generalizations and predictions. Without validated methods, predictions and generalizations would be approximate at best, and HF/E principles would present unfounded theories informed by common sense and by anecdotal observations and conclusions. HF/E methods are the investigative toolkits used to assess user and system characteristics as well as the resulting requirements imposed on the abilities, limitations, and requirements of each. HF/E methods are implemented via scientifically grounded empirical investigative

techniques that are categorized into experimental research, descriptive studies, and evaluation research. Further discrimination is made based on psychometric properties, practical issues, and descriptive, empirical, and evaluation research methodological processes.

By the very nature of its origin, HF/E is an interdisciplinary field of study comprising aspects of psychology, physiology, engineering, statistics, computer science, and other physical and social sciences. In a bibliometric analysis of the journal *Human Factors* a steady trend of more authors per paper was discovered (Lee et al., 2005). This may be an expression of the increasingly interdisciplinary nature of human factors research and a product of the need to address human interaction in more complex technology systems.

Why, then, bother with a chapter devoted to HF/E methods when it is an obvious union of well-documented methods offered by this wide array of subject matters? The answer is that the discipline of HF/E is concerned with understanding human-*integrated* systems. That is, HF/E researchers and practitioners strive to understand how the body, mind, machine, software, systems, rules, environment, and so on, work in harmony or dissonance and how to improve those relationships

and outcomes. Although HF/E is a hybrid of other disciplines concerned with designing systems and products for human use, it demonstrates unique characteristics that make it a distinct field of its own (Meister, 2004; Proctor and Vu, 2010).

Fitts's law provides a classic example of the application of basic psychological principles to HF/E problems. Fitts's law is a model of psychomotor behavior, predicting human reaction time. The model relates movement time, the size of two targets, and the distance between them (Fitts, 1954). Fitts's law is a principle that has been adapted and embedded within several HF/E studies of systems that employ similar types of psychomotor responses of the user. Examples include interface design selection, performance prediction, allocation of operators, and even movement time prediction for assembly line work.

However, the original Fitts's law was developed under certain assumptions in its application. This includes error-free performance and the assumption of one-dimensional movement along a single axis and specifies no guidelines for the input devices that control the movement (e.g., mouse, lever, etc.). This is due to the fact that the original Fitts's law was developed *independent* of any specific system, human variability, and other contextual factors. HF/E specialists therefore have to take these factors into account for the application of Fitts's law to various systems. An example of this is the work of MacKenzie, which extends Fitts's law to human–computer interaction (HCI) research and design. MacKenzie has, in basic research, manipulated Fitts's law to account for the use of a mouse and pointer, two-dimensional tasks, and more (MacKenzie, 1992a,b). Clearly, concessions were made in order to apply Fitts's law within the context of the HCI system.

As noted previously, Fitts's law supports the prediction of speeds but does not support the prediction of errors. Following the original work utilizing Fitts's law for HCI research, an error model for pointing has been developed. Wobbrock et al. (2008) developed an error model that can be used with Fitts's law to estimate and predict error rates along with speeds. In addition to predicting a new error model, their research also took a departure from Fitts's law. These researchers found that target size had a greater effect on error rate than target distance. This is contrary to what Fitts's law states—that target size and distance contribute proportionally to the index of difficulty (ID).

Perhaps the most demanding challenge encountered by HF/E investigators is deciding the most appropriate methodology to address their goals and research questions (Wickens et al., 2003a). Multitudes of basic scientific methodologies, principles, and metrics are available to HF/E researchers and practitioners. An accompanying challenge of HF/E methods emerges as how to use these methods to assess the human factor within the system context, a feature that is often absent in the traditionally basic research methods. The transformation of physical and social sciences into methodologies applicable to HF/E often generates conflicting goals and assumptions. The HF/E investigator is faced with trade-offs in how they select and actuate the methods. The distinctive

nature of the HF/E disciplines is certainly reflected in the methods used by both researchers and practitioners.

The appendix to this chapter is a set of tables that summarize the methods employed in a variety of HF/E investigations. This collection provides a snapshot of the current application of HF/E methods in the published scholarly journals *Ergonomics* and *Human Factors* for the period 2000–2010. Although the tables are not comprehensive, it is representative of the types of methods selected and how they are applied in the various subdisciplines of HF/E. The discipline(s), goal, and method(s) used by authors are underscored and labeled according to different methodological characteristics that are defined in this chapter. One can appreciate the variety of basic methods that HF/E investigators encounter in the generalization and prediction of human interaction with systems and physical machines.

The studies featured apply a range of methods, from highly controlled laboratory settings to loosely structured observational field studies. The variety of goals presented by these authors fit into two categories: (1) HF/E methods that develop and test scientific principles and theories as they apply to human-integrated systems and (2) HF/E methods that focus on applied problems, incorporating specific features of the target population, task environment, and/or the system. In both cases, the investigators are concerned with the human performance or behavior embedded in some system. HF/E researchers and practitioners should look regularly to the scientific literature to glean the types of methods that investigators select to achieve their goals.

No absolute right or wrong exists in method selection and application. However, some methods are more appropriate than others, influenced greatly by circumstantial factors. This is what makes this discipline of HF/E both exciting and frustrating at times for researchers and practitioners alike. The most common answer to the appropriateness of a method for a given HF/E objective is: *It depends*. The specific combination of methods used and control exerted depends heavily on task factors and study goals, among other key contextual factors. The studies listed in the HF/E literature sample and a handful of others referred to provide readers with real examples of the application of HF/E methods employed by investigators in an effort to realize a variety of goals under an assortment of assumptions. Four of the studies from the review have been selected as specific case studies of field studies, survey methodologies, empirical methods, and evaluation studies. Experience is one of the best tools to apply in the selection of methods. Much can be learned from the experiences of others, as their goals, methodologies, and procedures are published in the HF/E literature.

In this chapter we present the reader with critical issues in the selection and execution of HF/E methods such as the ethical issues of working with people, psychometrics issues, and practical constraints. We do not aim to provide readers with all the answers to problems in applying HF/E methods. Instead, readers should gain an improved sense of which questions they should ask prior to implementation of HF/E methods in research or practice. Moreover, we do not provide instructions for

the implementation of specific methodologies. Instead, we introduce the most relevant facets of methods for addressing critical issues. Classifications of methods are provided in ways that can direct HF/E investigators in their selection and planning. The data and information gathered through HF/E methods are analyzed and applied in ways unlike traditional psychology or engineering. Chapter 44 is devoted to HF/E outcomes.

## 2  HF/E RESEARCH PROCESS

Although a significant amount of improvisation is required of HF/E researchers to account for contextual factors while preserving experimental control, a general framework for HF/E investigations can be constructed. This framework is supported by psychometric attributes, previous research methods, principles, outcomes, and ethics of investigations. The foundation of this framework is the specific goals established for a given investigation. Figure 1 presents a schematic of the framework. Each decision point and its relative attributes are addressed throughout the chapter, beginning with assertion of HF/E goals.

### 2.1  Problem Definition

When evaluating human system interaction, first you must take into account the goals, knowledge, and procedures of the human operator; the system and its interface; and the operational environment (Bolton and Bass, 2009). What motivates HF/E investigations in addition to the underlying desire to improve human-integrated system safety and efficiency is usually the recognition of a problem by the HF/E researcher or specialist, management, or a funding agency. For example, the management team at a software call center may ask the usability group to evaluate the problematic issues of the installation of their software package. Alternatively, a government funding agency may issue a call for proposals to discover the source of errors in hospital staffs' distribution of medication to patients. HF/E investigators also come across ideas from reading relevant scientific literature, networking with colleagues and peers at work and conferences, observing some novel problem, or even attempting to reveal the source of unexplained variance in their or someone else's research. Problems usually stem from a gap in the research, contradictory sets of results, or the occurrence of unexplained facts (Weimer, 1995). These problems are influential in defining the purpose of the investigation at hand as well as subsequent decisions throughout the application of HF/E methodologies. The first important criterion is to determine the purpose or scope of the investigation. The goal of the investigation is critical. Methods selected have to be relevantly linked to the goal for the investigation to succeed.

Investigations may be classified as *basic* or *applied* (Weimer, 1995). Of course, as with most HF/E theories and principles, these are not completely dichotomous. Studies that are basic are explanation driven, with the purpose of contributing to the advancement of scientific knowledge. Basic research journals tend to be cited by other journals, thus demonstrating how instrumental basic research is in contributing to the scientific knowledge base (Lee et al., 2005). Basic investigations may

seem out of place because HF/E is so applied in nature. However, explanation-driven basic methods serve a critical role presenting solutions to real-world H/FE problems. Basic research is aimed at advancing the understanding of factors that influence human performance in human-integrated systems. The desired outcome of this type of research is to establish general principles and theories with which to explain them (Miller and Salkind, 2002). An example of basic research is a study of the impact of a variety of different feedback modalities on user performance in a series of drag-and-drop tasks with a desktop computer (Jacko et al., 2005). This publication reported the efficacy of some nonvisual and multimodal feedback forms as potential solutions for enhanced performance. Basic research may employ a variety of experimental methodologies. Basic investigations in HF/E are comparable to the nomothetic approaches used in psychology. The nomothetic approach uses investigations of large groups of people in order to find general laws that apply to everyone (Cohen and Swerdlik, 2002; Barlow and Nock, 2009).

Pollatsek and Rayner (1998) present classifications and explanations for several basic methodologies of tracking human behavior. These include psychophysical methods (subjective, discrimination, and tachiscopic methods), reaction time methods, processing time methods, eye movement methods, physiological methods, memory methods, question-answering methods, and observational methods. A characteristic of basic investigations is that the majority of these methods are operationalized in highly controlled settings, usually in an academic setting (Weimer, 1995). Most commonly, basic studies incorporate theories and principles stemming from behavioral research, especially experimental psychology. HF/E basic research goes beyond basic experimental psychology, conceiving the basic theories that explain the human–system interaction, not just the human in isolation (Meister, 1971). The development of human models of performance, such as ACT-R/PM (Byrne, 2001), and using them to evaluate software applications demonstrate the integration of basic theories of human information and physiological processes to create an improved understanding of human–system interactions relevant to a variety of applications.

Applied research directs the knowledge from basic research into real-world problems. Work in applied research is focused on system definition, design, development, and evaluation. Applied investigations are, in a sense, supplements to basic research. They would lack merit in the absence of basic research. For example, unlike basic research journals that tend to be cited by other journals, applied journals are likely to cite other journals (Lee et al., 2005). A characteristic of applied investigations is that the problems identified are typically too specific for their solutions to be generalizable. These investigations are implementation driven, with very specific goals to apply the outcomes relevant to specific applications, tasks, populations, environments, and conditions. Applied studies focus on system definition, system design, development, and evaluation. Investigations of the applied nature are used to assess problems, develop requirements for the human and/or machine,
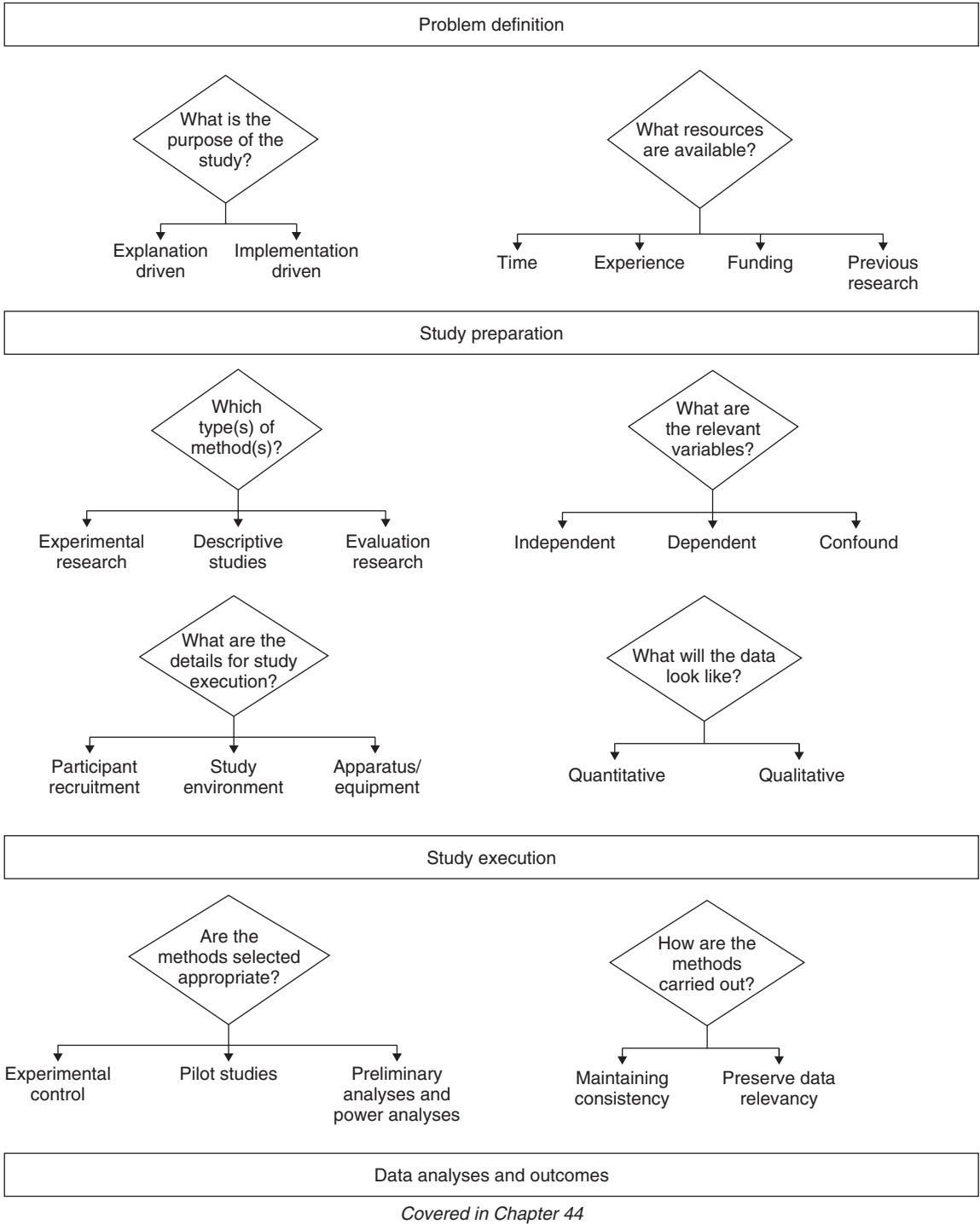
**Figure 1**  Framework of steps in selection and applications of HF/E methods.

and evaluate performance (Committee on Human Factors, 1983).

Applied or implementation-driven work is typically associated with work in the field. Participants, tasks, environments, and other extraneous variables usually need to closely match the actual real-world situation to truly answer the problems at hand. That said, applied methods are abundant and diverse and very much reflect the vast number of HF/E disciplines (mentioned above). Investigators often modify their techniques due to external demands and constraints of the operational system and environment (losing control over what is available for assessment and allowing confounding interactions to occur) (Wixon and Ramey, 1996). An example of this class of applied investigations from the literature summary is found in the analysis of interorganizational coordination after a railway accident (Smith and Dowell, 2000). Other studies have demonstrated that breakdown in the planning and execution of emergency response operations can have potentially disastrous consequences (Mendonca et al., 2001). These activities are often carried out under conditions of considerable time pressure and high risk emphasizing the need for interorganizational communication (Riley et al., 2006). Stanton and Young (1999) present a comparison of two car stereos to compare several applied methodologies. Some examples of applied methodologies include keystroke-level models (KLMs) (Card et al., 1983), checklists, predictive human error analyses, observations, questionnaires, task analyses, error analyses, interviews, heuristic evaluations, and contextual design.

Based on the framework presented in Figure 1, the development of a hypothesis is an important first step in both basic and applied research. Between the two types of research, the difference in the hypotheses is granularity. Hypotheses formulated for applied research are much more specific in terms of applied context. In either case, a hypothesis should be in the form of a proposition: *If A, then B*. Directional hypotheses are most commonly used in applied research. This is where an investigator makes a prediction hypothesis regarding the outcome of the research (Creswell, 2009). An example of a directional hypothesis would be "Scores for group B will be higher than for group A" following an intervention.

A problem must be testable if HF/E methods are to be applied. Generally, a problem is testable if it can be translated into the hypothesis format, and the likeliness of truth or falsity of that statement is attainable (Weimer, 1995). However, just the fact that a problem is testable does not ensure that results will be widely applicable or useful. Factors that can affect the applicability and acceptability of the results are discussed in subsequent sections.

## 2.2 Choosing the Best Method

The choice of HF/E method is influenced by several factors, as the decision to employ a specific methodology elicits several consequences relevant to the efficacy of that method in meeting the established goals. The ability to generalize the results of investigations is shaped by both the design/selection of methods and statistical analysis (see Chapter 44). The "study preparation" section of

the framework presented in Figure 1 illustrates the various factors influencing the selection of methods. The judicious selection and implementation of HF/E methods entails a clear understanding of what information will be collected or what will provide the information, how it will be collected, how it is analyzed, and how the method is presented as relevant to the predetermined objectives and hypotheses.

Several authors offer opinions on what the most important considerations should be in the selection of methods. Stanton and Young (1999) provide one of a handful of comparative examinations looking into the utility of various different descriptive methodologies for HF/E. They present case studies evaluating 12 different methodologies based on their use in the investigation of automobile radio controls. The authors evaluated the methods on the criteria of reliability, validity, resources required, ease of use, and efficacy. The accuracy required, criteria to be evaluated, acceptability of the method (to both participants and investigators), abilities of those involved in the process, and a cost–benefit analysis of the method are additional deciding factors for implementation. Later, Stanton et al. (2005) contributed to the literature again with a book detailing the results of a HF database with over 200 methods and techniques. At the very least, an investigator needs to be conscious of the attributes that are present in their chosen methods and the possible impact to avoid misrepresenting results and forming ill-conceived conclusions.

Kantowitz (1992) focused on reliability and validity by looking to problem representation, problem uniqueness, participant representativeness, variable representativeness, and setting representativeness (ecological validity). The specific selection of methodology is rarely covered in HF/E texts. Instead, most authors jump directly to method selection (e.g., descriptive, experimental, and evaluative) and look at a variable and metric definition. However, variable selection, definition, and the resulting validity are intertwined decisively with the methods selected and must link back in a relevant way to the investigation's objectives. In this section, methodological constraints are broken down into two categories:

1. *Practical concerns*
   - Intrusiveness
   - Acceptability
   - Resources
   - Utility
2. *Psychometric concerns*
   - Validity (uniqueness)
     - Construct validity
     - Content validity
     - Face validity
   - Reliability (representation)
     - Accuracy and precision
   - Theoretical foundation
   - Objectivity

Humans are by nature complex, unreliable systems. Kantowitz (1992) asserted that considered as a

stand-alone system human complexity supersedes that of a nuclear power plant. This creates an abundance of convolutions when considering the human system embedded within another system (social, manufacturing, technological). Humans can be inconsistent in their external and internal behaviors, which are often sensitive to extraneous factors (overt and covert). Undeniably, this creates a conundrum for the HF/E investigator. The impact of this variability can, to a certain extent, be mitigated through various strategies in method selection and implementation. This usually entails close examination and careful attention by the investigator to the relevant variables, including their definition, collection, and analysis. In addition to human fallibility, the selection of any given method and its execution is critically influenced by the objective of the study (basic/applied), objective clarification (i.e., hypothesis), experience of the investigator(s), resources (money, time, staff, equipment, etc.), and previously validated (relevant) research. Returning to the human side of method selection, the selection of the method is informed by the ethical and legal requirements of working with human beings as participants.

In practice, it is not feasible to comply with all of the psychometric and practical issues that occur in conjunction with HF/E methodologies. More often than not, the investigator must weigh the implications of their method choice on the desired outcome of the study. Investigators also prioritize the requirements placed on their work with respect to potential impact on the study. For example, the ethical treatment of human participants is of high priority, because an investigator's ethics approval organization [e.g., institutional review board (IRB) in the United States] or funding agency may choose to terminate the study if risks are posed by participation. In this section, we illustrate these potential issues further. It is important for the reader to have an awareness of these issues before our discussion of various methods. In this way, novices can examine the HF/E methods more critically with respect to practical constraints most relevant to their research and work.

### 2.2.1 Practical Concerns

Practical concerns for the application of HF/E methods should be fairly obvious to the HF/E investigator. However, they must be taken into account early in the planning process and revisited continually. Brief definitions for the practical concerns follow.

*Intrusiveness*   This is an appraisal of the extent to which the methodology used interferes with the system being measured. A measure that distracts the participant or interferes with their performance in other ways is intrusive. The extent to which an intrusive method causes covariance in recorded observations differs when applied to different scenarios (Rehmann, 1995). Task analysis is one of the least intrusive measures because in the initial stages of system design it can occur without a system being present to study (Wilson and Corlett, 2005).

*Acceptability*   This includes the appropriateness and relevance of the method as perceived by investigators, participants, and the HF/E community. For this, the investigator needs to perform an extensive literature review and also network with peers to understand their opinions of the method. Those who fund the research must also be accepting of the method (Meister, 2004).

*Resources*   This refers to the fact that many methods place prerequisites on investigator's resources. These include the time it takes for investigators to train and practice using the method, the number of people needed to apply the method, and any preliminary work that is needed before application of the method. In addition, the method may require the purchase of hardware and software or other special measurement instruments (Stanton and Young, 1999; Stanton et al., 2005). Investigations typically have limited financial assistance, which must be considered prior to the adoption of any methodology.

*Utility*   There are two types of utility relevant to HF/E methods: conceptual and physical utility (Meister, 2004). Research with conceptual utility yields results that are applicable in future research on human-integrated systems. Research that holds physical utility proves useful in the design and use of human-integrated systems. In general, investigators need to ensure usefulness and applicability of their proposed methods for the responsiveness of others to their results and easier dissemination of their findings in conference proceedings, journals, and texts.

### 2.2.2 Psychometric Concerns

In investigations of human-integrated systems, the methods used should possess certain psychometric attributes, including reliability, validity, and objectivity. Methods are typically used to apply some criteria or metric to a sample to derive a representation of the real world and subsequently link conclusions back to the established goals. As depicted in Figure 2, inferences are applied to the measurements to make generalizable conclusions about the real world. Interpretation is inclusive of statistical analyses, generalizations, and explanation of results (Weimer, 1995). The assignment of these inferences should be made to a unique set of attributes in the real world. How well these conclusions match the real world depends on a give and take between controlling for extraneous factors, without disrupting the important representative factors. For example, Figure 2 exemplifies a set of inferences in the shape of a square, which will not easily be matched up to the initial population sampled.

Several of the issues emergent in the selection and application of methods are attributable to representation and uniqueness (Kantowitz, 1992). Representation tends to inform issues of reliability, or the "consistency or stability of the measures of a variable over time or across representative samples" (Sanders and McCormick, 1993, p. 37). A highly reliable method will capture metrics with relatively low errors repeatedly over time. Attributes of reliability include accuracy, precision, detail, and resolution. Human and system reliability, which is a reference to failures in performance, is a topic apart from methodological issues of reliability.

Validity is typically informed by the issues of uniqueness. Validity is the index of truth of a measure,
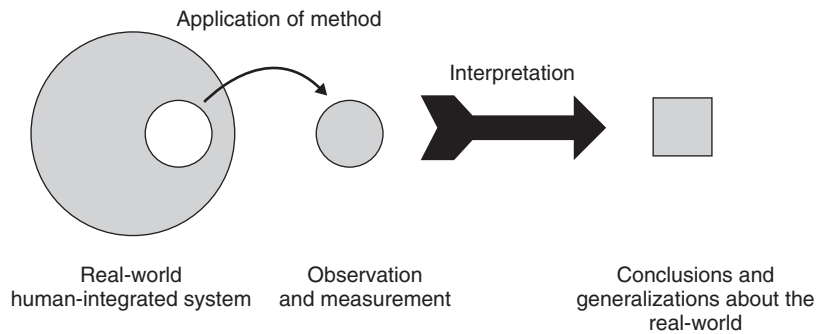
**Figure 2**  Application of HF/E methods.

or, in other words, if it actually captured what it set out to and not observing the extraneous (Kantowitz, 1992; Sanders and McCormick, 1993; Kanis, 2000; Wilson and Corlett, 2005). Both concepts are alluded to by many, defined by few, and measured by an even more select group of HF/E researchers and practitioners, with many different interpretations for this basic concept (Kanis, 2000). Despite the disparity in definitions and interpretations of the terms, it is generally agreed that these concepts are multifaceted.

Reliability and validity in HF/E are not dichotomous but lucid concepts as they may appear in the social sciences. The evolutionary nature of the HF/E discipline does not support such a "neatly organized" practice (Kanis, 2000). Reliability relates to how well the measure relates to itself, but validity relates to how well a measure correlates with external phenomena (Wilson and Corlett, 2005). A method must be reliable to be valid, but the reverse is not always true (i.e., reliable methods are not necessarily valid) (Gawron, 2000). Stanton and Young (1999) found this to be the case in their evaluation of hierarchical task analyses. The predictive validity of this method was found to be robust, but the reliability was less so. The authors concluded that the validity of the technique could not be accepted because of the underlying shortcomings in terms of

reliability. Types of validity include face, content, and construct. In this section we discuss different psychometric properties of both reliability and validity, how to control for them in the selection of methods, limitations imposed by the practical issues, and the resulting trade-offs. One disclaimer needs to be made before the discussion of validity and reliability that ensues: Although validity and reliability of methods enhance acceptance of the conclusion, they do not guarantee widespread utility of the conclusion (Kantowitz, 1992).

***Key Characteristics of Reliability***    Characteristics of reliability include accuracy and precision, which influence the consistency of the methodology over representative samples and the degree to which the methods and results are free from error. Accuracy is a description of how near a measure is to a standard or true value. Precision details the degree to which several methods provide closely related results, observable through distribution of the results. Test–retest reliability is a way to assess the precision of a given method. This is simply an assessment of correlations between separate applications of the methods. Sanders and McCormick (1993) report that for HF/E test–retest reliability scores of 0.80 and above are usually satisfactory. This score should be taken in context, however, because what determines an acceptable test–retest reliability score is intertwined with the specific contextual factors of the investigation.

The level of precision and/or accuracy sought in HF/E method selection and implementation is heavily contextually dependent. The investigator needs to select the method with reliability that is consistent with the requirements alluded to in the goals and problems. The KLM introduced by Card et al. (1983) was one of the first predictive methods for the field of HCI. This method predicts the time to execute a task given error-free performance using four motor operators, one mental operator, and one system response operator. KLM predicts error-free behavior, so the functions to calculate the time for the operators would probably be consistent between the applications of the method. The accuracy of the KLM method is purportedly high for certain tasks (Stanton and Young, 1999; Stanton et al., 2005). However, the accuracy of the method could deviate drastically from what Stanton and Young observed in their evaluation of car stereo designs,
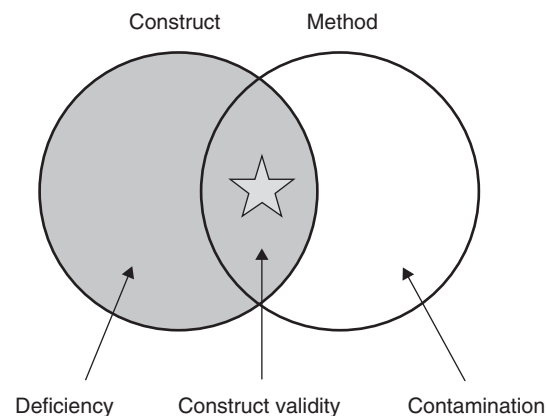


**Figure 3**  Construct validity. (Adapted from Sanders and McCormick, 1993.)

when KLM is applied to a different scenario, with the overall precision of the method constant. Additional disadvantages of KLM include the fact that it only models error-free expert performance, does not take context into account, and can only deal with serial, not parallel, activity (Stanton et al., 2005).

***Face Validity*** Face validity is defined as the extent to which the results look as though the method captured what is intended, which is the degree of consensus that a measure actually represents a given concept (Sanders and McCormick, 1993; Wilson and Corlett, 2005). It is a gauge of the perceived relevance of the methods to the identified goals of the investigation without any explanation by the investigator. Research based on actual tasks evaluated by real users has high face validity and is more likely to generalize to similar systems (Sage and Rouse, 2009). Face validity is important not only with respect to acceptance of the results reported by the scientific and practicing communities but also from the perspective of the participants in the study. If a measure seems irrelevant or inappropriate to the participants, it may affect their motivation in a negative way. People may not take their participation seriously if the methods seem disconnected from the purported goals. This can be mitigated first by briefing the participants on the purpose of the methods used or by collecting measures of the performance in the background, so the participant is not exposed to the specifics of the study.

***Content Validity*** The content validity of a method is essentially the scope of the assessment relevant to the domain of the established goals of the investigation. The analysis of Web logs provides an example of content validity. For example, consider an investigation with the goal to report employee use of a corporate intranet portal, which provides information on insurance benefits. Simply reporting the number of hits the portal receives does not possess high content validity. This is because this method provides no indication if the employees are actually pulling content from the intranet site. The Web logging methodology could instead look at various facets of employee activity on the intranet site to illustrate a more complete representation of use as well as actually to talk to some employees to get verbalizations and perceptions about the intranet site.

***Construct Validity*** Construct validity is best defined as the degree to which a method can be attributed to the underlying paradigm of interest. Figure 3 exemplifies the concept of construct validity and other relevant features. The gray-shaded circle on the left represents the model or theory under scrutiny, and the white circle on the right represents the space that is assessed by the selected measures. The star marks the intersection of these two spaces, which represents the construct validity of the measure. It represents the aspects of the target construct that are actually captured by the methods used. This measure leaves out elements of the construct (because it cannot account for the entirety of the concept), called the *deficiency* of a measurement. Equally, there are areas unrelated to the construct that the measure captures. This undesired or unintended area measured is termed the *contaminant*.

Physiological methods, such as heart rate, are especially prone to construct validity issues. For example, heart rate may be affected by caffeine, medications, increased mental workload, age, or physical stressors such as exercise. Consider a study that aims to measure the mental workload experienced by drivers while driving under different weather conditions by recording each driver's heart rate. The investigation will not necessarily detect the changes in mental workload as affected by the conditions. Instead, the investigator will also have captured a measure of the effects of coffee consumption, age, recent physical activity, medications, and mental workload, so that the effects of mental workload are virtually inseparable from the other *extraneous* contamination variants. The investigator can mitigate these contaminants through exercising control in the applications of their methods. In this case, specifying inclusion criteria for subjects' selection and participation in the study would be advantageous. Methods of control are discussed further in this section.

***Controlling for Reliability and Validity*** If not controlled during the selection and application of HF/E methods, problems with validity and reliability can prove detrimental to the generalizability and predictive value of conclusion. The following issues in the results are probably ascribable to matters of validity and reliability:

- A lack of correlation between reality and the criteria used
- A correlation of the criteria with unknown bias(es), so even if changes are detected, the absolute value of the factor(s) cannot be determined
- Multivariate correlations, because the construct of interest is actually affected by several factors
- Interference from extraneous factors may inappropriately suggest causal relationships when it is in fact just a correlation

Psychometric issues may be mitigated through the control of extraneous factors that can affect the construct and collecting data/observations in representative environments. That being said, fundamental conflicts often arise in trying to ascertain control without sacrificing critically representative aspects of the system, task, or population. Furthermore, time, financial, and practical constraints can make it impossible to ascertain desired levels of validity of H/FE methods. There are methods and approaches for the analysis of HF/E outcomes, discussed in Chapter 44, which can potentially account for some of the validity and reliability issues. Yet much like HF/E in the design process, the earlier changes are made in the selection and applications of a methodology to correct for validity and reliability issues, the more easily the changes are implemented and the greater positive impact they will have on the methodological outcome.

***Control*** Control in the selection and application of methods strives to challenge the sources of variance to which HF/E is highly prone. Sources of variance can include noise from the measurements, unexpected variance of the construct, and unexpected participant

behavior (Meister, 2004). It is the regulation of standard conditions to reduce the sources of variance. Variance is detrimental to HF/E methods because it restricts the certainty of inferences made. Control removes known confounding variables by making sure that the extraneous factors do not vary freely during the investigation (Wickens et al., 2003a). Control is not absolute. An investigator can exercise various levels of control to ensure minimal effects of confounding variables. Methods that lack control can lead to data that are virtually uninterpretable (Meister, 2004).

Ways to reduce variance include choosing appropriate participants, tasks, contexts, and measures; eliminating confounding variables to reduce covariance effects; implementing methods consistently; and increasing the structure with which the methodology is utilized. For example, control was exercised in research conducted to predict performance on a computer-based task for people with age-related macular degeneration (AMD). Great control was exercised in the selection of participants who had AMD and age-matched controls (Jacko et al., 2005). The selection of participants controlled for the exclusion of any person who had any ocular dysfunction other than AMD. Great care was also taken in ensuring that the variation of age between the experimental and control groups was consistent. If age had not been controlled for in recruitment, the differences between the two groups could be a result of interactions between age and ocular disease. Methods of control will be introduced in relation to experimental studies, descriptive studies, and evaluations.

***Participant Representativeness***    As stated earlier, human behavior is sensitive to a variety of factors, and interactions often surface between specific characteristics of the participant and the environment. The extent to which the results of investigations are generalizable depends on those connections between the characteristics of those observed and the actual population. Prior to beginning HF/E research, it is critical to know whether the users in the actual population will be beginners or experts, the frequency of system use, and the degree of discretion they will have in using the system (Wilson and Corlett, 2005). Although it is not always necessary only to sample participants from the actual population (Kantowitz, 1992), consistency checks should be made. In a study of age-related differences in training on home medical devices, presented in the literature summary table (Mykityshyn et al., 2002), the investigators needed to recruit persons from the aging population so that their age-related capabilities, mental and physical, would be consistent with that in the general population. Relevant aspects or attributes should be identified and present in the same proportion as the real population in order to specify and characterize the target user group (Sanders and McCormick, 1993; Wilson and Corlett, 2005).

***Variable Representativeness***    The selection of methods mandates the selection of necessary measures and variables. For HF/E, measurement is the assignment of value to attributes of human-integrated systems. Assigning value can be accomplished through various methods, such as nominal, ordinal, interval, and ratio scales. In Chapter 44 we discuss HF/E outcome measurements and their analysis in detail. A given measure affords a specific set of statistical summary techniques and inferences that have implications on validity and reliability.

Measurement selection and its assignment of value to events in human-integrated systems should be guided by theory and previous studies. In HF/E it is most germane to include more than one measure. Three classes of variables have been identified as necessary to capture human-integrated systems: (1) *system descriptive criteria*, which evaluate the engineering aspects of a system; (2) *task performance criteria*, which indicate the global measure of the interaction such as performance time, output quantities, and output qualities; and (3) *human criteria*, which capture the human's behavior and reactions throughout task performance through performance measures (e.g., intensity measures, latency measures, duration measures), physiological measures, and subjective responses (Sanders and McCormick, 1993).

Table 1 presents examples of task performance, human criteria, and system criteria. System descriptive criteria tend to possess the highest reliability and validity, followed by task performance criteria. Human criteria are the noisiest, with the most validity and reliability issues. Note that human criteria demonstrate the broadest classification of measurements. This is due to the inherent variability (and noise) in human data. These metrics—performance, physiological, and subjective responses—portray a more complete characterization of human experience when observed in combination. Performance optimization should not be pursued, say, at the cost of high levels of workload observed through heart rate and subjective measures using the NASA-TLX subjective assessment of mental workload. A useful guide in the selection of specific human measures is Gawron's *Human Performance, Workload, and Situational Awareness Measures Handbook* (2008), where over 100 performance, workload, and situational awareness measures are defined operationally for application in different methodologies.

"The utility of human factors research is linked intimately to the selection of measures" (Kantowitz,

**Table 1   Classification of Criteria Addressed in HF/E Methods**

| System Descriptive Criteria | Task Performance Criteria | Human Criteria |
|---|---|---|
| Reliability | Quantity of output | Performance |
| Quality measures | Output rate | Frequency |
| Operation cost | Event frequency | Latency |
| Capacity | Quality of output | Duration |
| Weight | Errors | Reliability |
| Bandwidth | Accidents | Physiological |
| | Variation | Cardiovascular |
| | Completion time | Nervous system |
| | Entire task | Sensory |
| | Subtask time | Subjective opinion |
| | | Situation awareness |
| | | Mental workload |
| | | Comfort |
| | | Ease of use |
| | | Design preference |

1992, p. 387). Validity and reliability of results are best substantiated when the three classes of criterion are addressed in the methodology. In fact, the selection of measurements or metrics is second in importance to method selection. Measurement and methodology are not independent concepts, and the selection of both are closely related and often iterative (Drury, 2005). The interpretation of results is greatly influenced by the combination of measures chosen in a particular test plan. Robust measurement techniques can make the interpretation process a much more streamlined process.

**Objectivity**   A second issue of variable representativeness is the level of objectivity in its definition and measurement. Objectivity is a function of the specific techniques employed in collecting and recording data and observations. Data and observations recorded automatically are the most objective approach. Objective variables can be captured without probing the participant directly. In contrast, in the collection of highly subjective variables, the participant is the *medium of expression* for the variable (Meister, 2004). The investigator may also interject subjectivity. Investigators can impose subjectivity and bias in how they conduct the investigation, which participants they choose to collect data from, and what they attend to, observe, and report.

For instance, three levels of objectivity can be demonstrated in capturing task time for a person to complete the assembly of widgets on a manufacturing line. A highly objective method may involve the use of the computer to register and store task time automatically based on certain events in the process (e.g., the product passes by a sensor on the manufacturing line). A less objective method would be to have the investigator capture the assembly with a stopwatch, where the investigator determines the perceived start and completion of the assembly. Finally, the least objective, most subjective method would be to ask the participant, without using a clock, to estimate the assembly time. Clearly, the level of objectivity in the methods influences the accuracy and precision of the outcomes. Subjective measures are not unwarranted and are in fact quite important. In the assembly example, if the worker perceives the assembly time for a certain component as very long, even if a more objective assessment method does not detect it as long, the workers' perceptions still affect the quality of the work they produce and the amount of workload they perceive, ultimately affecting the quality of the system and its resulting widgets. In eliciting covert mental processes from people for measurement and assessment, the investigator does, however, need to ensure that the results obtained will adequately answer the questions defined in the formulation of a project's goals.

Finally, how researchers choose to interpret the results of methods may also induce subjectivity into the outcome. By decreasing the level of involvement of either participant or investigator in the expression of performance, an increase in the overall objectivity of the method and outcomes can be realized. Tables 2 and 3 provide a taxonomy of methods and measures that summarize Meister's (2004) work. Subjective and objective methods are outlined with a specific example of each class of methods. In conclusion, both subjective and objective measurements have their place in HF/E investigations (Wickens et al., 2003a)

**Setting Representativeness**   Setting representativeness is the coherence between the environment where methods are performed and the real-world environment of the target situation where the results are to be applied. This is not necessarily a judgment of realism but, rather, the level of comparability between how the participants' physical and psychological processes are affected by the context of the study (Kantowitz, 1992). This informs the investigator's consideration for collecting data in the laboratory versus in the field. Another term for setting representativeness is the *ecological validity* of the study. Ecological validity influences the generalizability of the results but can also influence the behavior of those who participate in the investigation. Research has shown that participants can exhibit different behavior when they know they are being observed, a phenomenon known as the *Hawthorne effect* [for a complete overview of the Hawthorne effect, see Gillespie (1991)]. The more representative the task and environment and the less intrusive the investigations to the participants' behaviors, the better this effect can be mitigated. The investigator may retain a more complete picture of human behavior with a complex system in the actual, operational environment (Meister, 2004), which supports fewer objective measures.

HF/E investigators must ultimately decide where the best location is to collect data: in the field or in the laboratory. The collection of data in a field study versus in a highly controlled laboratory setting is a trade-off that HF/E practitioners and researchers continually debate in the execution of methods. Field research typically provides an investigation of the means to look at the system in order to shape their assumptions about the construct, in a way that informs the selection and implementation of other methods (Wixon and Ramey, 1996). Field studies naturally include context from the environment, supervision, motivation, and circumstances, although the investigator must keep in mind that their presence will also add an additional context, which could potentially invalidate the study (Wilson and Corlett, 2005). Wixon and Ramey (1996) claim further that most field studies are best suited for situations about which little is known, saving time in the laboratory studying the wrong problem. Conversely, fieldwork serves the purpose as an executable setting to validate theories and principles developed in more controlled, laboratory environment environments. Case Study 1 provides a summary of HF/E work from the literature in which field studies have been used.

## CASE STUDY 1: Effects of Task Complexity and Experience on Learning and Forgetting

The goal of this study by Nembhard (2000) was to investigate how task complexity and experience affect individual learning and forgetting in manual sewing tasks using worker-paced machinery that placed high demands on manual dexterity and hand–eye coordination.

**Table 2 Taxonomy of HF/E Objective Method**

Objective Methods

| Outcome Measure | Method | Description | Example(s) |
|---|---|---|---|
| Performance measures | Unmanaged performance | Measures of human-integrated system performance through unnoticed observations<br><br>*Assumptions*: (1) System of measurement is completely functional; (2) mission, procedures, and goals of the system are fully documented and available; (3) expected system performance is available in quantitative criteria to link human performance with system criteria | Evaluation of an advance brake warning system in government fleet vehicles (Shinar, 2000) |
| | Empirical assessment | Comparison of conditions of different system and human characteristics in terms of treatment conditions<br><br>*Assumption*: Experimental control of variables and representativeness will enable conclusions about the correlation between manipulated conditions for valid, generalizable results | Investigation of multimodal feedback conditions on performance in a computer-based task (Vitense et al., 2003) |
| | Predictive models of human performance | Applying theories of cognition and physiological processes and statistics to predict human performance (involves no participation by human participants)<br><br>*Assumptions*: (1) The model explains the cognitive processes to a reasonable degree; (2) the model incorporates contextual factors relevant to the operational environment | ACT-R/PM, a cognitive architecture to predict human performance using drop-down menus (Byrne, 2001) |
| | Analysis of archival data | Aggregated of data sets aimed at the representation of a particular facet of human-integrated systems; HF/E archival data from journal articles, subdivided into subtopics such as computers, health systems, safety, and aging<br><br>*Assumptions*: (1) Differences between individual study situations are small in nature; (2) error rate predicts performance with validity; (3) the models will be informed continually by new data studies | Anthropometric differences among occupational groups (Hsiao et al., 2002) |

**Methods** A notable amount of related research had been conducted in this domain through laboratory studies. Although the previous research was strong in finding causal inferences, it could not validate the findings for real-world situations. This gap in the knowledge base motivated the author to study the effects of task complexity and experience on performance in the factory. In the design of the study, the author ascribed task complexity and worker experience (e.g., training for their task) as two prominent factors determining the trends of learning and forgetting during task performance. Task complexity was measured by three variables: complexity of the method, machine, and material. Over the course of one year, the study captured 2853 episodes of learning/forgetting from all the workers. User performance was sampled 10 times per week and averaged to derive the learning/forgetting. The complexity variables and the worker experience variable were recorded in combination with each learning and forgetting episode.

**Analyses** Based on the data collected, the parameters, and the variables derived (e.g., prior expertise, steady-state productivity, rate of learning, and degree of forgetting), a mathematical model of learning and forgetting was developed. Then, using statistical methods such as Kolmogorov–Smirnov, analysis of variance (ANOVA), pairwise comparisons, and regression, the effects of learning and the effects of task complexity and experience on learning/forgetting were extrapolated.

Methodological Implications

1. *Expensive to Conduct*. As this study shows, a field study requires a larger number of samples (i.e., 2853 episodes) or observations to mitigate extraneous confounds. Thus, it can take more time and be costly.

2. *Strongly Valid*. Because the research hypotheses and questions are tested under real situations, the validity of the argument is usually strong. In fact, this served as the major motivation for this

**Table 3 Taxonomy of HF/E Subjective Methods**

Subjective Methods

| Outcome Measure | Method | Description | Example(s) |
|---|---|---|---|
| Observational measures | Observations | Information about what happened and is happening in the human-integrated system; status of the person(s) and system components and their characteristics of the outcomes<br>*Assumptions*: (1) What the observation recorded is the essence of what actually happened; (2) observers record the situation veridically; (3) interobserver differences are minimal with respect to reliability; (4) no questions are probed during observation | Task analyses of automated systems with which humans interact (Sheridan, 2002) |
| | Inspection | Similar to observation, but objects have a role in what is considered; comparisons between the object at hand and a predetermined guideline for the required characteristics of both the object and the target user<br>*Assumption*: The object of inspection has some deficiency, and the standards provided are accurate in representing what is truly required. | Ergonomic redesign of a poultry plant facility; data collected on tracking employee musculoskeletal disorders are compared against Occupational Safety and Health Administration (OSHA) guidelines (Ramcharan, 2001) |
| Self-reported measures | Interviews and questionnaires | Direct questioning for the participant(s) to express converts mental processes, including reasoning of perceptions of their interaction<br>*Assumptions*: (1) People can validly describe their response to different stimuli; (2) the words and phrasing used in the questions accurately capture what is intended; (3) credibility of the respondent in their ability to answer the questions; (4) formality of structure required in responses | Case study of disaster management (Smith and Dowell, 2000) |
| | Concurrent verbal protocol | Verbal protocols to elicit convert participant information processing while executing a task; participant explains and justifies actions while performing the task<br>*Assumption*: People can better explain processes when there are aspects of the ecological validity | Study of mental fatigue on a complex computer task (van der Linden et al., 2003) |
| Judgmental measures | Psychophysical methods | Questions that determine thresholds for discrimination perceptual qualities and quantities; size weight, distance, loudness, and so on<br><br>*Assumptions*: (1) The participant has a conceptual frame of reference for evaluations; (2) the judgment is a result of analysis of internal stimuli | Investigation of time estimation during sleep deprivation (Miro et al., 2003) |

study. This enables improved implementation of the results of this study back into the field more easily than those laboratory-based conjectures of potential causal relationships.

3. Complex *Analyses*. The data from a field study are naturally large and complex because they were captured under real situations. That is, they are subject to a lot of extraneous noise in the system observed. Therefore, strategic analytical methods are quite useful. For example, this study simplified the presentation of data by introducing a mathematical model of learning/forgetting.

Typically, the data from field methods are more subjective (coming mostly from surveys and observations), which affects the analysis of the outcomes. Three widely recognized field methods are (1) ethnography (Ford and Wood, 1996; Woods, 1996), (2) participatory design (Wixon and Ramey, 1996), and (3) contextual design (Holtzblatt and Beyer, 1996). These types of methods typically illustrate the big picture and do not provide the investigator with a simple yes-or-no answer (Wixon and Ramey, 1996). Instead, the data gathered tend to be information rich, somewhat subjective, and highly qualitative.

### 2.2.3 Trade-Offs

**Control versus Representation** The need for experimental control and representative environments, tasks, and participants creates a fundamental conflict. It is impossible to have both full control and completely representative environments (Kantowitz, 1992). This is because in representative environments participants control their environment, as they want to, barring any artificial constraints. Both highly controlled and representative methods serve important roles in HF/E. The answer of which to sacrifice, when faced with this predicament, is entirely dependent on the objectives of the study and the availability of resources. Ideally, an investigation should be able to incorporate aspects of both. The selection of H/FE methodologies is largely directed by the investigator's needs and abilities in terms of objectivity and control and how the results are to be applied (Meister, 2004). In fact, investigators often include specific aspects of the operational system while controlling aspects of the testing environment. The applied nature of HF/E (even in basic research) leads investigators to simulate as much as they can in a study while maintaining control on extraneous factors (Meister, 2004).

Ideally, HF/E researchers and practitioners strive to generalize the results of investigations to a range of tasks, people, and contexts with confidence. Intuitively, it becomes necessary to apply methods to a range of tasks, people, and contexts to achieve this. Conflict often arises in terms of the available resources for the investigations (time and money). That said, the level of representation achieved through a method should be selected consciously, addressing the set goals, what is known about the method, and the practical limitations. Of course, the larger the sample size, the more confidence in results, but the larger sample usually entails higher financial and time investments. Furthermore, human limitations such as fatigue, attention span, and endurance may impose constraints on the amount of information to be gathered. It is also critical to consider the implications of method choice in terms of the analysis used. For example, investigations which collect and manipulate a large numbers of variables can take months to mine and analyze the data. Qualitative data or videos can take a significant amount of time to code for analysis. For this reason, the reader is encouraged to review Chapter 44 prior to using HF/E methods.

There is a point of diminishing returns when it comes to increasing the size of the sample or number of observations. In other words, the amount of certainty or knowledge gained from the additional observation may or may not be worth the time and effort spent in its collection and analysis. Sanders and McCormick (1993) introduced three factors that can influence sample size:

1. *Degree of Accuracy Required*. The greater the required accuracy, the larger the sample size required.
2. *Variance in the Sample Population*. The greater the variance, the larger the sample size required.
3. *Statistic To Be Estimated*. A greater number of samples are required to estimate the median than

the mean with the same degree of accuracy and certainty.

### 2.2.4 Incorporating Theory and Previous Work

The number of factors to consider in the selection of HF/E methodologies may seem an impossible task. However, method selection is greatly informed by theory as well as previous applications of the methodologies, as documented in the literature. The investigator needs to examine the existing knowledge base critically as well as talk with other HF/E investigators to gain practical insight. This serves as one of the best ways to justify the selection of methods for the problem at hand. In the selection of methods, only those that offer evidence of practicality and validity should be selected.

When conducting a critical examination of the literature, readers should be cognizant of the following factors, adapted from Weimer (1995):

- What are the authors' goals, both explicit and inferred from the text?
- What prior research do they reference and how do they interpret it?
- What are their hypotheses?
- How are the methods linked to the hypothesis?
- What are the variables (independent, dependent, and control) and, operationally, how are they defined?
- Are extraneous variables controlled and how?
- What are the relevant characteristics of the participant population?
- How did the authors recruit the participant population and how many people did they use?
- Was the research done in a laboratory or in the field?
- Did they use any special measurement equipment, technologies, surveys, or questionnaires?
- What statistical tests were run?
- What was the resulting statistical power?
- How do the authors interpret the results?
- How well do the authors' results fit with the existing knowledge base?
- Are there any conflicts in the interpretation of data between authors of different studies?

Investigators who are able to find literature relevant to their objective problem(s) can apply methods based on what others have applied successfully. However, because the subject matter and context of HF/E are so varied, care should be taken in this extrapolation. In the application of historically successful methods, the investigator must justify any deviation he or she made from the accepted status quo of the method. Basic research typically has the most to gain from such literature reviews.

Applied research is somewhat more problematic, as the methods are more diversified according to the conditions specified in the target application. Additionally, there are issues in the documentation of

applied methods (Committee on Human Factors, 1983). The historical memory of human factors methods resides largely in the heads and thick report files of the practitioners. Probing colleagues and other HF/E practitioners for their practical experience in using applied methods is therefore useful.

The survey of relevant literature is such a critical step in investigation that it can, by itself, serve as a complete method. A thorough literature search for theory and practice combined with discussions of methodological issues prevent investigators from reinventing the methodological approach. It may also save time through the avoidance of the common pitfalls in the execution of certain methods and analyses. In fact, literature reviews can substitute the need for the application of methods if the experimental questions have already been addressed. A *meta-analysis* is a specific method for the combination of statistical results from several studies (Wickens et al., 2003a).

Basic HF/E investigations, theories, and principles from psychology, physiology, and engineering all merit review. Theory is especially important because it can direct attention in complex systems as to where to focus resources. Knowledge of existing theories provides blueprints for the selection and application of methods as well as the explanation of results (Kantowitz, 1992). Theory is essential in the planning process because of the need to link the methodological processes and hypotheses strongly to the problem and goals of the investigation. However, when applying theories and principles, the investigator must ensure that the end results are in line with the theories employed.

The guidelines for critical examinations of the literature serve another role for the selection and application of HF/E methods. Investigators should realize that others will, one day, examine their work critically in a similar fashion. That said, consideration of these systematic evaluations in the design of methods can save undue hardship later during analysis and especially in the course of result interpretation. Similar to the design of systems, it is easier to make changes in the beginning steps of method formulation than to make changes to, or draw logical and meaningful conclusions from, ill-conceived outcomes of the method.

The decision of what methods are generally accepted by the community as standard and valid is difficult. The field of human factors, relative to more basic research, is much less grounded in terms of methodology. With the continued introduction of new technologies and systems, investigators are continually deriving new methods, metrics, and inferences. Despite this ongoing development, it is the investigator's responsibility to consider the issues of validity, reliability, and practical issues of a method before implementing it and reporting the ensuing results. Furthermore, in the report of results HF/E investigators need to clearly mark what is informed by theory and what, in actual fact, is their own speculation. Speculations are easily mistaken for fact when not labeled explicitly as such (Meister, 2004). This is true even when examples of a method exist in the literature, as the new method of application must be validated. What should be considered is *how* the

authors in the scientific literature justified a method (and those who do not justify should be looked at with some skepticism).

## 2.3 Working with Humans as Research Participants

Many HF/E methodologies require humans to serve as participants, providing data needed in the analysis of the system. As researchers (and humans beings ourselves), we are bound to the ethical handling of participants and their data. The foundation of ethical concerns is to ensure that the investigators do not sacrifice participants' general health, welfare, or well-being in lieu of achieving results for their research goals. Professional and federal agencies have assembled specific guidelines aimed at the appropriate treatment of people and their data in research and analysis funded by U.S. federal monies. The federal code of regulations for the protection of human subject (U.S. Department of Health and Human Services, 2009) (for investigators in the United States) and the American Psychological Association ethical guidelines for research with human participants (American Psychological Association, 2010) should be familiar to anyone conducting research with people as participants. Basically, these principles entail (1) guarding participants from mental or physical harm, (2) guarding participants' privacy with respect to their actions and behavior during the study, (3) ensuring that participation in the research is voluntary, and (4) allowing the participant the right to be informed about the nature of the experimental procedure and any potential risks (Wickens et al., 2003a).

Although the associated risk of HF/E investigations may seem minor, the rights of participants should not be taken lightly. Several historical events have informed the development of codes of conduct under which participants experienced undue mental and/or physical harm. Perhaps the most widely known is the Nuremberg Code, written by American judges in response to scientific experiments (mental and physical) in which prisoners were exposed to extreme medical and psychological tests in Nazi concentration camps. The Nuremberg Code was the first of its kind and mandates that the duties of those conducting research have the responsibility to protect the welfare of the participant ["Nuremberg Code (1947)," 1996].

Even after the Nuremberg Code, several instances of unethical treatment of human participants were documented. In 1964 the Declaration of Helsinki was developed to provide guidance to those conducting medical trials (World Medical Association, 2002). Finally, in 1979 the Belmont Report (Office for Human Research Protections, 1979) was released partly in response to inappropriately conducted U.S. human radiation experiments (U.S. Department of Energy, 2004). The three principles emergent from the Belmont Report were (1) *respect* in recognition of the personal dignity and autonomy of individuals and special protection for those with diminished autonomy, (2) *beneficence* by maximizing the anticipated benefits of participation and minimizing the risks of harm, and (3) *justice* in the fair

distribution of research benefits and burdens to participants (Office for Human Research Protections, 1979).

To aid researchers in ethical conduct, many institutions have an institutional review board (IRB) to provide guidance and approval for the use of human participants in research. Each protocol must be approved by the IRB before experimentation can begin. The IRB may review, approve, disapprove, or require changes in the research activities proposed. Approval is based on the disclosure of experimental details by the investigator(s). For example, many IRBs request the following information:

- Completion of educational training for research involving human participants by all persons (investigators and support staff) involved in the study
- A description of the research in lay language, including the scientific significance and goals:
  - Description of participant recruitment procedures (even copies of advertisements)
  - Inclusion/exclusion criteria for participant entry and justification for the specific exclusion of minorities, women, or minors
  - Highlights of the potential benefits and risks
  - Copies of all surveys and questionnaires
  - Vulnerable groups such as minors
- Funding of the research
- Location of the research
- How the data will be archived and secured to ensure participant privacy

In addition, researchers are instructed to create an informed consent form for the participants to sign. This document, approved by the IRB, explains the nature and risks of the study, noting voluntary participation and stating that withdrawal from the study is possible at any time without penalty.

Although the documentation and certification to ensure the welfare of participants may impose a lot of paperwork, these factors do have implications as to the quality of results in HF/E. The more comfortable participants are, the more likely they are to cooperate with the investigator during human–system investigations (and return for subsequent sessions). This contributes to the acceptability of a method by participants, one of the practical criteria to be used in method selection described in Section 2.2.

### 2.4   Next Steps in Method Selection

Operational methods are most commonly classified into three categories: (1) experimental studies, (2) descriptive studies, and (3) evaluative studies. The selection of methodology from one of these categories will lead the investigator through a series of directed choices, as depicted in Figure 1. These decisions include:

- What are the relevant variables?
  - How are they defined?
  - How are they captured?

- What will the actual measurements look like?
  - Is it qualitative?
  - Is it quantitative?
  - Is it a combination of both?
- What levels of experimental control and representativeness will be exercised?
  - Who are the participants?
  - Where will the study be conducted?
  - What equipment and measurement tools are needed?

The selection of method type, variables, measurements, and experimental control factors are much intertwined. There is no specific order to be followed in answering these questions except what is directed by the priorities established in the problem definition phase. The choices available in response to each question are limited, according to the method that is applied. Furthermore, the psychometric and practical issues introduced in this section must be verified routinely during the selection of specific plans for improved robustness of predictability and generalizability of the investigation outcomes.

In the remainder of this chapter we introduce the three different operational approaches. Specific examples of methodologies in each category are provided, along with answers to the questions outlined above and in Figure 1. The execution of each method will also serve as a point of discussion. Although the number of issues to consider in HF/E methods is sizable, the implications of improper use of methods can be far reaching. The careless application of HF/E methods may result in lost time, lost money, health detriments, discomfort, dissatisfaction, injury, stress, and loss of competitiveness (Wilson and Corlett, 2005).

## 3   TYPES OF METHODS AND APPROACHES

The taxonomy of HF/E methodologies is not straightforward, as there are areas that overlap within these defining characterizations (Meister, 2004). However, a classification enables guidance in methodology selection. There are several different classifications of methods in the literature, each author presenting the field in different scope, point of view, and even terminology. One of the more detailed and comprehensive taxonomies is that of Wilson and Corlett (2005). The authors classify methods as (1) general methods, (2) collection of information about people, (3) analysis and design, (4) evaluation of human–machine system performance, (5) evaluation of demands on people, and (6) management and implementation of ergonomics into group and subgroup. The authors then detail 35 groups of methods each with subgroup classifiers. Finally, the authors present techniques that are used in each method and common measures and outcomes.

Other authors, this handbook included, present a more simplified classification of methodological processes. Although the taxonomy presented by Wilson and Corlett (2005) has utility, that level of detail is beyond

the scope of this chapter. Instead, methods are broken down in a manner similar to those of Meister (1971), Sanders and McCormick (1993), and Wickens et al. (2003a).

Methodologies will be classified as *descriptive, experimental*, and *evaluation based*. Thus far, classifications of basic and applied research goals and attributes that methods can possess in terms of validity, reliability, and objectiveness have been covered. Each of the three classes of research best serves a different goal while directing the selection of research setting, variables, and participants (to meet the demands of validity, reliability, and objectiveness). Although some overlap exists between descriptive, experimental, and evaluation-based methods, HF/E research can usefully be classified into one of the three (Sanders and McCormick, 1993).

## 3.1 Descriptive Methods

Descriptive methods assign certain attributes to features, events, and conditions in an attempt to identify the variables present and their values in order to characterize a specific population and sometimes determine the relationships that exist (Sanders and McCormick, 1993; Gould, 2002). Descriptive methods do not involve the manipulation of an independent variable but instead focus on nonexperimental strategies (Smith and Davis, 2008). The investigator is typically interested in describing a population in terms of attributes, identifying any possible parallels between attributes (or variables). The variables of interest encompass who, what, when, where, and how. The objective of descriptive research is to obtain a "snapshot" of the status of an attribute or phenomenon. The results of descriptive methods do not provide causal explanation of attributes. Correlation is the only relationship between variables that can be determined unless the specific attributes of relationships are captured.

There are three primary rationales for descriptive research: no alternative exists to natural observation, unethical behavior would be involved if certain factors were manipulated, and finally it is advantageous in the early stages of research to conduct descriptive research prior to experimental manipulation (Gould, 2002). The third reason demonstrates the utility of descriptive research, in that it provides a basis for conducting additional, more specific investigations. Descriptive methods are identified by the characterization of system states, populations, or interactions in its most natural form, without manipulation of conditions (as in the case of empirical methods). The results of descriptive research methods often serve as motivation for experimental or evaluative research. Furthermore, assumptions of populations, environments, and systems underlie just about any research. These assumptions may be implicit or explicit, well founded, and in some cases unfounded. Descriptive methodologies clear up assumptions by providing investigators with an improved characterization of the target population, environment, or system.

Descriptive studies may be cross-sectional or longitudinal. *Cross-sectional descriptive studies* take a one-time snapshot of the attributes of interest. The collection of anthropometric data from schoolchildren and dimensions of their school furniture was a cross-sectional assessment of these two attributes (Agha, 2010). The majority of available anthropometric data in the scientific knowledge base is, in fact, cross-sectional, representative of a single population (usually military) at one point in time (the 1950s).

*Longitudinal studies* follow a sample population over time and track changes in the attributes of that population. A longitudinal study asks the same question or involves observations at two or more times. There are four different types of longitudinal studies, dictated by the type of sampling used in the repeated methodology (Menard, 2002):

1. *Trend Studies*. The same inquiries are made to different samples of the target population over time.
2. *Cohort Studies*. Track changes in individuals with membership in an identified group that experiences similar life events (e.g., organizational, geographical groups) over time.
3. *Panel*. The same inquiries are made to the same people over time.
4. *Follow-Up*. Inquiries are made to the participants after a significant amount of time has passed.

### 3.1.1 Variables

Descriptive studies ascribe values to characteristics, behaviors, or events of interest in a human-integrated system. The variables captured can be qualitative (such as a person's perceived comfort) and/or quantitative (such as the number of female employees). These variables sort out into two classes: (1) criterion variables and (2) stratification variables. Criterion variables summarize characteristic behaviors and events of interest for a given group (such as the number of lost-time accidents for a given shift). Stratification variables are predictive variables that are aimed at the segmentation of the population into subgroups (e.g., age, gender, and experience).

### 3.1.2 Key Concern: Sampling

As noted by the classification of longitudinal descriptive studies, the approach to selecting participants is a critical factor in descriptive studies. The plan used in sampling or acquiring data points directs the overall validity of the method. To establish a highly representative sample, the investigator can try to ensure equal probability for the inclusion of each member of a population in a study through random sampling of the target population. However, this is not always feasible to do, as monetary and time constraints sometimes compel investigators to "take what they can get" in terms of participants. Still, if sampling bias has occurred, it can skew data analysis and suggest inferences that lack validity and reliability.

The solution is to review prior research, theories, and their experience to estimate the potential impact of bias factors on the variables of interest. Keep in mind that both analyst and participant bias can adversely affect reliability and validity of a study (Stanton et al., 2005).

A classic example of sampling bias occurs in telephone-administered surveys. This method neglects the proportion of the population whose socioeconomic status does not afford a home telephone. This can translate into bias in the variables gathered. Investigators need to weigh the potential impact on this "participant misrepresentation" in the potential confounding of their data.

Common types of bias issues include the following (Arleck and Settle, 1995):

- *Visibility Bias*. Bias results when some units of a population are more visible than others (e.g., the telephone example provided above).

- *Order Bias.* This occurs when the log of potential participants is in a specific order, such as birth dates or alphabetical order.

- *Accessibility Bias.* When measures are collected in the field, certain persons in the population are more accessible than others (e.g., teachers vs. the administrative staff).

- *Cluster Bias*. When a method targets clusters of participants from the sample frame, some clusters may be interrelated such that they share similar opinions, experiences, and values (e.g., workers from the third shift of a manufacturing operation).

- *Affinity Bias.* Usually a problem in fieldwork, the investigator may be more likely to select people based on extraneous physical and personality traits (e.g., approaching only those who seem to be friendly and cooperative).

- *Self-Selection Bias*. Persons in the population can, by choice, elect to participate in the descriptive methodology (e.g., people who respond to customer feedback surveys are only those who have a complaint).

- *Nonresponse Bias*. Typically associated with mail or e-mail surveys, those who elect not to respond could do so at random or due to some feature of the survey (e.g., the amount of personal information requested was too intrusive for some participants).

### 3.1.3   Techniques Employed

Observational techniques, surveys, and questionnaires are techniques most commonly associated with descriptive research methods. Descriptive methods may collect data in the field or laboratory or through survey methods. Participants must be recruited from the real world for representation sake, but the actual methods may be carried out in the laboratory (Sanders and McCormick, 1993). Typically, methods are conducted in a laboratory when the measurement equipment is too difficult to transport to a participant. This is often the case for anthropometric studies.

*Surveys, questionnaires*, and *interviews* embody the second class of methods used most often in descriptive studies. They are information-gathering tools to characterize user and system features. The data collected with the surveys can be qualitative, from open-ended

response questions, or employ quantitative scales. Surveys and questionnaires are very challenging to design with the assurance of valid, reliable results (Wickens et al., 2003a). They are susceptible to bias attributable to the investigator's wording and administration of questions as well as to the subjective opinions of those being questioned.

Survey and questionnaire design and administration are topics complicated enough for an entire handbook of its own. In fact, for more detailed explanations of questionnaire and survey design, readers are encouraged to review texts such as *The Survey Research Handbook* (Arleck and Settle, 1995) or *Survey Methodology* (Groves et al., 2009). In the scope of this chapter, the advantages and common pitfalls of surveys and questionnaires will be introduced. Interviews can be considered similar to questionnaires and surveys because they share the element of question and response (Meister, 2004), with the exception of aural administration (in most cases). Interviews are characterized by their ability to be conducted with more than one investigator or respondent and the range of formality they may take on. Interviews typically take more time, and for this reason, questionnaires are often used in lieu of interviews.

Interviews and questionnaires are useful for their ability to extract the respondent's perceptions of the system and their performance and behaviors for descriptive studies. That said, the construct validity of participant responses and both the inter- and intrarater reliability of responses are difficult to validate and confirm before analysis of the data collected. Case Study 2 provides examples of surveys and interviews employed in descriptive research.

### CASE STUDY 2: Computer use among older adults in a naturally occurring retirement community

***Goal***   The goal of this study by Carpenter and Buday (2007) was to examine patterns of computer use and barriers to computer use for older adults living in a naturally occurring retirement community (NORC).

***Methods***   Adults over age 65 in the NORC were invited to participate in an interview regarding their current service needs, recent service use, and preferences. Door-to-door solicitation, media announcements, booths at health fairs and local grocery stores, presentations to community groups, and direct mailings were used to recruit participants.[*] The participants were subsequently screened by phone to ensure that they met the residency and age requirements. Subsequent in-home interviews lasted approximately 2 h. The interview contained questions about demographics, social impairment, physical and mental health, and cognitive impairment. Questions were also asked regarding computer use, frequency of use, purpose, and barriers to computer usage. They interviewed a total of 324 older adults, of which 115 were computer users and 209 were nonusers.

[*] For an example of surveys through electronic media, see Rau and Salvendy (2001)

***Analyses***   Various analytical techniques were used, such as paired *t*-test and hierarchical regression analysis. Particularly, hierarchical regression analysis was used to identify characteristics that differentiated computer users from nonusers.

Methodological Implications:

*More Expensive to Conduct*. Because the interview had to take place in each participant's home, this method is more costly than a survey in terms of time and money.

*Benefits of Interviews:* By utilizing trained interviewers, consistency is likely. Unlike a survey, which can be distributed using various means, the structured interview format remains consistent. The interview format also allows for individuals who both use the computer and do not use computers to be sampled, wheras an electronic survey would miss the portion of the population that does not use a computer.

*Judicious Design of Interview Content*. Authors of this study identified two different types of participants: those who use computers and those who do not. Despite the two different categories, the interviewer asked both sets of individuals what prevents them from using the computer or from using it more than they currently do.

---

The selection of which questions to ask; how, procedurally, the questions are presented; and the responses collected are therefore critical in a situation so prone to bias. The responses given are dependent on the participants' psychological abilities, especially their memory (Meister, 2004). Respondents may interpret words and phrases in ways that may produce invalid responses. They may also have trouble in the affirmation and expression of their internal states. Finally, the level of control exercised over responses is important. Structured responses, such as multiple-response questions, can pigeonhole respondents into an ill-fit self-categorization. An alternative is the use of free-response answers and less formal interviews, which create data that are highly subjective and qualitative, making analysis and comparisons difficult. Specific scaling techniques for self-reported measures are detailed in Chapter 44.

*Observation techniques* consist of an investigator sensing, recording, and interpreting behavior. They capture covert behavior with clarity but demonstrate difficulty in estimates of tacit behavior. Meister (2004) asserts: "Observation is the most 'natural' of the measurement methods, because it taps a human function that is instinctive" (p. 131). Observations may be casual and undirected or direct with a definite measurement goal and highly structured. In planning observational methods, Wickens et al. (2003a) suggest that the investigator predefine the variables of interest, the observational methods, how each variable will be recorded, the specific conditions that afford a specific observation, and the observational time frame. These observational categories form a taxonomy of specific pieces of information

to be collected during the course of the observation. Furthermore, defining observational scenarios can enable the investigator to sample only at times when events are occurring that are relevant to the study's goals. This prevents the investigator from filtering through extraneous events and data related with the situation. Observational methods often use some data-recording equipment to better enable the investigator to return to specific events, code data postobservation, and archive raw data for future investigations. Commercial software programs are available that enable the investigator to flag certain events in the video stream for frequency counts and to return for a closer observation.

Observation-based methods are appropriate when working under constraints that limit contact with the participant or interference with the task, such as observing a team of surgeons in an operating room. There are also times when observations are useful because the population cannot express their experience accurately in alternative terms. This is especially the case when working with children, who are sometimes unable to use written surveys or respond to questionnaires. As with most descriptive methods, observation is useful in conceptual research as the precursor to empirical or evaluative research.

Some important factors in the implementation of observation include the amount of training required of observers, inter- and intraobserver reliability of the recorded data, the intrusiveness of the observation on the situation of interest, and how directly observable the variable of interest is (e.g., caller frustration is more difficult to measure than the frequency of calls someone makes to a technical support center). Table A1 in the appendix to this chapter provides the reader with real-world examples of descriptive studies that have been published in the past five years.

## 3.2   Empirical Methods

*Empirical research methods*, also known as *experimental methods*, assess whether relationships between system, performance, and human measures are due to random error or there is a causal relationship. The question in empirical research is: "If *x* is changed, what will happen to *y*?" at different levels of complexity. In empirical research the investigator typically manipulates one or more variables to appraise the effects on human, performance, or system criteria. The investigator manipulates the system directly to invoke an observable change (Drury, 2005).

Empirical methods are beneficial because the manipulations of variables enable the observation of circumstances that may occur infrequently in the operational (i.e., real) world. What's more, this manipulated situation allows for the application of more robust measurement approaches by removing the negative consequences of employing invasive implications to safety- or time-critical situations. The ability to exercise control in the situation to reduce variability may also provide the advantage of smaller sample sizes.

However, these benefits are not without cost. Drury (2005) asserts that face validity is sacrificed in the use of empirical methods; much more persuasion is necessary

for acceptance of the studies. Sanders and McCormick (1993) state that increases in precision, control, and replication are coupled with a loss of generalizability. It is then difficult to make an argument for applicability when dealing with theoretical questions. For this reason, many HF/E practitioners take their inferences and theories developed in highly controlled empirical research and confirm them via field-based descriptive and evaluation-based studies (Meister, 2004). Representation and validity of results are often the most problematic concerns with conducting empirical methods.

### 3.2.1  Variables

For investigators to hypothesize potential relationships between human and system components, they must select variables. *Independent variables* are those factors that are manipulated or controlled by the investigator and are expected to illicit some change in system and/or human behavior in an observable way. Independent variables can be classified as task related, environmental, or participant related and occur at more than one level. *Dependent variables* are measures of the change imposed by the independent variable(s). Extraneous variables are those factors that are not relevant to the hypotheses but that may influence the dependent variable. If extraneous variables are not controlled, their effect on the dependent variable could confound the observed changes triggered by the dependent variable.

Dependent variables are much like the criterion variables used in descriptive studies, with the exception that physical traits such as height, weight, and age are uncommon. Of course, the independent and dependent variables should be linked back to the hypotheses and goals. The best approach, when possible, is the assessment of human behavior in terms of performance, physiological, and subjective dependent measurements to tap accurately into the construct of interest. The goal of empirical research is to detect variance in the dependent variables triggered by different levels of the independent variable(s). Variable selection and definition play a key part in the structuring of an experimental plan.

### 3.2.2  Selecting Participants

While descriptive methods typically require sampling from an actual population, empirical research directs the investigator to select participants who are representative of those in the target population. Certain traits of the population are more important than others, depending on the task and the physical and mental traits exhibited by the target population. The HF/E investigator needs to seriously contemplate if the participant population will be influenced by the independent variable in the same ways as the target population and which factors are extraneous. To determine this, a review of previous theory and literature is once again valuable. Of additional value in narrowing the scope of participant characteristics are descriptive studies: observations, interviews, and questionnaires. These studies can characterize the target population and help an investigator to incorporate the necessary subjective features.

In some circumstances, members of the target population are so highly skilled and trained in their behavior and activities that it is difficult to match participants of similar skill levels. The investigator may find that they can circumvent this by training. Learning is typically estimated by an exponential model; there is an asymptotic point where there is little improvement in the knowledge or skill acquired (Gawron, 2008). The investigator may train participants to a certain point so that their interactions can match more closely those of the target population. Training can be provided to the participants through a specific regimen (e.g., subjects are exposed to three practice trials), they may be trained until they ascertain a specific performance level (e.g., accuracy, time to complete), or the training can be self-directed (i.e., the participant trains until he or she has attained a self-perceived comfort level). Circumstances exist where the amount of time and money to train is prohibitive or training for the construct of interest is simply unrealistic. For example, in studies that employ flight simulators, it is not feasible to train a group of undergraduates to the same level as that of rated pilots although they can be compared to beginning pilots who have not had any flight training (e.g., Pritchett, 2002; Donderi et al., 2010).

Another issue that surfaces in conjunction with the selection of participants is interparticipant variability. This could be age, experience, formal training, or skill. When variability is attributable to differences in knowledge and skill level, the same approach can be taken as mentioned above to train participants to a certain skill level. Variability among participants can create confounding variables for the analysis and interpretation of the data. If this variability is indicative of the actual population (and is desired), the investigator can take specific measures in assigning participants to the experimental conditions.

### 3.2.3  Key Concern: Experimental Plan

The experimental plan is the blueprint for empirical research. It outlines in detail how the experiment will be implemented (Wickens et al., 2003a). The key components of the experimental plan include:

- Defining variables in quantifiable terms in order to determine:
  - The experimental task
  - The levels of manipulation for the independent variable (e.g., the experimental conditions)
  - Which aspects of the behavior to measure: the dependent variable
  - The strategy for controlling confounding variables
  - The type of equipment used for data collection (e.g., pencil and paper, video, computer, eye-tracker)
  - The types of analytical methods that can be applied (e.g., parametric vs. nonparametric statistical)

- Specification of the experimental design in order to determine:
  - Which participants will be exposed to different experimental conditions
  - Order of exposure to treatments
  - How many replications will be used
  - The number of participants required and recruitment methods (e.g., statistical power)

Experimental designs represent (1) different methods for describing variation in treatment conditions, (2) the assignment of participants to those conditions, and (3) the order in which participants are exposed to treatments (Williges, 1995; Meister, 2004). The basic concept of experimental design is discussed here, but for a more thorough discussion the reader is encouraged to review Williges (1995). A more statistically based account may be found in Box et al. (1978). The assignment of participants to treatment conditions is accomplished by means of two-group designs, multiple-group designs, factorial designs, within-subject designs, and between-subject designs. Two-group, multiple-group, and factorial designs describe ways in which the independent variable(s) of interest are broken down into quantifiable, determinant units. Between- and within-subject designs detail how the levels are assigned to the participants. Following is a brief description for each type of design and the conditions that are best supported by each design [Wickens et al. (2003a), compiled from Williges (1995)].

1. *Two-Group Design*. An evaluation is conducted using one independent variable with two conditions or treatment levels. The dependent variable is compared between the two conditions. Sometimes there is a *control condition*, in which no treatment is given. Thus the two levels are the presence or absence of treatment.

2. *Multiple-Group Design*. One independent variable is specified at more than two levels to gain more information (often, more diagnostic) on the impact of the independent variable.

3. *Factorial Design*. An evaluation of two or more independent variables is conducted so that all possible combinations of the variables are evaluated to assess the effect of each variable in isolation and in interaction.

4. *Between-Subject Design*. Each experimental condition is given to a unique group of participants, and participants experience only one condition. This is used widely when it is problematic to expose participants to more than one condition and if time is an issue (e.g., fatigue, learning, order effects).

5. *Within-Subject Design*. Each participant is exposed to every experimental condition. This is called *repeated-measure design* because each participant is observed more than once. It typically reduces the number of participants required.

6. *Mixed-Subject Design*. Variables are explored within and between subjects.

Factorial designs are the most comprehensive type of experimental design. Furthermore, this design enables the variation of more than one system attribute during a single experiment; it is more representative of the real-world complexity of keeping track of the interactions between factors. Factorial designs are common in HF/E empirical work. Terminology is used to explain factorial designs quickly in the reporting of results. For example, if an empirical result has employed four independent variables, the design would be described as a four-way factorial design. If each of three of the independent variables has two levels, and the fourth has three levels, these levels are disclosed by describing the experiment as using a $2 \times 2 \times 2 \times 3$ factorial design.

Between-subject designs are highly susceptible to variation between groups on extraneous factors. This variation can impose constraints on the interpretation and generalizability of the results and can impose the risk of concluding a difference between the two groups based on the independent variable when in fact it is the other intragroup variation. The randomized allocation of participants to groups does not ensure the absence of intergroup variation on factors such as education, gender, age, and experience. These factors are identified through experience, preliminary research, or literature reviews. If extraneous factors have a potential influence on the dependent variable, the investigators should do their best to distribute the variation among the experimental groups. Randomized blocking is a two-step process of separating participants based on the intervening factors; an *equal* number of participants from each block are randomly assigned a condition.

Within-subject designs are prone to order effects. That is, participants might exhibit different behaviors depending on the sequence in which they are exposed to the conditions. Participants may exhibit improved performance over consecutive trials due to learning effects or degraded performance over the consecutive trials due to fatigue or boredom. Unfortunately, fatigue and learning effects do not tend to balance each other out. In terms of fatigue, the investigator may offer the participants rest breaks between sessions or schedule several individual sessions with each participant over time (Gawron, 2008). Learning effects can be mitigated if the participants are trained to a specified point using the techniques mentioned previously.

Investigators may also use specific strategies for assigning the order of conditions to participants. If each condition is run at a different place in the sequence among participants, the potential learning effects may be averaged out; this is called *counterbalancing* (Wickens et al., 2003a). This can be accomplished through randomization, which requires a large number of participants to be effective (imbalance of assignment is likely with a smaller sample). Alternatively, there are structured randomization techniques, which ensure that each "sequence" is experienced. However, to mitigate the learning impact effectively, the number of participants needs to be a multiplier of the number of sequences (which is difficult in studies with many variables), and this may be implausible, depending on the constraints of the study.

Carryover effects are also a possible effect in with-subject designs if conditions are consecutively run repeatedly. Say that in an experimental design there are four conditions, $a$, $b$, $c$, and $d$, and the following orders have been determined for the experimental runs of four participants:

- Order 1: $a-b-c-d$
- Order 2: $b-c-d-a$
- Order 3: $c-d-a-b$
- Order 4: $d-a-b-c$

Note that in these four orders each condition is in a sequentially unique position each time. However, condition $a$ always precedes $b$, $b$ always precedes $c$, and $c$ always precedes $d$. Features of one condition could potentially influence changes in the participants' behaviors under subsequent conditions. As an example, consider an empirical study that strives to understand the visual search strategies employed by quality control inspectors in the detection of errors under a variety of environmental conditions. If one condition is more challenging and it takes an inspector longer to find an error, the inspectors might well change their visual search strategies in reaction, based on the prior difficulties. An investigator may therefore have to employ a combination of random assignment and structured assignment of conditions.

Empirical research methods are typically conducted in the lab but can be gathered in the field as well. The field offers investigators higher representation, but their control of independent and extraneous variables diminishes significantly. The advantages of working in a laboratory setting include the high level of control an investigator can exercise in the specification of independent variable levels and the blocking of potential confounding variables.

The importance in empirical research of running a pilot study cannot be understated. This provides the investigator with a preview of potential issues with equipment, participants, and even the analysis of data. Even with a thorough experimental plan, investigators can encounter unplanned sources of variability in data or unknown confounding variables. This "practice run" can help an investigator to circumvent such problems when collecting actual data. The potential sunk cost of experimental trials that yield contaminated data drives the need for pilot studies.

Empirical investigations possess many advantages in terms of isolating the construct of interest, but the amount of control applied to the empirical setting can drastically limit the generalizability of the results. Empirical research is typically more *basic* in nature, for it drives the understanding of principles and theories which can then be applied to (and validated by) real-world systems. Case Study 3 provides readers with a review of one empirical investigation using a mixed factorial design. In addition, Table A2 provides several more examples of contemporary empirical work in HF/E.

## CASE STUDY 3: Multimodal Feedback Assessment of Performance and Mental Workload

***Goal*** The goal of this study by Vitense et al. (2003) was to establish recommendations for multimodal interfaces using auditory, haptic, and visual feedback.

***Methods*** To extract and assess the complexity of HCI with multimodal feedback in a quantifiable way, the authors conducted a highly controlled empirical study. Thirty-two participants were selected carefully in order to control extraneous factors and to meet hardware requirements. These inclusion criteria were right-handedness, normal visual acuity, and near-normal hearing capability. Appropriate software and hardware were developed and purchased to generate the multimodal feedback to match both the real world and research published previously.

To investigate three different modalities and all possible combinations of the modalities, this study used a $2 \times 2 \times 2$ factorial, within-subject design. Participants used a computer to perform drag-and-drop tasks while being exposed to various combinations of multimodal feedback. Training sessions were conducted to familiarize participants with the experimental tasks, equipment, and each feedback condition. NASA-TLX and time measurement were employed to capture the workload and task performance of participants quantitatively.

***Analyses*** A general linear model repeated-measures analysis was run to analyze the various performance measures and the workload. Interaction plots were also used to present and explain some significant interaction among visual, auditory, and haptic feedback.

Methodological Implications:

*A small number of observations is required*. By controlling uninteresting factors from an experiment, unnecessary variability can be decreased. Thus, as you can see in this case study, empirical studies generally employ smaller numbers of participants than do other types of studies (e.g., descriptive).

*Factors are difficult to control*. Controlling extraneous factors is not an easy task. As this case study shows, careful selection of participants and training were necessary to reduce contaminant variability.

*A covert, dynamic HF/E phenomenon is easier to capture*. Human subjects are easily affected by various extraneous factors, making isolated appraisal of the construct difficult. In this example, the authors conducted a highly controlled experiment in an attempt to extract subtle differences in the interactions among feedback conditions.

## 3.3 Evaluation Methods

Evaluation methods are probably the most difficult to classify because they embody features of both descriptive and empirical studies. Many of the techniques and

tools used in evaluation methods overlap with descriptive and empirical methods. Evaluation methods are chosen specifically because the objectives mandate the evaluation of a design or product, the evaluation of competing designs or products, or even the evaluation of methodologies or measurement tools. The goals of evaluation methods also match closely both descriptive and field methods, but with more of an applied flavor. Evaluation methods are a critical part of system designs, and the specific evaluation methodology used depends on the stage of the design. These methods are highly *applied* in nature, as they typically reference real-world systems.

The purpose of evaluation research embodies (1) understanding the effect of interactions for system or product use (akin to empirical research), (2) descriptions of people using the system (akin to descriptive research), and (3) assessment of the outcomes of system or product use compared to the system or product goal (akin to descriptive research), to confirm intended and unintended outcomes of use (unique to evaluation methods).

Evaluation research is part of the design process. Evaluations assess the integrity of a design and make recommendations for iterative improvements. Therefore, they can be used at a number of points during the design process. The stage of the product or system, including concept, design, prototype, and operational products, is the authority in mandating which techniques to use. Stanton and Young (1999) usefully categorized 12 evaluation methods according to applicability to the various product stages. Table 4 presents a summary of their classification. It is interesting to note that the further along the design process is, the greater the number of applicable techniques. The ease with which methods are applied is therefore a function of the abstraction in the design process. Those products with a physical presence or systems that are tangible are compatible with a wider variety of methodological techniques. This does not imply greater importance in using evaluation methods at later design stages. In fact, evaluations can have the greatest impact in the conceptual stage of product design, when designers express the most flexibility and acceptance of change.

Evaluation methods typically have significant constraints placed on their resources in terms of time, money, and staff. Therefore, these factors, combined with the goal of the evaluation, direct the selection of methods. The relevant questions to consider when selecting a method for evaluation include:

- Resource-specific criteria
  - What is the cost−benefit ratio for using this method?
  - How much time is available for the study?
  - How much money is available for the study?
  - How many staff members are available for implementation and analysis of the study?
  - How can designers be involved in the evaluation?
- Method-specific criteria

- What is the purpose of the evaluation?
- What is the state of the product or system?
- What will the outcome of the evaluation be? (e.g., a report, presentation, design selection)

Evaluation methods typically serve three roles: (1) functional analysis, (2) scenario analysis, and (3) structural analysis (Stanton and Young, 1999). *Functional analyses* seek to understand the scope of functions that a product or system supports. *Scenario analyses* seek to evaluate the actual sequence of activities that users of the system must step through to achieve the desired outcome. *Structural analysis* is the deconstruction of the design from a user's perspective. The selection of variables for evaluation research methods is influenced largely by the same factors that influence variable selection in both descriptive and empirical studies. Quantifiable, objective criteria of system and human performance are most useful in making comparisons of competing designs, systems, or products.

### 3.3.1 Key Concern: Representation

The research setting, tasks, and participants need to be as close to the real world as possible. A lack of generalizability of evaluation research to the actual design, users, tasks, and environment would mean significant gaps in the inferences and recommendations to be made. Sampling of participants should follow those guidelines outlined previously for descriptive studies.

The research setting should be selected based on the constraints listed above. The research needs to ask: "Do you gain more from watching the interactions in context than what you lose from lack of control (Woods, 1996)?" In evaluation studies, field research can provide an in-depth understanding of the goals, needs, and activities of users. But pure field methods such as ethnographic interviews create extensive challenges in terms

**Table 4 Assessment Techniques in the Product Design Process**

| Product Phase | Assessment Techniques |
| --- | --- |
| Concept | *5/12 methods applicable:* checklists, hierarchical task analysis (HTA), repertory grids, interviews, heuristics |
| Design | *10/12 methods applicable:* KLM, link analysis, checklists, protective human error analysis (PHEA), HTA, repertory grids, task analysis for error identification (TAFEI), layout analysis, interviews, heuristics |
| Prototype | *12/12 techniques applicable:* KLM, link analysis, checklists, PHEA, observation, questionnaires, HTA, repertory grids, TAFEI, layout analysis, interviews, heuristics |
| Operational | *12/12 techniques applicable:* KLM, link analysis, checklists, PHEA, observation, questionnaires, HTA, repertory grids, TAFEI, layout analysis, interviews, heuristics |

of budgets, scheduling, and logistics (Woods, 1996). Evaluation methods often succumb to constraints of time, financial support, and expectations as to outcomes. That said, investigators must leverage their resources to best meet those expectations. Practitioners must adopt creative techniques to deal with low-fidelity prototypes (and sometimes no prototype) and limited population samples. The prioritization of methodological decisions must be clearly aligned with the goals and expectations of the study.

Case Study 4 provides readers with an example of one evaluation study. Additionally, Table A3 provides several more examples of evaluation studies.

### CASE STUDY 4: Fleet Study Evaluation of an Advance Brake Warning System

**Goal**   The goal of this study by Shinar (2000) is to evaluate the effectiveness of an advance brake warning system (ABWS) under true driving conditions.

**Methods**   This case study is one of several evaluation studies of the ABWS. Prior to study execution, a simulation study proved the ABWS effective in decreasing the possibility of rear-end crashes (from 73 to 18%). However, the assumptions made in developing the simulation caused limitations in the applicability of its results and conclusions to real-world situations. The inadequacies in the previous study motivated this longitudinal field study investigating 764 government vehicles. Half of the vehicles were equipped with an ABWS, and the other vehicles were without an ABWS. Over four years, the 764 vehicles were used as government fleet vehicles and all crashes involving the vehicles were tracked. This tracking process was carried out unbeknownst to the vehicle drivers. Because the accidents happened for a variety of reasons, it was difficult to distinguish whether a collision was relevant or not. Although the assessment of causality could not be objective, judgments by the investigator were made conservatively in order to improve the validity and integrity of study results.

**Analyses**   A paired $t$-test was used to detect a statistically significant difference in the number of accidents between the two automobile systems. Because the data were gathered under real circumstances, some uncontrolled factors potentially confounded the results. For example, the average distance driven by the control group was different from that of the treatment group.

Those factors were accounted for by introducing more diagnostic evaluative measures, such as the number of rear-end collisions per kilometer for a specific region.

Methodological Implications:

*Specific to a Certain Design or System*. One salient characteristic of evaluation studies is that they target a specific design or system. In this case study the target system is an ABWS.

*Emphasizing Representation*. The major reason that the author was not satisfied with the simulation method was that it lacked representation of the real world, due to some assumptions. Thus, the second study was conducted using actual operating vehicles. Often, investigators must do both for a more complete understanding of the HF/E related to the systems of interest.

*Lacking Control*. Representation is not obtained without cost. Under real situations, experimenters cannot control extraneous, confounding factors. As a result, this author had difficulty distinguishing relevant crashes from irrelevant ones. To compensate, the experiment took place over the course of four years, which increased the sample size to a more acceptable level for analysis.

## 4   CONCLUSIONS

The selection and application of HF/E methods are part art, part science. There is a certain creative skill for the effective application of HF/E methods. Furthermore, that creative skill is acquired through practice and experience. HF/E investigators must be knowledgeable in several areas, be able to interpret theories and principles of other sciences, and integrate them with their own knowledge and creativity in valid, reliable ways to meet the investigation's goals. Of course, all this is to be accomplished within the constraints of time and resources encountered by researchers and practitioners. An awareness of HF/E methods—their limitations, strengths, and prior uses—provides an investigator with a valuable toolkit of knowledge. This and practical experience lend the investigator the ability to delve into the complex phenomena associated with HF/E.

## APPENDIX: EXEMPLARY STUDIES OF HF/E METHODOLOGIES

**Table A1  Examples of Descriptive Studies**

| Study | HF/E Subdiscipline | Goal | Methodology | Analysis Methods |
|---|---|---|---|---|
| School furniture match to students' anthropometry in the Gaza Strip (Agha, 2010) | Anthropometry | To compare primary school students' anthropometry to the dimensions of school furniture whether the furniture used matches the students' anthropometry | *Field study:* Measured anthropometric data from randomly selected 600 male schoolchildren from five schools were used to assess ratio measures of stature | Stature categorized in quartiles; derivative ratios calculated to obtain a range of anthropometric proportions |

**Table A1** (*continued*)

| Study | HF/E Subdiscipline | Goal | Methodology | Analysis Methods |
|---|---|---|---|---|
| Comparing safety climate in naval aviation and hospitals: implications for improving patient safety (Singer et al., 2010) | Safety | To compare results of safety climate survey questions from health care respondents with those from naval aviation, a high-reliability organization | *Survey:* Designed to include comparable items were used — PSCHO survey for hospital personnel and command safety assessment (CSAS) for naval aviators. They received 13,841 completed surveys in U.S. hospitals, 5,511 in VA hospitals, and 14,854 among naval aviators for a total of 34,206 individuals. | Calculated percent of problematic response (PRR) for each survey item and for safety climate overall on average for U.S. hospitals, VA hospitals, and naval aviators. Significance tests were performed for all comparisons of categorical data with $p < 0.001$ for all but one question which used $p < 0.0445$. |
| Effects of driver fatigue monitoring — an expert survey (Karrer and Roetting, 2007) | Fatigue | To evaluate how experts perceive future driver fatigue monitoring; identify objectives and predicted effects of these systems | *Survey:* The survey was distributed to two expert groups: researchers working in the field of driver fatigue monitoring and professional drivers in order to examine differences between the expert groups. Questions were asked to rank the objective of driver fatigue monitoring. On the second portion of the survey they were asked to rank whether or not they agreed or disagreed with possible positive and negative outcomes of driver fatigue monitoring while taking into consideration three different types of automated feedback types. Based on the criteria they set for their ''experts,'' they used 19 researcher surveys and 52 surveys from professional drivers. | ANOVA |
| Anthropometric differences among occupational groups (Hsiao et al., 2002) | Anthropometry | To identify differences in various body measurements between occupational groups in the United States | Archival data collected in the third National Health and Nutrition Examination Survey (NHANES III 1988–1994) were analyzed. | Two-tailed *t*-test |
| Comprehending product warning information: age-related effects and roles of memory, inferencing, and knowledge (Hancock et al., 2005) | Aging, warning perception | To show age-effect comprehension of warnings by comparing younger (18–23 years) and older (65–75 years) adults | The first experiment measured younger and older adults' comprehension of real-world warnings through a verification test presented either immediately after reading the warnings or after a delay. In the second experiment younger and older adults read fabricated warnings that were inconsistent with real-world knowledge. There were 52 younger and 47 older participants. | Univariate ANOVA |

**Table A1** (*continued*)

| Study | HF/E Subdiscipline | Goal | Methodology | Analysis Methods |
|---|---|---|---|---|
| Ergonomics of electronic mail address systems: related literature review and survey of users (Rau and Salvendy, 2001) | E-mail, user-centered design | To obtain information on preferences, dislikes, and difficulties associated with the e-mail address system | *Survey*: Conducted through e-mail and a newsgroup. Seventy questions were administered regarding respondents' use of and attitude toward their electronic mail systems. 160 electronic questionnaires were returned. | Analysis of correlations |
| Effects of task complexity and experience on learning and forgetting: a field study (Nembhard, 2000) | Cognitive processes (learning, memory) | To examine the effects of task complexity and experience on parameters of individual learning and forgetting | *Longitudinal and field study*: 2853 learning and forgetting episodes were captured over the course of a year. The episodes captured and averaged performance data, task complexity, and task experience for each employee on a weekly basis. | Mathematical model development; Kolmogorov–Smirnov; ANOVA; pairwise comparison; regression analysis |
| Computer use among older adults in naturally occurring retirement community (Carpenter and Buday, 2007) | HCI, aging | To examine patterns of computer use and barriers to use among older adults | *Field interviews*: Demographic data were collected and social impairment was obtained with the Older American Resources and Services Assessment Questionnaire (OARS) Social Resources Rating Scale. Physical health was assessed using the Cornell Medical Index Health Questionnaire and items from the Duke Older Americans Resources and Services survey. Cognitive impairment was measured using the blessed orientation–memory–concentration test (BOMC). Mental health was assessed using the short form of the Geriatric Depression Scale. Questions were also asked regarding computer use, frequency of use, purpose, and barriers to computer usage. | *t*-Test; regression analysis; |
| Case study of co-ordinative decision making in disaster management (Smith and Dowell, 2000) | Organizational behavior | To report a case study of interagency coordination during the response to a railway accident in the United Kingdom | Interviews were conducted to capture workers' accounts of a railway incident. Interviews were audiorecorded. Investigators also reviewed documentation taken in relation to the accident. | Critical decision method |
| Traffic sign symbol comprehension: a cross-cultural study (Shinar et al., 2003) | Surface transportation system (highway design), warning perception | To understand the cultural difference in comprehending sign symbol among five countries | One thousand unpaid participants were recruited. Participants were presented with 31 multinational traffic signs and asked their comprehension of each. | ANOVA; arcsin $\sqrt{p}$ transformation |

**Table A2  Examples of Empirical Studies**

| Study | HF/E Subdiscipline | Goal | Methodology | Analysis Methods |
|---|---|---|---|---|
| Detection of temporal delays in visual–haptic interfaces (Vogels, 2004) | Displays and controls, multimodality | To address the question of how large the temporal delay between a visual and a haptic stimulus can be for the stimuli to be perceived as synchronous | Three different experiments were conducted to remove unintended methodological factors and to investigate deeply. Learning effect was controlled through training. | Mathematical model developed and applied to quantify the data; ANOVA used to analyze the quantified mathematical model |
| Adjustable typography: an approach to enhancing low-vision text accessibility (Arditi, 2004) | Accessibility, typography | To show that adjustable typography enhances text accessibility | Participants who had low vision were allowed to adjust key font parameters (e.g., size and spacing) of text on a computer display monitor. After adjustment, the participants' accuracy on a reading task was collected. Participants completed the experiment within a single experimental session due to the fatigue experienced in this predominantly older population (mean age = 68.6 years). | Box-and-whisker plot; regression analysis |
| No evidence for prolonged latency of saccadic eye movements due to intermittent light of a CRT computer screen (Jainta et al., 2004) | Psychomotor processes (eye movement) | To show that there is no clear relationship between latency of saccadic eye movements and the intermittency of light of cathode-ray tubes | A special fluorescent lamp display was used to control the refresh rate. An eye tracker captured saccadic eye movements. | ANOVA with repeated measures; Greenhouse–Geisser adjusted error probabilities |
| Physical workload during use of speech recognition and traditional computer input devices (Juul-Kristensen et al., 2004) | Work physiology (physical workload), HCI | To investigate musculoskeletal workload during computer work using speech recognition and traditional computer input devices | The workload of 10 participants while performing text entry, editing, and reading aloud with and without the speech recognition program was studied. Workload was measured using muscle activity (EMG). | Nonparametric statistics (e.g., Wilcoxon's ranked-sign test, Mann–Whitney test) |
| Attentional models of multitask pilot performance using advanced display technology (Wickens et al., 2003b) | Aerospace systems, attention | To compare air traffic control presentation of auditory (voice) information regarding traffic and flight parameters with advanced display technology presentation of equivalent information | Pilots were exposed to both auditory and advanced display technology conditions. Performance with the information presented in each condition was assessed. A Latin square design was used to counterbalance order effects. | Within-subjects ANOVA and regression analysis |

*(continued overleaf)*

**Table A2  (*continued*)**

| Study | HF/E Subdiscipline | Goal | Methodology | Analysis Methods |
|---|---|---|---|---|
| Role of spatial abilities and age in performance in an auditory computer navigation task (Pak et al., 2008) | Aging, displays, and controls | To investigate the relationship between spatial ability and performance in a nonvisual computer-based navigation task for participants in three age groups: younger (ages 18–39), middle aged (40–59) and older (60–91) | Over the course of two days, prescreened participants were given cognitive battery tests and ability tests concerning their working memory, attention , and spatial abilities..On the second day, participants dialed into a fictional interactive voice response system to complete tasks such as obtaining information in a banking or electric utility context. | Regression analysis |
| Measuring the fit between human judgments and automated alerting algorithms: a study of collision detection (Bisantz and Pritchett, 2003) | Aviation, automation | To evaluate the impact of displays on human judgment using the *n*-system lens model; to explicitly assess the similarity between human judgments and a set of potential judgment algorithms for use in automated systems | Using a flight simulator, the approach of an oncoming aircraft was manipulated. Data were collected on the performance of the automation system and its effect on pilot judgments. A time-sliced approach was used to capture wider environmental conditions. | Within-subjects ANOVA used on a transformation of the data |
| Bimodal displays improve speech comprehension in environments with multiple speakers (Rudmann et al., 2003) | Displays and controls | To prove that showing additional visual cues from a speaker can improve speech comprehension | Twenty-four participants were exposed to voice recordings with and without visual cues. In some trials, noise distracters were introduced. The level of participant comprehension was assessed while listening to the recording, and eye movement data were collected. | Within-subjects ANOVA |
| Performance in a complex task and breathing under odor exposure (Danuser et al., 2003) | Displays and controls (olfactory displays) | To investigate the influence of odor exposure on performance and breathing | Fifteen healthy individuals were each exposed to different odors. To capture the emotional status, a self-assessment manikin (SAM) was used. | ANOVA with repeated measures; Wilcoxon's ranked-sign test |
| Time estimation during prolonged sleep deprivation and its relationship to activation measures (Miro et al., 2003) | Fatigue | To investigate the effect of prolonged sleep deprivation for 60 h on time estimation | *Longitudinal*: For 60 h of sleep deprivation, time estimations were measured every 2 h. Skin resistance level, body temperature, and Stanford sleepiness scale scores were collected. | ANOVA with repeated measures; regression analysis (linear, quadratic, quintic, and sextic) |

**Table A2** (*continued*)

| Study | HF/E Subdiscipline | Goal | Methodology | Analysis Methods |
|---|---|---|---|---|
| Impact of mental fatigue on exploration in a complex computer task: rigidity and loss of systematic strategies (Van Der Linden et al., 2003) | Fatigue; HCI | To investigate the impact of mental fatigue on how people explore in a complex computer task | Sixty-eight participants (psychology students) performed a complex computer test using the think-aloud protocol. Data were collected on mental fatigue and performance through observation, videotaping, the activation–deactivation checklist, and the rating scale mental effort. | Multivariate test ($\alpha = 0.10$); univariate/post hoc tests ($\alpha = 0.05$) |
| What to expect from immersive virtual environment exposure: influences of gender, body mass index, and past experience (Stanney et al., 2003) | Virtual reality | To investigate potential adverse effects, including sickness, associated with exposure to virtual reality and extreme responses | Of the 1102 subjects recruited for participation, 142 (12.9%) dropped out because of sickness. Qualitative measurement tools were used to assess motion sickness with the motion history questionnaire and simulator sickness. Sessions were videotaped for archival purposes. | Spearman's correlation test; Kruskal–Wallis nonparametric test; chi-squared test |
| Control and perception of balance at elevated and sloped surfaces (Simeonov et al., 2003) | Work physiology | To investigate the effects of the environment characteristics of roof work (e.g., surface slope, height, and visual reference) on standing balance in construction workers | Twenty-four participants were recruited. The slope of a platform, on which they stood, was varied. At each slope the participant performed a manual task and were asked afterward to rate their perceived balance. Instrumentation measured the central pressure movement. Each subject received the same 16 treatments ($4 \times 2 \times 2$). Balanced to control order effects. | ANOVA with repeated measures and the Student–Newman–break Keuls multiple-range test used when ANOVA indicated significance |
| Multimodal feedback: an assessment of performance and mental workload (Vitense et al., 2003) | Displays and controls (multimodality), HCI | To establish recommendations for the incorporation of multimodal feedback in a drag-and-drop task | The NASA-TLX was used to assess workload. Time measures, such as trial completion time and target highlight time, were used to capture performance as it was affected by multimodal feedback. | Interaction plots |
| Contribution of apparent and inherent usability to a user's satisfaction in a searching and browsing task on the Web (Fu and Salvendy, 2002) | Usability, World Wide Web | To investigate the impact of inherent and apparent usability on user's satisfaction of Web page designs | The questionnaire for user interaction satisfaction was used to measure the levels of users' satisfaction with a browsing task completed on one of four interfaces. | ANOVA; stepwise regression analysis |

**Table A3  Examples of Evaluation Studies**

| Study | HF/E Subdiscipline | Goal | Methodology | Analysis Methods |
|---|---|---|---|---|
| Handle dynamics predictions for selected power hand tool applications (Lin et al., 2003) | Biomechanics | To test a previously developed model of handle dynamics by collecting muscle activity (EMG) data | Muscle activity (EMG) was collected to calculate the magnitude of torque that participants experienced. | Regression analysis |
| Learning to use a home medical device: mediating age-related differences with training (Mykityshyn et al., 2002) | Aging, training | To examine the differential benefits of instructional materials, such as a user manual or an instructional video, for younger and older adults learning to use a home medical device | *Longitudinal*: The NASA task load index (TLX) was used to assess the workload associated with instructional methods. A longitudinal study, there was a two-week retention session used between training and measurement. | ANOVA |
| The influence of distraction and driving context on driver response to imperfect collision warning systems (Lees and Lee, 2007) | Surface transportation system (driver behavior) | To determine how false (FA) and unnecessary alarms (UA) impact collision warning system (CWS) effectiveness | A driving simulator was used to investigate the influence of accurate, UA, FA, and no warnings from the CWS effect driver performance during noncritical and critical events. A total of 64 drivers between the ages of 20 and 35 were prescreened for simulator sickness prior to participation. The CWS utilized an auditory alarm to warn the driver of impending rear-end collisions. . | ANOVA; Tukey-Kramer |
| Fleet study evaluation of an advance brake warning system (Shinar, 2000) | Accident, surface transportation system | To prove the effectiveness of an advanced brake warning system (ABWS) | *Longitudinal and field study*: Traced accidents of vehicles with ABWS-installed government vehicles (382) and non-ABWS vehicles (382) for 4 years. | *t*-Test; chi-square test |
| Continuous assessment of back stress (CABS): a new method to quantify low-back stress in jobs with variable biomechanical demands (Mirka et al., 2000) | Work physiology | To compare three different back-stress modeling techniques for the continuous assessment of lower back stress in various situations and to incorporate them into a hybrid model | *Field study and observation*: 28 construction workers were observed and videotaped while on the job. Based on this videotaped observation, three models of each task that induced biomechanical stress were created. Each model was evaluated based on the extent to which it modeled the actual biomechanical stress experienced. | Time-weighted histograms to aggregate and compare the output from three different models |

# REFERENCES

Agha, S.R. (2010) School furniture match to students' anthropometry in the Gaza Strip. Ergonomics. 2010; 53 (3): 344–54.

American Psychological Association (2010), "Ethical Principles of Psychologists and Code of Conduct 2010 Amendments," available: http://www.apa.org/ethics/code/index.aspx, accessed September 27, 2010.

Arditi, A. (2004), "Adjustable Typography: An Approach to Enhancing Low Vision Text Accessibility," *Ergonomics*, Vol. 47, No. 5, pp. 469–482.

Arleck, P. L., and Settle, R. B. (1995), *The Survey Research Handbook*, Irwin, Chicago.

Barlow, D. H., and Nock, M. K. (2009), "Why Can't We Be More Idiographic in Our Research?" *Perspectives on Psychological Science*, Vol. 4, No. 1, pp. 19–21.

Bisantz, A. M., and Pritchett, A. R. (2003), "Measuring the Fit between Human Judgments and Automated Alerting Algorithms: A Study of Collision Detection," *Human Factors*, Vol. 45, No. 2, pp. 266–280.

Bolton, M.L., and Bass, E.J. (2009), "A Method for the Formal Verification of Human-Interactive Systems," *Human

*Factors and Ergonomics Society Annual Meeting Proceedings*, Vol. 53, No. 12, pp. 764–768.

Box, G. E., Hunter, W. G., and Hunter, J. S. (1978), *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, Wiley, New York.

Byrne, M. D. (2001), "ACT-R/PM and Menu Selection: Applying a Cognitive Architecture to HCI," *International Journal of Human–Computer Studies*, Vol. 55, pp. 41–84.

Card, S., Moran, T., and Newell, A. (1983), *The Psychology of Human–Computer Interaction*, Lawrence Erlbaum Associates, Mahwah, NJ.

Carpenter, B. D. and Buday, S. (2007). Computer use among older adults in a naturally occurring retirement community. *Comput. Hum. Behav.*, 23: 3012–3024.

Cohen, R. J., and Swerdlik, M. (2002), *Psychological Testing and Assessment: An Introduction to Tests and Measurement*, 5th ed., McGraw-Hill, Burr Ridge, IL.

Committee on Human Factors (1983), *Research Needs for Human Factors*, Commission on Behavioral and Social Sciences and Education, National Academy of Sciences, Washington DC.

Creswell, J. R. (2009), *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, Sage Publications, Los Angeles, CA.

Danuser, B., Moser, D., Vitale-Sethre, T., Hirsig, R., and Krueger, H. (2003), "Performance in a Complex Task and Breathing under Odor Exposure," *Human Factors*, Vol. 45, No. 4, pp. 549–562.

Donderi, D. C., Niall, K. K., Fish, K., and Goldstein, B. (2010), "Effects of Screen Resolution and Training Variation on a Simulated Flight Control Task," in *Vision and Displays for Military and Security Applications*, K. Niall, Ed., Springer, New York, Chapter~11, pp. 141–161.

Drury, C. G. (2005), "Designing Ergonomic Studies and Experiments," in *Evaluation of Human Work*, 3rd ed., J. R. Wilson and E. N. Corlett, Eds., Taylor & Francis, Bristol, PA, pp. 113–140.

Fitts, P. M. (1954), "The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement," *Journal of Experimental Psychology*, Vol. 47, No. 6 pp. 381–391.

Ford, J. M., and Wood, L. E. (1996), "An Overview of Ethnography and System Design," in *Field Methods Casebook for Software Design*, D. Wixon and J. Ramey, Eds., Wiley Computer Publishing, New York, pp. 269–282.

Fu, L., and Salvendy, G. (2002), "The Contribution of Apparent and Inherent Usability to a User's Satisfaction in a Searching and Browsing Task on the Web," *Ergonomics*, Vol. 45, No. 6, pp. 415–424.

Gawron, V. J. (2000). *Human performance measures handbook*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gawron, V. J. (2008), *Human Performance, Workload, and Situational Awareness Measures Handbook*, Taylor & Francis Group, Boca Raton, FL.

Gillespie, R. (1991), *Manufacturing Knowledge: A History of the Hawthorne Experiments*, Cambridge University Press, New York.

Gould, J. E. (2002), *Concise Handbook of Experimental Methods for the Behavioral and Biological Sciences*, CRC Press, Boca Raton, FL.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., and Singer, E. (2009), *Survey Methodology*, Wiley Hoboken, NJ.

Hancock, H. E., Fisk, A. D., and Rogers, W. A. (2005). "Comprehending Product Warning Information: Age-Related Effects and the Roles of Memory, Inferencing, and Knowledge," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 47, No.2, pp. 219–234.

Holtzblatt, K., and Beyer, H. (1996), "Contextual Design: Principles and Practice," in *Field Methods Casebook for Software Design*, D. Wixon and J. Ramey, Eds., Wiley Computer Publishing, New York, pp. 301–354.

Hsiao, H., Long, D., and Snyder, K. (2002), "Anthropometric Differences among Occupational Groups," *Ergonomics*, Vol. 45, No. 2, pp. 136–152.

Jacko, J. A., Moloney, K. P., Kongnakorn, T., Barnard, L., Edwards, P., Lenard, V. K., Sainfort, F., and Scott, I. U. (2005), "Multimodal Feedback as a Solution to Ocular Disease-Based User Performance Decrements in the Absence of Functional Visual Loss," *International Journal of Human–Computer Interaction*, Vol. 18, No. 2, pp. 183–218.

Jainta, S., Jaschinski, W., and Baccino, T. (2004), "No Evidence for Prolonged Latency of Saccadic Eye Movements Due to Intermittent Light of a CRT Computer Screen," *Ergonomics*, Vol. 47, No. 1, pp. 105–114.

Juul-Kristensen, B., Laursen, B., Pilegaard, M., and Jensen, B. R. (2004), "Physical Workload during Use of Speech Recognition and Traditional Computer Input Devices," *Ergonomics*, Vol. 47, No. 2, pp. 119–133.

Kanis, H. (2000), "Questioning Validity in the Area of Ergonomics/Human Factors," *Ergonomics*, Vol. 43, No. 12, pp. 1947–1965.

Kantowitz, B. H. (1992), "Selecting Measures for Human Factors Research," *Human Factors*, Vol. 34, No. 4, pp. 387–398.

Karrer, K., and Roetting, M. (2007), "Effects of Driver Fatigue Monitoring - An Expert Survey," in *Engineering Psychology and Cognitive Ergonomics*, Vol. 4562 of *Lecture Notes in Computer Science*, D. Harris, Ed., Springer, Berlin/Heidelberg, pp. 324–330.

Lee, J. D., Cassano-Pinché, A., and Vicente, K. J. (Winter 2005), "Bibliometric Analysis of Human Factors (1970-2000): A Quantitative Description of Scientific Impact," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 47, No. 4, pp. 753–766.

Lees, M. N., and Lee, J. D. (2007), "The Influence of Distraction and Driving Context on Driver Response to Imperfect Collision Warning Systems," *Ergonomics*, Vol. 50, No. 8, pp. 1264–1286.

Lin, J.-H., Radwin, R. G., and Richard, T. G. (2003), "Handle Dynamics Predictions for Selected Power Hand Tool Applications," *Human Factors*, Vol. 45, No. 4, pp. 645–656.

MacKenzie, I. S. (1992a), "Fitts' Law as a Research Design Tool in Human–Computer Interaction," *Human–Computer Interaction*, Vol. 7, No. 1992, pp. 91–139.

MacKenzie, I. S. (1992b), "Extending Fitts' Law to Two Dimensional Tasks," in *Human Factors in Computing Systems*, ACM Press, New York, pp. 219–226.

Meister, D. (1971), *Human Factors: Theory and Practice*, Wiley-Interscience, New York.

Meister, D. (2004), *Conceptual Foundations of Human Factors Measurement*, Lawrence Erlbaum Associates, Mahwah, NJ.

Menard, S. (2002), *Longitudinal Research*, 2nd ed., Sage Publications, Newbury Park, CA.

Mendonca, D., Beroggi, G.E.G., and Wallace, W. A. (2001), "Decision Support for Improvisation during Emergency Response Operations," *International Journal of Emergency Management*, Vol. 1, pp. 30–38.

Miller D. C., & Salkind, N. J. (2002). *Handbook of research design and social measurement* (6th ed.). Thousand Oaks, CA: Sage.

Mirka, G. A., Kelaher, D. P., Nay, D. T., and Lawrence, B. M. (2000), "Continuous Assessment of Back Stress (CABS): A New Method to Quantify Low-Back Stress in Jobs with Variable Biomechanical Demands," *Human Factors*, Vol. 42, No. 2, pp. 209–225.

Miro, E., Cano, M. C., Espinosa-Fernandez, L., and Buela-Casal, G. (2003), "Time Estimation during Prolonged Sleep Deprivation and Its Relation to Activation Measures," *Human Factors*, Vol. 45, No. 1, pp. 148–159.

Mykityshyn, A. L., Fisk, A. D., and Rogers, W. A. (2002), "Learning to Use a Home Medical Device: Mediating Age-Related Differences with Training," *Human Factors*, Vol. 44, No. 3, pp. 354–364.

Nembhard, D. A. (2000), "The Effects of Task Complexity and Experience on Learning and Forgetting: A Field Study," *Human Factors*, Vol. 42, No. 2, pp. 272–286.

"Nuremberg Code (1947)" (1996), *British Medical Journal*, Vol. 313, No. 7070, p. 1448.

Office for Human Research Protections (1979), "The Belmont Report," available: http://ohsr.od.nih.gov/guidelines/belmont.html, accessed May 5, 2005.

Panwaree, J., Klaphajone, J., and Kovindha, A. (2009), "Neck Pain in Computer Users: A Study of Ergonomic Factors," *Journal of Thai Rehabilitation Medicine*, Vol. 19, No. 1, available: http://thailand.digitaljournals.org/index.php/JTRM/article/view/2731/0.

Pollatsek, A., and Rayner, K. (1998), "Behavioral Experimentation," in *A Companion to Cognitive Science*, W. Bechtel and G. Graham, Eds., Blackwell, Malden, MA, pp. 352–370.

Pritchett, A. R. (2002), "Human–Computer Interaction in Aerospace," in *Handbook of Human–Computer Interaction*, J. A. Jacko and A. Sears, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 861–882.

Proctor, R.W., and Vu, K.L. (2010), "Cumulative Knowledge and Progress in Human Factors," *Annual Review of Psychology*, Vol. 61., pp. 623–651.

Ramcharan, D. (2001), "Ergonomic Redesign of a Hanging Table for Poultry Processing," in *Applied Ergonomics*, D. Alexander and R. Rabourn, Eds., Taylor & Francis, New York, pp. 126–134.

Rau, P.-L. P., and Salvendy, G. (2001), "Ergonomics of Electronic Mail Address Systems: Related Literature Review and Survey of Users," *Ergonomics*, Vol. 44, No. 4, pp. 382–401.

Rehmann, A. J. (1995), *Handbook of Human Performance Measures and Crew Requirements for Flightdeck Research*, Federal Aviation Administration, Washington, DC.

Riley, J. M., Endsley, M. R., Bolstad, C. A., and Cuevas, H. M. (2006), "Collaborative Planning and Situation Awareness in Army Command and Control," *Ergonomics*, Vol. 49, No. 12, pp. 1139–1153.

Rudmann, D. S., McCarley, J. S., and Kramer, A. F. (2003), "Bimodal Displays Improve Speech Comprehension in Environments with Multiple Speakers," *Human Factors*, Vol. 45, No. 2, pp. 329–336.

Sage, A. P., and Rouse, W. B. (2009), *Handbook of Systems Engineering and Management*, Wiley, Hoboken, NJ.

Sanders, M. S., and McCormick, E. J. (1993), *Human Factors in Engineering and Design*, McGraw-Hill, New York.

Sheridan, T. B. (2002), *Humans and Automation: System Design and Research Issues*, Wiley, New York.

Shinar, D. (2000), "Fleet Study Evaluation of an Advance Brake Warning System," *Human Factors*, Vol. 42, No. 3, pp. 482–489.

Shinar, D., Dewar, R. E., Summala, H., and Zakowska, L. (2003), "Traffic Sign Symbol Comprehension: A Cross-Cultural Study," *Ergonomics*, Vol. 46, No. 15, pp. 1549–1565.

Simeonov, P. I., Hsiao, H., Dotson, B. W., and Ammons, D. E. (2003), "Control and Perception of Balance at Elevated and Sloped Surfaces," *Human Factors*, Vol. 45, No. 1, pp. 136–147.

Singer, S. J., Rosen, A., Zhao, S., Ciavarelli, A. P., and Gaba, D. M. (2010), "Comparing Safety Climate in Naval Aviation and Hospitals: Implications for Improving Patient Safety," *Health Care Management Review*, Vol. 35, No. 2, pp. 134–46

Smith, R. A. and Davis, S. F. (2008) The Changing Face of Research Methods, in Handbook of Research Methods in Experimental Psychology (ed S. F. Davis), Blackwell Publishing Ltd, Oxford, UK. doi:10.1002/9780470756973.ch6

Smith, W., and Dowell, J. (2000), "A Case-Study of Coordinative Decision-Making in Disaster Management," *Ergonomics*, Vol. 43, No. 8, pp. 1153–1166.

Stanney, K. M., Hale, K. S., Nahmens, I., and Kennedy, R. S. (2003), "What to Expect from Immersive Virtual Environment Exposure: Influences of Gender, Body Mass Index, and Past Experience," *Human Factors*, Vol. 45, No. 3, pp. 504–520.

Stanton, N. A., and Young, M. A. (1999), *A Guide to Methodology in Ergonomics*, Taylor & Francis, New York.

Stanton, N. A., Salmon, P. M., Walker, G. H, Baber, C., and Jenkins, D. P. (2005), *Human Factors Methods: A Practical Guide for Engineering and Design*, TJ International, Padstow, Cornwall.

U.S. Department of Energy (2004), "DOE Openness: Human Radiation Experiments," available: http://tis.eh.doe.gov/ohre/, accessed May 5, 2005.

U.S. Department of Health and Human Services (2009), "Code of Federal Regulations, Part 46: Protection of Human Subjects," available: http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm, accessed September 27, 2010.

van der Linden, D., Frese, M., and Sonnentag, S. (2003), "The Impact of Mental Fatigue on Exploration in a Complex Computer Task: Rigidity and Loss of Systematic Strategies," *Human Factors*, Vol. 45, No. 3, pp. 483–494.

Vitense, H. S., Jacko, J. A., & Emery, V. K. (2003). Multimodal feedback: An assessment of performance and mental workload. *Ergonomics*, *46*, 68–87.

Vogels, I. M. (2004), "Detection of Temporal Delays in Visual–Haptic Interfaces," *Human Factors*, Vol. 46, No. 1, pp. 118–134.

Weimer, J. (1995), "Developing a Research Project," in *Research Techniques in Human Engineering*, J. Weimer, Ed., Prentice Hall, Englewood Cliffs, NJ, pp. 35–54.

Wickens, C. D., Lee, J. D., Liu, Y., and Gordon-Becker, S. (2003a), *An Introduction to Human Factors Engineering*, Prentice Hall, Englewood Cliffs, NJ.

Wickens, C. D., Goh, J., Helleberg, J., Horrey, W. J., and Talleur, D. A. (2003b), "Attentional Models of Multitask Pilot Performance Using Advanced Display Technology," *Human Factors*, Vol. 45, No. 3, pp. 360–380.

Williges, R. C. (1995), "Review of Experimental Design," in *Research Techniques in Human Engineering*, J. Weimer, Ed., Prentice Hall, Englewood Cliffs, NJ, pp. 49–71.

Wilson, J. R., and Corlett, E. N. (2005). *Evaluation of Human Work*, Taylor & Francis Group, Boca Raton, FL.

Wixon, D., and Ramey, J. (1996), *Field Methods Casebook for Software Design*, Wiley Computer Publishing, New York.

Wobbrock, J. O., Myers, B. A., and[check] Aung, H. H. (2008). The performance of hand postures in front- and back-of-device interaction for mobile computing. *International Journal of Human-Computer Studies*, *66* (12), 857–875. doi:DOI: 10.1016/j.ijhcs.2008.03.004

Woods, L. (1996), "The Ethnographic Interviews in User Centered Work/Task Analysis," in *Field Methods Casebook for Software Design*, D. Wixon and J. Ramey, Eds., Wiley Computer Publishing, New York, pp. 35–56.

World Medical Association (2002), "Police: World Medical Association Declaration of Helsinki Ethical Principles for Medical Research Involving Human Subjects," available: http://www.wma.net/e/policy/b3.htm, accessed June 23, 2004.

# CHAPTER 11

# ANTHROPOMETRY FOR PRODUCT DESIGN

**Kathleen M. Robinette**
**Air Force Research Laboratory**
**Wright-Patterson Air Force Base, Ohio**

## 1 INTRODUCTION

Did you know that designing for the 5th percentile female to the 95th percentile male can lead to poor and unsafe designs? If not, you are not alone. These and similar percentile cases, such as the 99th percentile male, are the only cases presented as anthropometry solutions to many engineers and ergonomics professionals. In this chapter we review this and other anthropometry issues and present an overview of practical effective methods for incorporating the human body in design.

*Anthropometry*, the study of human body measurement, is used in engineering to ensure the maximum benefit and capability of products that people use. The use of anthropometric data in the early concept stage can minimize the size and shape changes needed later, when modifications can be very expensive. To use anthropometry knowledge effectively, it is also important to have knowledge of the relationships between the body and the items worn or used. The study of these relationships is called *fit mapping*. Databases containing both anthropometry and fit-mapping data can be used as a lessons-learned source for development of new products. Therefore, anthropometry and fit mapping can be thought of as an information core around which products are designed, as illustrated in Figure 1.

A *case* is a representation of a combination of body measurements, such as a list of measurements on one subject, the average measurements from a sample, a three-dimensional scan of a person, or a two- or three-dimensional human model. If the relationship between the anthropometry and the fit of a product is simple or known, cases may be all that is needed to arrive at an effective design. However, if the relationship between the anthropometry and the fit is complex or unknown, cases alone may not suffice. In these situations, fit mapping with a prototype, mock-up, or similar product is needed to determine how to accommodate the cases and to predict accommodation for any case.

The chapter is divided into three sections. Section 2 deals with the selection of cases for characterizing anthropometric variability. Section 3 covers fit-mapping methods. Section 4 is devoted to some of the benefits of the newest method in anthropometric data collection, three-dimensional anthropometry.

## 2 ANTHROPOMETRY CASES: ALTERNATIVES, PITFALLS, AND RISKS

When a designer or engineer asks the question "What anthropometry should I use in the design?" he or she is essentially asking, "What cases should I design around?" Of course, the person would always like to be given one case or one list of measurements and be told that nothing else is needed. That would make it simple. However, if the question has to be asked, the design is probably more complicated than that, and the answer is correspondingly more complicated as well. In this section we discuss alternative ways to determine which cases to use.

### 2.1 Averages and Percentiles

Since as early as 1952, when Daniels (1952) presented the argument that no one is average, we have known that anthropometric averages are not acceptable for many applications. For example, an average human head is not appropriate to use for helmet sizing, and an average female shape is not appropriate for sizing apparel. In addition, we have known since Searle and Haslegrave (1969) presented their debate with Ed Hertzberg that the 5th and 95th percentile people are no better. Robinette and McConville (1982) demonstrated that it is not even possible to construct a 5th or 95th percentile human figure: The values do not add up. This means that 5th or 95th percentile values can produce very unrealistic figures that do not have the desired 5th or 95th percentile size for some of their dimensions.

The impact of using percentiles can be huge. For example, for one candidate aircraft for the T-1 program, the use of the 1st percentile female and 99th percentile male resulted in an aircraft that 90% of females, 80% of African-American males, and 30% of white males
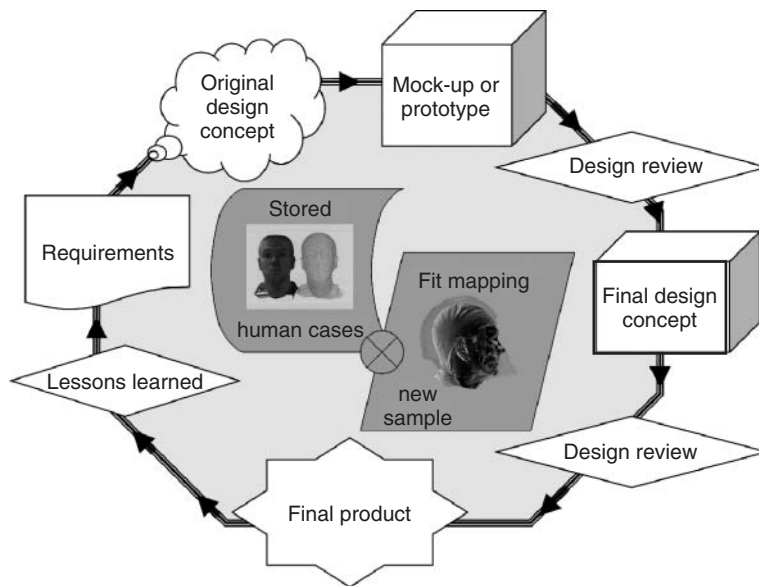
**Figure 1**  Anthropometric information as a design core.

could not fly. The problem is illustrated in Figure 2. The pilots needed to be able, simultaneously, to see over the nose of the plane and operate the yoke, a control that is similar to a steering wheel in a car. For the 99th percentile seated eye height, the seat would be adjusted all the way down to enable the pilot to see over the nose. For the 1st percentile seated eye height, the seat would be adjusted all the way up. Since the design used all 1st percentile values for the full-up seat position, it accounted for only a 1st percentile or smaller female thigh size when the seat was all the way up. As a result, it did not accommodate most female pilots' thigh size without having the yoke interference as pictured in Figure 2. For designs such as this, where there are conflicting or interacting measurements or requirements, percentiles will not be effective. Cases that have combinations of small and large dimensions are needed.

To understand when to use and when not to use averages and percentiles, it is important to understand what they are and what they are not. Figure 3 illustrates average and percentile values for stature and weight. Sample frequency distributions for these two measurements are shown for the female North American data from the CAESAR survey (Harrison and Robinette, 2002) in the form of histograms. The averages for 5th and 95th percentiles are indicated. The *frequency* is the count of the number of times that a value or range of values occurs, and the vertical bars in Figure 3 indicate the number of people who had a stature or weight of the size indicated. For example, the one vertical bar to the right of the 95th percentile weight indicates that approximately 10 people have a weight between 103 and 105 kg. Percentiles indicate the location of a particular cumulative frequency. For example, the 50th percentile is the point at which 50% have a smaller value, and the 95th percentile is the point at which 95% have a smaller value.



**Figure 2**  Problem that occurred when using 1st percentile female and 99th percentile male.

The *average* is a value for one measurement that falls near the middle of the distribution for that measurement. In this case the arithmetic average is shown. Another kind of central value is the 50th percentile, which will be
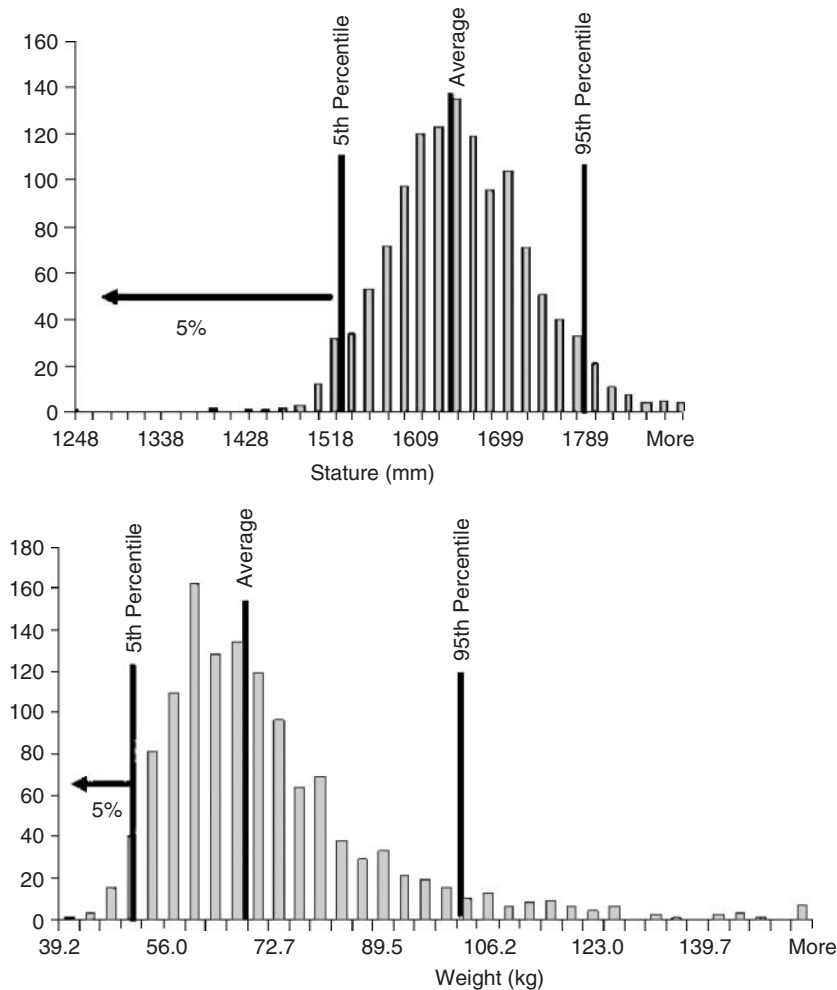
**Figure 3**   Stature and weight univariate frequencies, CAESAR, U.S. females.

the same as the arithmetic average when the frequency distribution is symmetric. The stature distribution shown in Figure 3 is approximately symmetrical, so the average and the 50th percentile differ by just 0.5%, whereas the weight distribution is not symmetric, so the average and the 50th percentile differ by 5.4%.

### 2.1.1   Percentile Issues

Percentile values refer only to the location of the cumulative frequency of one measurement. This means that the 95th percentile weight has no relationship to the 95th percentile stature. This is illustrated in Figure 4, which shows the two-dimensional frequency distribution for stature and weight along with the one-dimensional frequency distributions that appeared in Figure 3. Stature values are represented by the vertical axis and weight by the horizontal. The histogram from Figure 3 for stature is shown to the right of the plot, and the histogram from Figure 3 for weight is shown at the top of the

plot, each with its respective 5th and 95th percentile values. Each of the circular dots in the center of the plot indicates the location of one subject from the sample of CAESAR U.S. females. The ellipse toward the center that surrounds many of the dots is the 90% ellipse; in other words, it encircles 90% of the subjects.

If a designer uses the "5th percentile female to the 95th percentile female" approach, only two cases are being used. These two cases are indicated as black squares, one at the lower left and the other at the upper right of the two-dimensional plot in Figure 4. The one at the lower left is the intersection of the 5th percentile stature and the 5th percentile weight. The one at the upper right is the intersection of the 95th percentile stature and the 95th percentile weight. The stature range from the 5th to 95th percentile falls between the two horizontal 5th and 95th percentile lines and contains approximately 90% of the population. The weight range from the 5th to 95th percentile falls between the
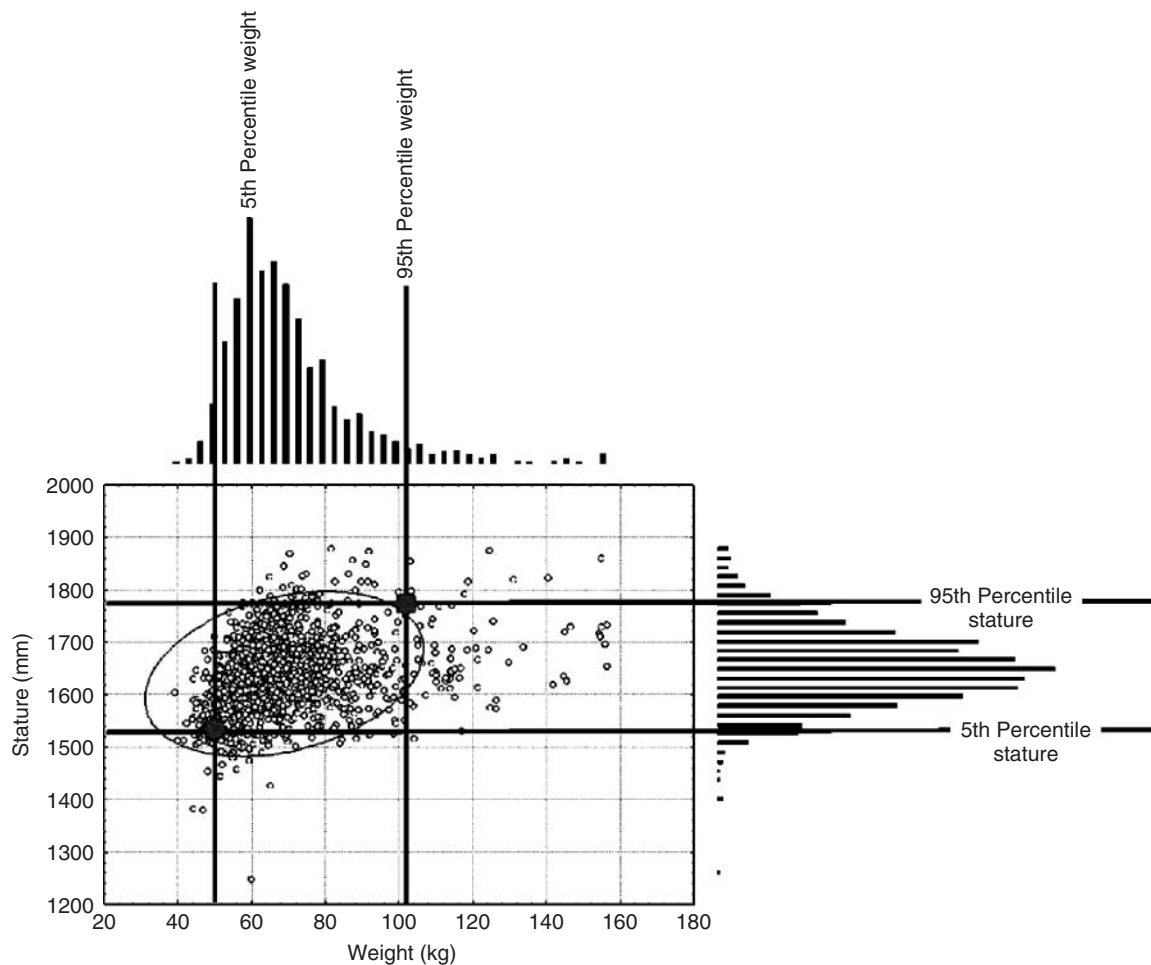
**Figure 4** Bivariate frequency distribution of stature and weight, CAESAR, U.S. females.

two vertical 5th and 95th percentile lines and contains approximately 90% of the population. The people who fall between the 5th and 95th percentiles for both stature and weight are only those people who fall within the intersection of the vertical and horizontal bands. The intersection contains only approximately 82%. If a third measurement is added, it makes the frequency distribution three-dimensional, and the percentage accommodated between the 5th and 95th percentiles for all three measurements will be fewer still.

The bar in the weight histogram in Figure 3 that is just to the right of the 95th percentile is the same as the bar that is just to the right of the 95th percentile weight in Figure 4. As stated above, approximately 10 subjects fall at this point. If you look at the two-dimensional plot, you can see that these 10 subjects who are at approximately the 95th percentile weight have stature values from as small as 1500 mm to as large as 1850 mm. In other words, women in this sample who have a 95th percentile weight have a range of statures that

extends from below the stature 5th percentile to above the stature 95th percentile. This means that if the product has conflicting requirements, the 5th to 95th percentile cases would not work effectively. For example, suppose that a zoo has an automatically adjusting platform for an exhibit that adjusts its height based on the person's weight, and the exhibit designers want to make a window or display large enough so that the population can see the exhibit. If this is designed for the 5th percentile person to the 95th percentile person, they would design for (1) the 5th percentile stature with the 5th percentile weight as one case and (2) the 95th percentile weight with the 95th percentile stature as the other case. Let us use the same female population data to see what the 5th and 95th percentiles would accommodate. At the 5th percentile case the platform would be full up and the stature accommodated would be 1525 mm. At the 95th percentile case the platform would be full down and the stature accommodated would be 1767 mm. This range of stature is 1767 − 1525 = 242

mm. This will accommodate the 5th percentile female to the 95th percentile female, but not the 5th percentile stature with an average weight or the 5th percentile weight with the average stature. The female who has a 5th percentile weight of 49.2 kg but an average stature of 1639 mm would need 114 mm more headroom. The female who has an average weight but a 5th percentile stature may not be able to see the display because the weight-adjusted platform is halfway down. This is illustrated in Figure 5.

### 2.1.2 When to Use Averages and Percentiles

Averages, percentiles, and other one-dimensional summary statistics such as the standard deviation, minimum, and maximum are very useful for comparing measurements captured in different ways or for comparing samples from different populations to determine if there are size and variability differences. For example, Krul et al (2010) provide a good example of the use of summary statistics for comparing self-reported values to measured values for stature, weight and body mass index. In Table 1, one-dimensional summary statistics from the U.S. CAESAR sample (Harrison and Robinette, 2002) are compared with summary statistics from the U.S. ANSUR survey (Gordon et al., 1989) illustrating another example of the proper use of summary statistics. The CAESAR sample was taken from a civilian population, whereas the ANSUR sample was taken from a military population, the U.S. Army. The U.S. Army has fitness and weight limitations for its personnel. As a result, the ANSUR sample has a more limited

range of variability for weight-related measurements. The effect of this can be seen by examining the differences in the weight and buttock–knee length ranges (minimum to maximum) versus the ranges for the other measurements that are less affected by weight.

You might also notice that the difference between CAESAR and ANSUR females in buttock–knee length is greater than the difference in CAESAR and ANSUR males in buttock–knee length. This highlights a key difference between men and women. Women tend to gain weight in their hips, buttocks, and thighs, whereas men tend to gain weight or bulk in their waists and shoulders.

The ANSUR/U.S. CAESAR differences in Table 1 are contrasted with another comparison of anthropometric data in Table 2. This compares the U.S. CAESAR data with those collected in The Netherlands on the Dutch population (TN). The Dutch claim to be the tallest people in Europe, and this is reflected in all the heights and limb lengths. Both the male and female Dutch subjects are more than 30 mm taller on average than their U.S. counterparts.

Averages and percentiles and other one-dimensional statistics can also be very useful for products that do not have conflicting requirements. In these instances the loss in accommodation with each additional dimension can be compensated for by increasing the percentile range for each dimension. For example, if you want to ensure 90% accommodation for a simple design problem (one that has no interactive measurements) with five key measurements, you can use the 1st and 99th percentile
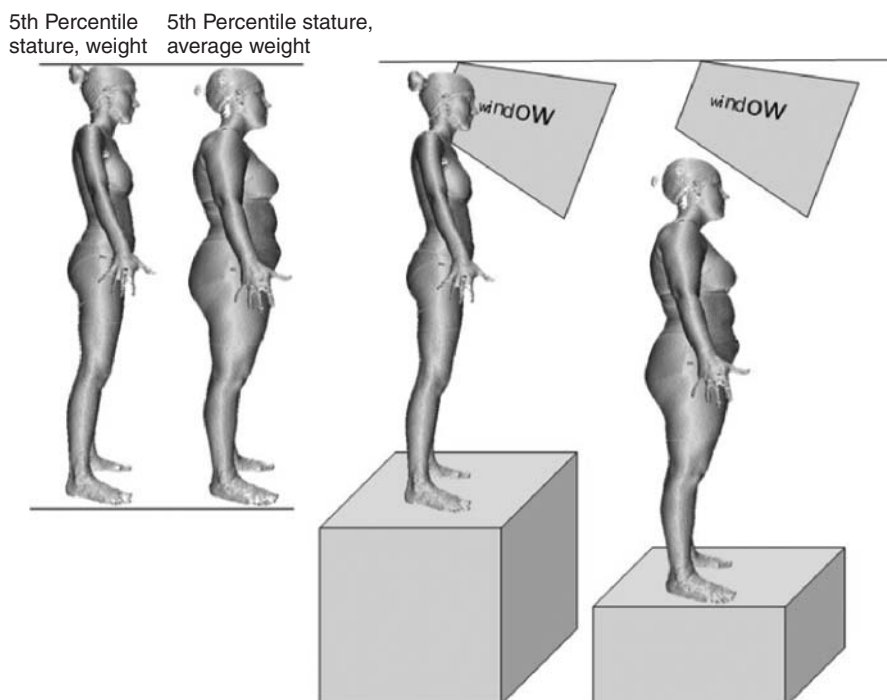


**Figure 5** Woman with 5th percentile stature and average weight is not accommodated.

**Table 1  Comparison of U.S. Civilian Summary Statistics (CAESAR Survey) with U.S. Army Statistics (ANSUR Survey)**

|  | N | Mean | Minimum | Maximum | Std. Dev. |
|---|---|---|---|---|---|
| *Acromion Height, Sitting (mm)* | | | | | |
| Females | | | | | |
| CAESAR | 1264 | 567.42 | 467.00 | 672.00 | 29.76 |
| ANSUR | 2208 | 555.54 | 464.06 | 663.96 | 28.65 |
| Males | | | | | |
| CAESAR | 1127 | 607.21 | 489.00 | 727.00 | 34.19 |
| ANSUR | 1774 | 597.76 | 500.89 | 694.94 | 29.59 |
| *Buttock–Knee Length (mm)* | | | | | |
| Females | | | | | |
| CAESAR | 1263 | 586.97 | 489.00 | 805.00 | 37.43 |
| ANSUR | 2208 | 588.93 | 490.98 | 690.88 | 29.63 |
| Males | | | | | |
| CAESAR | 1127 | 618.93 | 433.00 | 761.00 | 35.93 |
| ANSUR | 1774 | 616.41 | 505.97 | 722.88 | 29.87 |
| *Sitting Eye Height (mm)* | | | | | |
| Females | | | | | |
| CAESAR | 1263 | 755.34 | 625.00 | 878.00 | 34.29 |
| ANSUR | 2208 | 738.71 | 640.08 | 864.11 | 33.24 |
| Males | | | | | |
| CAESAR | 1127 | 808.07 | 681.00 | 995.00 | 39.24 |
| ANSUR | 1774 | 791.97 | 673.10 | 902.97 | 34.21 |
| *Sitting Knee Height (mm)* | | | | | |
| Females | | | | | |
| CAESAR | 1264 | 509.06 | 401.00 | 649.00 | 28.28 |
| ANSUR | 2208 | 515.41 | 405.89 | 632.97 | 26.33 |
| Males | | | | | |
| CAESAR | 1127 | 562.35 | 464.00 | 671.00 | 31.27 |
| ANSUR | 1774 | 558.79 | 453.90 | 674.88 | 27.91 |
| *Sitting Height (mm)* | | | | | |
| Females | | | | | |
| CAESAR | 1263 | 865.02 | 720.00 | 994.00 | 36.25 |
| ANSUR | 2208 | 851.96 | 748.03 | 971.04 | 34.90 |
| Males | | | | | |
| CAESAR | 1127 | 925.77 | 791.00 | 1093.00 | 40.37 |
| ANSUR | 1774 | 913.93 | 807.97 | 1032.00 | 35.58 |
| *Stature (mm)* | | | | | |
| Females | | | | | |
| CAESAR | 1264 | 1639.66 | 1248.00 | 1879.00 | 73.23 |
| ANSUR | 2208 | 1629.38 | 1427.99 | 1869.95 | 63.61 |
| Males | | | | | |
| CAESAR | 1127 | 1777.53 | 1497.00 | 2084.00 | 79.19 |
| ANSUR | 1774 | 1755.81 | 1497.08 | 2041.91 | 66.81 |
| *Thumb Tip Reach (mm)* | | | | | |
| Females | | | | | |
| CAESAR | 1264 | 738.65 | 603.30 | 888.00 | 39.55 |
| ANSUR | 2208 | 734.61 | 605.03 | 897.89 | 36.45 |
| Males | | | | | |
| CAESAR | 1127 | 813.99 | 694.60 | 1027.00 | 44.11 |
| ANSUR | 1774 | 800.84 | 661.92 | 979.93 | 39.17 |
| *Weight (kg)* | | | | | |
| Females | | | | | |
| CAESAR | 1264 | 68.84 | 39.23 | 156.46 | 17.60 |
| ANSUR | 2208 | 62.00 | 41.29 | 96.68 | 8.35 |
| Males | | | | | |
| CAESAR | 1127 | 86.24 | 45.80 | 181.41 | 18.00 |
| ANSUR | 1774 | 78.47 | 47.59 | 127.78 | 11.10 |

**Table 2  Comparison of the U.S. (US) and Dutch (TN) Statistics from the CAESAR Survey**

|  | N | Mean | Minimum | Maximum | Std. Dev. |
|---|---|---|---|---|---|
| *Acromion Height, Sitting* (*mm*) | | | | | |
| Females | | | | | |
| US | 1264 | 567.42 | 467.00 | 672.00 | 29.76 |
| TN | 687 | 589.53 | 490.98 | 709.93 | 33.20 |
| Males | | | | | |
| US | 1127 | 607.21 | 489.00 | 727.00 | 34.19 |
| TN | 559 | 629.89 | 544.07 | 739.90 | 35.96 |
| *Buttock–Knee Length* (*mm*) | | | | | |
| Females | | | | | |
| US | 1263 | 586.97 | 489.00 | 805.00 | 37.43 |
| TN | 688 | 608.07 | 515.87 | 728.98 | 31.15 |
| Males | | | | | |
| US | 1127 | 618.93 | 433.00 | 761.00 | 35.93 |
| TN | 558 | 636.22 | 393.95 | 766.06 | 37.44 |
| *Sitting Eye Height* (*mm*) | | | | | |
| Females | | | | | |
| US | 1263 | 755.34 | 625.00 | 878.00 | 34.29 |
| TN | 676 | 774.46 | 664.97 | 942.09 | 35.82 |
| Males | | | | | |
| US | 1127 | 808.07 | 681.00 | 995.00 | 39.24 |
| TN | 593 | 825.35 | 736.09 | 957.07 | 39.71 |
| *Sitting Knee Height* (*mm*) | | | | | |
| Females | | | | | |
| US | 1264 | 509.06 | 401.00 | 649.00 | 28.28 |
| TN | 676 | 510.67 | 407.92 | 600.96 | 28.78 |
| Males | | | | | |
| US | 1127 | 562.35 | 464.00 | 671.00 | 31.27 |
| TN | 549 | 557.09 | 369.06 | 680.97 | 35.97 |
| *Sitting Height* (*mm*) | | | | | |
| Females | | | | | |
| US | 1263 | 865.02 | 720.00 | 994.00 | 36.25 |
| TN | 687 | 884.64 | 766.06 | 1049.02 | 38.06 |
| Males | | | | | |
| US | 1127 | 925.77 | 791.00 | 1093.00 | 40.37 |
| TN | 559 | 941.85 | 823.98 | 1105.92 | 42.54 |
| *Stature* (*mm*) | | | | | |
| Females | | | | | |
| US | 1264 | 1639.66 | 1248.00 | 1879.00 | 73.23 |
| TN | 679 | 1672.29 | 1436.12 | 1947.93 | 79.02 |
| Males | | | | | |
| US | 1127 | 1777.53 | 1497.00 | 2084.00 | 79.19 |
| TN | 593 | 1808.08 | 1314.96 | 2182.88 | 92.81 |
| *Thumb Tip Reach* (*mm*) | | | | | |
| Females | | | | | |
| US | 1264 | 738.65 | 603.30 | 888.00 | 39.55 |
| TN | 690 | 751.22 | 632.97 | 889.25 | 37.71 |
| Males | | | | | |
| US | 1127 | 813.99 | 694.60 | 1027.00 | 44.11 |
| TN | 564 | 826.7 | 488.70 | 1055.62 | 53.55 |
| *Weight* (*kg*) | | | | | |
| Females | | | | | |
| US | 1264 | 68.84 | 39.23 | 156.46 | 17.60 |
| TN | 690 | 73.91 | 37.31 | 143.23 | 15.81 |
| Males | | | | | |
| US | 1127 | 86.24 | 45.80 | 181.41 | 18.00 |
| TN | 564 | 85.57 | 50.01 | 149.73 | 17.28 |
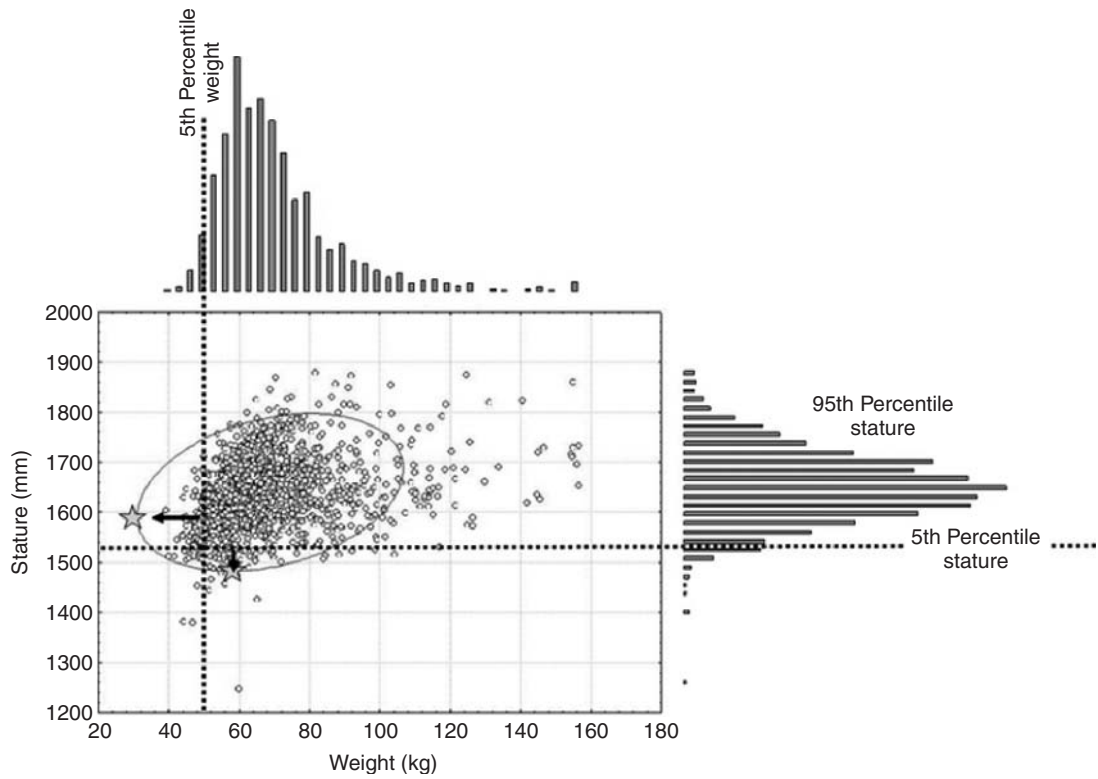
**Figure 6**  Moving the percentile values out farther to get a desired joint accommodation of 90%.

values instead of the 5th and 95th. Each of the five measurements restricts 2% of the population, so at most you would have $5 \times 2\% = 10\%$ disaccommodated. This approach is illustrated in Figure 6. The bars represented by the 5th percentile values would be moved to where the stars are in the figure.

To summarize, percentiles represent the proportion accommodated for one dimension only. When used for more than one dimension, the combination of measurements will accommodate less than the proportion indicated by the percentiles. If the design has no conflicting requirements, you can sometimes compensate by moving out the percentiles. However, if the design has conflicting requirements, using percentiles may accommodate very few people, and an alternative set of cases is required.

## 2.2  Alternative Methods

There are two categories of alternatives to percentiles: (1) use a sample of human subjects as fit models or (2) select a set of cases or representations of people with relevant size and shape combinations. Generally, using a random sample with lots of subjects is not practical, although new three-dimensional modeling and CAD technologies may soon change this. Therefore, the selection of a small number of cases that effectively represent the population is preferable. The first method is typical practice in the apparel industry, but only

for one subject (a sample of one) for a central size referred to as the base size. They adjust for the rest of the population using a process called grading. The grade is expressed as the increment of change from one size to the next starting with the base size for a list of measurements. Therefore grading uses the second method or what we refer to as cases. The success or failure of the fit of the garment for the target population is dependent upon the selection of the initial subject as well as the selection of the cases, the grade increments.

### 2.2.1  Case Selection

The purpose of using a small number of cases is to simplify the problem by reducing to a minimum the amount of information needed. Generally, the first thing reduced is the number of dimensions. This is done using knowledge of the product and by examining the correlation of the dimensions that are related to the product. The goal is to keep just those that are critical and have as little redundant information as possible. For example, eye height, sitting, and sitting height are highly correlated; therefore, accommodating one could accommodate the other, and only one would be needed in case selection. The risk is that something important will be missed and therefore will not be well accommodated.

It is easiest if the number of critical dimensions can be reduced to four or fewer, because all combinations

of small and large proportions need to be considered. If there are two critical dimensions, the minimum number of small and large combinations is four: small–small, small–large, large–small, and large–large. If there are three critical dimensions, the minimum number of small and large combinations is eight: small–large–large, small–small–large, small–large–small, small–small–small, large–small–small, large–small–large, large–large–small, and large–large–large. With more than four the problem gets quite complex, and a random sample may be easier to use.

The next simplification is a reduction in the combinations used. Often in a design, only the small or large size of a dimension is needed. For example, for a chair hip breadth, sitting might be one of the critical dimensions but only the large size is needed to define the minimum width of the seat. If it does not have any interactive or conflicting effect with the other critical dimensions, it can be used as a stand-alone single value. Also, if two or more groups have overlapping cases, such as males and females in some instances, it is possible to drop some of the cases.

This process is best explained using an example of a seated workstation with three critical dimensions: eye height, sitting; buttock–knee length; and hip breadth, sitting. The minimum seat width for this design should be the largest hip breadth, sitting, but this is the only seat element that is affected by hip breadth, sitting, so it interferes with no other dimension. The desired accommodation overall is 90% of the male and female population. First the designer selected the hip breadth, sitting value by examining its summary statistics for the large end of its distribution. These are shown in Table 3. As can be seen from the table, the women have a larger hip breadth, sitting than the men. Therefore,

**Table 3 Hip Breadth, Sitting Statistics from U.S. CAESAR Survey (mm)**

|         | N    | Mean | 95th Percentile | 99th Percentile | Maximum |
|---------|------|------|-----------------|-----------------|---------|
| Females | 1264 | 410  | 501             | 556             | 663     |
| Males   | 1127 | 376  | 435             | 483             | 635     |

the women's maximum value will be used. If the 99th percentile is used, approximately 1% of U.S. civilian women would be estimated to be larger. If 90% are accommodated with the remaining two dimensions, only 89% would be expected to be accommodated for all three. It would be simplest to use the maximum and then accommodate 90% in the other two. This was the approach used by Zehner (1996) for the JPATS aircraft. An alternative is to assume some risk in the design and to select a smaller number than the maximum. This is a judgment to be made by the manufacturer or customer.

Next we examine the two-dimensional (also called *bivariate*) frequency distribution for eye height, sitting and buttock–knee length. The distribution for female subjects from the CAESAR database is shown in Figure 7, and the distribution for male subjects is shown in Figure 8. The stars in Figures 7 and 8 represent the location of the 5th and 95th percentiles, and the probability ellipses enclose 90% of each sample. To achieve the target 90% accommodation, cases that lie on the elliptical boundary are selected. Boundary cases chosen in this way represent extreme combinations of the two measurements. For example, in Figure 7, cases 1 and 3 represent the two extremes for buttock–knee length, and cases 2 and 4 represent the two extremes for eye height, sitting. Note that the cases are moderate
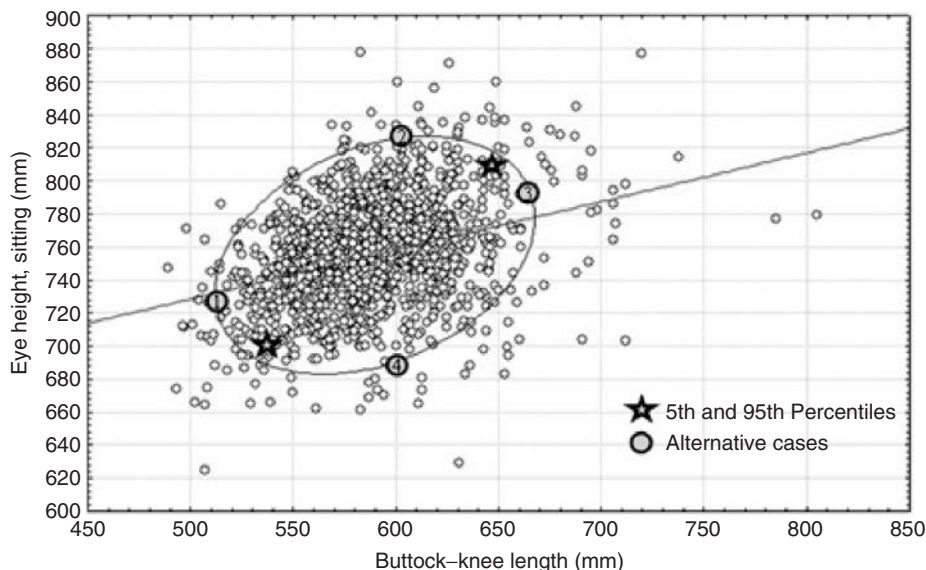


**Figure 7** Bivariate frequency distribution for eye height, sitting and buttock–knee length for CAESAR U.S. female sample (N = 1263) with four female cases shown; 90% probability ellipse.
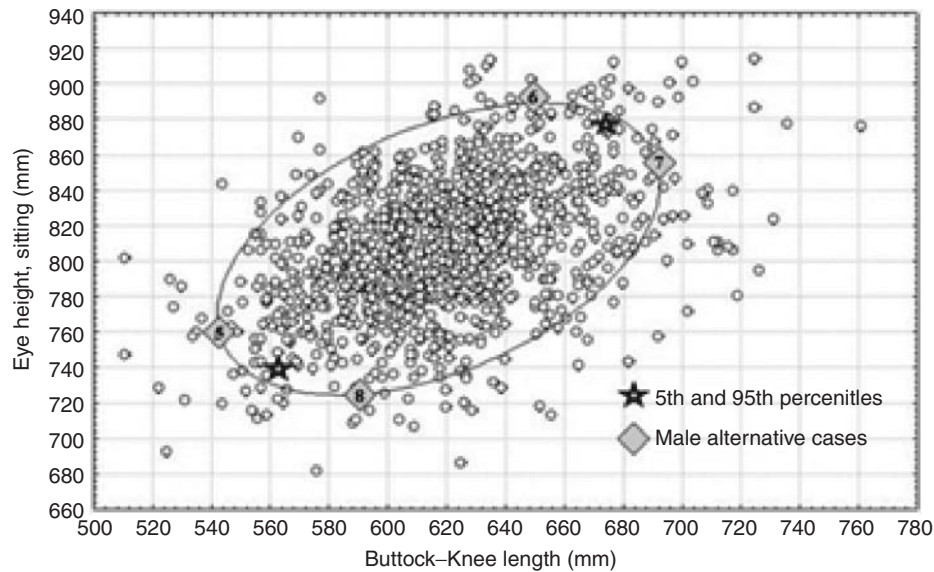
**Figure 8** Bivariate frequency distribution for eye height, sitting and buttock–knee length for U.S. male sample (*N* = 1125); 90% probability ellipse.

in size for one dimension but extreme for the other. The boundary ellipse provides combinations that are not captured in the range between small–small (5th/5th) or large–large (95th/95th) percentiles. Case selection of this type makes the assumption that if the boundary cases are accommodated by the design, so are all those within the probability ellipse. Although this assumption is valid for workspace design, where vision, reach, and clearance from obstruction are key issues, it is not necessarily true for design of clothing or other gear worn on the body. In the latter application, an adequate number of cases must be selected to represent the inner distribution of anthropometric combinations, which should be given much more emphasis than the boundary cases.

The dimensions for the eight cases represented in Figures 7 and 8 are shown in Table 4. Note that this table includes the same hip breadth, sitting for all cases. This is the hip breadth, sitting taken from Table 3 and represents the smallest breadth that should be used in the design. The dimensions for each case must be

applied to the design as a set. For example, the seat must be adjustable to accommodate a buttock–knee length of 510 mm at the same time that it is adjusted to accommodate an eye height, sitting of 725 mm and a hip breadth, sitting of 663 mm to accommodate case 1.

An option for reducing the number of cases is to drop those that are overlapping or redundant. If the risk is so small that differences in men and women will affect the design significantly, it is possible to drop some of the overlapping cases and still accommodate the desired proportion of the population. For example, male cases 5 and 8 are not as extreme as female cases 1 and 4, and the accommodation risk due to dropping them is small. The bivariate distribution in Figure 9 illustrates buttock–knee length and eye height, sitting for both men and women. The final set of anthropometric cases is shown, as well as the location of the dropped cases, 5 and 8.

### 2.2.2 Distributing Cases

As introduced in Section 2.2.1, all of the prior examples make the assumption that if the outer boundaries of the distribution are accommodated, all of the people within the boundaries will also be accommodated. This is true for both the univariate case approach (upper and lower percentile values) and the multivariate case approach (e.g., bivariate ellipse cases, as above). For products that come in sizes or with adjustments that are stepped rather than continuous, this may not be a valid assumption. Imagine a T-shirt that comes in only X-small and XX-large sizes. Few people would be accommodated. For these kinds of products, it is necessary to select, or distribute, cases both *at* and *within* the boundaries.

For distributing cases it is important that there be more cases than expected sizes or adjustment steps to

**Table 4  Case Dimensions for Seated Workstation Example (mm)**

| Females | Case 1 | Case 2 | Case 3 | Case 4 |
|---|---|---|---|---|
| Buttock–knee length | 510 | 600 | 660 | 600 |
| Eye height, sitting | 725 | 820 | 795 | 690 |
| Hip breadth, sitting | 663 | 663 | 663 | 663 |
| Males | Case 5 | Case 6 | Case 7 | Case 8 |
| Buttock–knee length | 541 | 655 | 690 | 595 |
| Eye height, sitting | 760 | 890 | 855 | 725 |
| Hip breadth, sitting | 663 | 663 | 663 | 663 |

ensure that people are not missed between sizes or steps. A good example of distributed cases is shown by Harrison et al. (2000) in their selection of cases for laser eye protection (LEP) spectacles. They used three key dimensions: face breadth for the spectacle width, nose depth for the distance of the spectacle forward from the eye, and eye orbit height for the spectacle height. They used bivariate plots for each of these dimensions with the other two and selected 30 cases to characterize the variability for all three. They also took into account the different ethnicities of subjects when selecting cases to ensure adequate accommodation of all groups. One of their bivariate plots with the cases selected is shown in Figure 10.

For the LEP effort, the critical dimensions were used to select individual subjects, and their three-dimensional scans were used to characterize them as a case for implementation in the spectacle design. Figure 11 illustrates the side view of the three-dimensional scan for one of the cases. By using distributed cases throughout the critical dimension distribution, a broader range is covered than using the equivalent number of subjects in
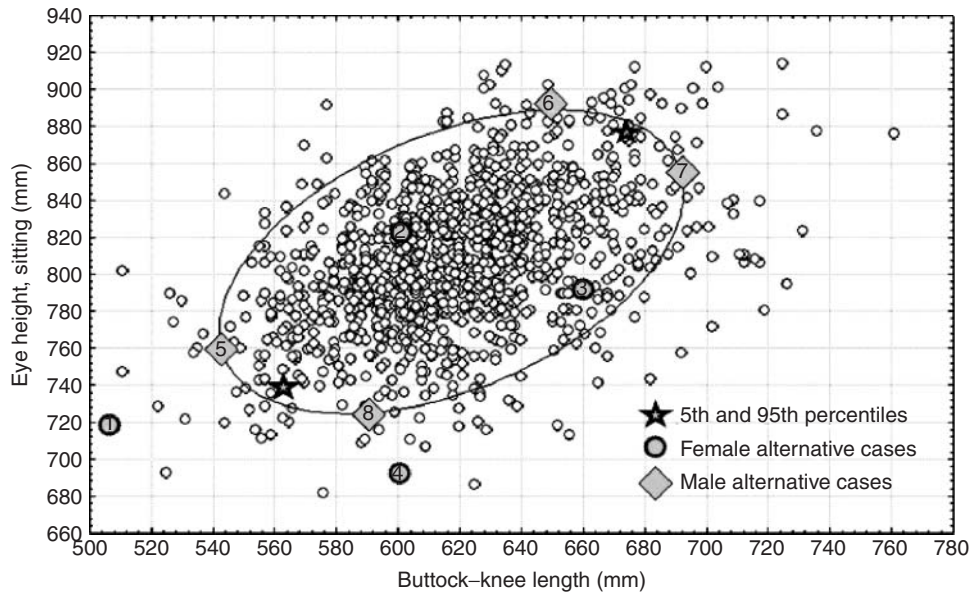


**Figure 9**  Bivariate frequency distribution for eye height, sitting and buttock–knee length for U.S. male ($N = 1127$) and female ($N = 1263$) sample. Cases 5 and 8 were not included in the final set due to proximity to cases 1 and 4.
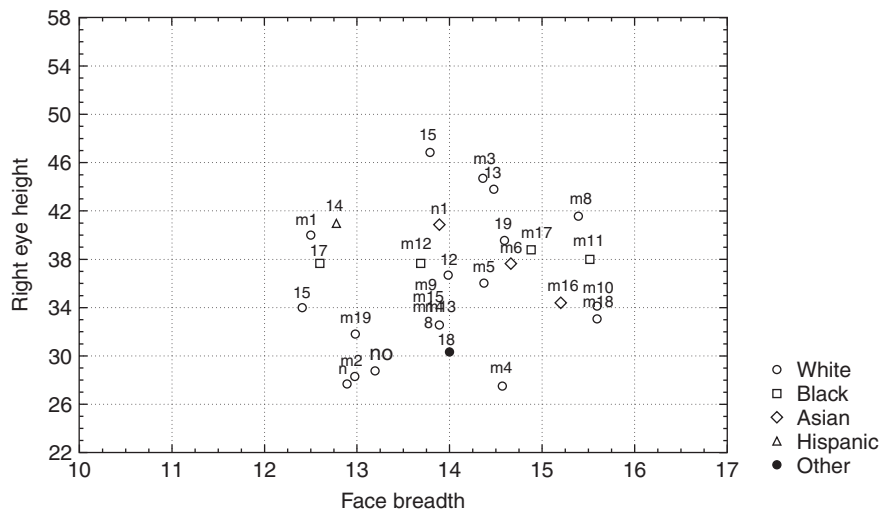


**Figure 10**  Plot 2 of the three critical dimensions and the cases for the LEP.
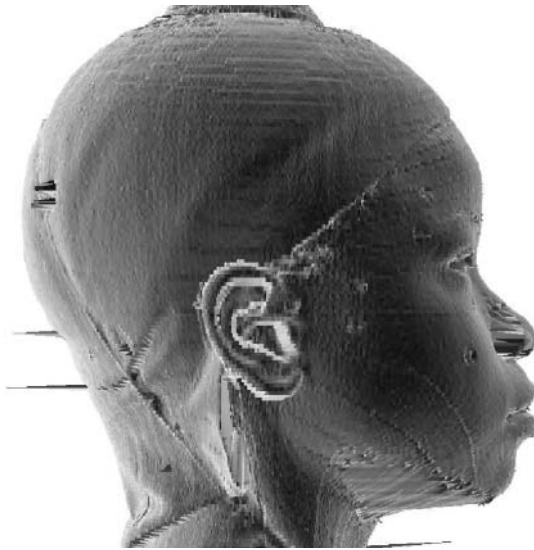
**Figure 11**  Side view of one LEP case.

a random sample, and no assumption about the range of accommodation within one size is made. This permits evaluation of the range of fit within a size and the degree of size overlap during the design process.

### 2.2.3  Principal-Components Analysis

In the examples above, the set of dimensions was reduced using judgment based on knowledge of the problem and the relationship between measurements. Principal-components analysis (PCA) can be helpful in both understanding the relationship between relevant measurements and reducing the set of dimensions to a small, manageable number. This technique has been used effectively for aircraft cockpit crew station design (Bittner et al., 1987; Zehner et al., 1993; Zehner, 1996).

Human dimensions often have some relationship with each other. For example, sitting height and eye height, sitting are highly correlated. The relationships between a set of dimensions can be expressed as either a correlation or a covariance matrix. PCA uses a correlation or covariance matrix and creates a new set of variables called *components*. The total number of components is equal to the number of original variables, and the first component will always represent the greatest amount of variation in the distribution. The second component describes the second greatest, and so on. An examination of the relative contributions, or correlations, of each original dimension and a particular component can be used to interpret and "name" the component. For example, the first component usually describes overall body size and is defined by observing a general increase in the values for the original anthropometric dimensions as the value, or score, of the first component increases.

The premise in using PCA for accommodation case selection is that if most of the total variability in the relevant measurements can be represented in the first

two or three components these components can be used to reduce and simplify your case selection. For example, to write the anthropometric specifications for cockpit design in Joint Primary Air Training System (JPATS) aircraft, Zehner (1996) used the first and second components from a PCA on six cockpit-relevant anthropometric dimensions. The first two components explained 90% of the total variability for all six combined measurements. This was approximately the same for each gender (conducted in separate analyses). Zehner then used a 99.5% probability ellipse on the first two principal components to select the initial boundary cases. One of the genders is shown in Figure 12. Combining the initial set of cases from both genders (with some modification) resulted in a final set of JPATS cases that offered an accommodation of 95% for the women and 99.9% for the men. The first principal component was defined as size; the second was a contrast between limb length and torso height (short limbs/tall torso vs. long limbs/short torso).

Unlike compiled percentile methods (or compiled bivariate approaches when there are more than two variables), multivariate PCA takes into account the simultaneous relationship of three or more variables. However, with PCA the interpretation of the components may not always be clear, and it can be more difficult to understand what aspect of size is being accommodated. An alternative way to use PCA is to use it only to understand which dimensions are correlated with others and then select the most important single dimensions to represent the set as a key dimension. In this way, the key dimension is easier to understand.

The chief limitation of PCA is that all of the dimensions are accepted into the analysis as if they have equal design value and PCA has no way to know the design value. As a result, accommodating the
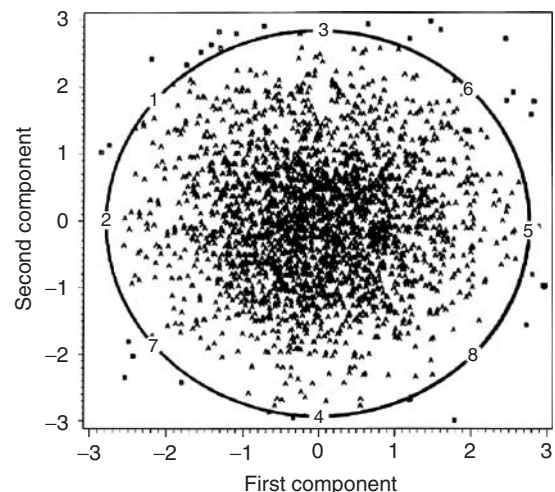


**Figure 12**  PCA bivariate and 99.5% boundary for the first and second principal components for one gender of the JPATS population. Initial cases 1–8 for this gender are regularly distributed around the boundary shape.

components will accommodate some of the variability of the less important dimensions at the expense of the more important ones. Also, PCA is affected by the number of correlated dimensions used of each type. For example, if 10 dimensions are used and 9 of them are strongly correlated with one another, the one dimension that is not correlated with any other may end up being component 4. Since it is one of 10 dimensions, it represents 10% of the total variability. So it is possible to accommodate 90% of the variability in the first three components and not accommodate the most important dimension. Therefore, when using PCA it is important to (1) include only dimensions that are both relevant and important and (2) check the range of accommodation achieved in the cases for each individual dimension.

## 3  FIT MAPPING

Fit mapping is a type of design guidance study that provides information about who a product fits well and who it does not. When anthropometry is used in product design without the knowledge of fit, many speculations must be made about how to place the anthropometry in the design space and the range of accommodation. As a result, even with digital human models and computer-aided design, it is often the case that the first prototypes do not accommodate the full range of the population and may accommodate body size regions that do not exist in the population.

Fit mapping combines performance testing of prototypes or mock-ups with anthropometric measurement to "map" the fit effectiveness of a product for different body sizes and shapes. Fit effectiveness means that the desired population is accommodated without wasted sizes or wasted accommodation regions. Because most performance-based fit tests cannot be done on digital models, fit mapping involves using human subjects to do the assessments. The following is a list of things needed for a fit-mapping study:

1. Human subjects drawn to represent a broad range of variability
2. A prototype or sample of the product (multiple samples of each size is desirable)
3. A testable concept-of-fit definition
4. An expert fit evaluator or one who is trained to be consistent using the concept-of-fit definition
5. Anthropometry measuring equipment
6. Multivariate analysis software and knowledge
7. Survey data from the target population with relevant measurements

The study process consists of:

1. Scoring the fit for each size that the subject can don against the concept of fit
2. Measuring the subjects
3. Analyzing the data to determine:
   a. The key size-determining dimensions
   b. The range of accommodation for each size with respect to the key dimensions
   c. General design or shaping issues
   d. Size or shape gaps in target population coverage
   e. Size or shape overlaps in target population coverage

The end result of the study is that wasted sizes or adjustment ranges are dropped, sizes or adjustment ranges are added where there are gaps, and design and reshaping recommendations are provided to make the product fit better overall. One example of the magnitude of the improvement that can be achieved with the use of fit mapping was demonstrated in the Navy women's uniform study (Mellian et al., 1990; Robinette et al., 1990). The Navy women's uniform consisted of two jackets, two skirts, and two pairs of slacks. The fit mapping consisted of measuring body size and assessing the fit of each of the garments on more than 1000 Navy women. Prior to the study, the Navy had added odd-numbered sizes in an attempt to improve fit, because 75% of all Navy recruits had to have major alterations. The sizes included sizes 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 20, and 22, with three lengths for each, for a total of 42 sizes.

The results indicated three important facts. First, there was 100% overlap in some of the sizes. For each of the items, sizes 7 and 8, 9 and 10, 11 and 12, 13 and 14, and 15 and 16 fit the same subjects equally well. Second, the size of best fit was different for nearly every garment, with some women wearing up to four different sizes. For example, one woman had the best fit in a size 8 for the blue skirt, size 10 for the white skirt, size 12 for the blue slacks, and size 14 for the white slacks. Third, most women did not get an acceptable fit in any size.

The size overlap was examined and it was determined that the difference between the sizes was less than the manufacturing tolerance for a size, which was $\frac{1}{2}$ in. Therefore, the manufacturers had actually used exactly the same pattern for sizes 7 and 8, 9 and 10, 11 and 12, 13 and 14, and 15 and 16. Therefore, sizes 7, 9, 11, 13, and 15 in all three of their lengths could be removed with no effect on accommodation.

The difference in which size fits a given body was resolved by renaming the sizes for some of the garments, to make them consistent. This highlights the fact that the size something is designed to be is not necessarily the size it actually is. Fabric, style, concept of fit, function, and many other factors affect fit. Many of these cannot be known without fit testing on human subjects.

Finally, the women who did not get an acceptable fit in any size were proportioned differently than the size range. They had either a larger hip for the same waist or a smaller hip for the same waist as the Navy size range. This is an example of an interaction or conflict in the dimensions. All of the sizes were in a line consisting of the same shape scaled up and down. This is consistent with common apparel sizing practice. Most apparel companies start with a base size, such as a 10 or a 12, and scale it up and down along a line. The scaling is called *grading*. This is illustrated in Figure 13. The grading line is shown in bold in Figure 13. The sizes

that fall along this line are similar to those used in the Navy women's uniform. Note the overlapping of the odd-numbered sizes with the even-numbered sizes in one area. This is the area where there were more sizes than necessary. Also, note that above and below the grading line, no sizes are available.

Figure 14 illustrates the types of changes made to the sizing to make it more effective. (Note that the sample of women shown is that of the civilian CAESAR survey, not Navy women, who do not have the larger waist sizes.) The overlapping sizes have been dropped. The sizes shown above the grade line have a larger hip for
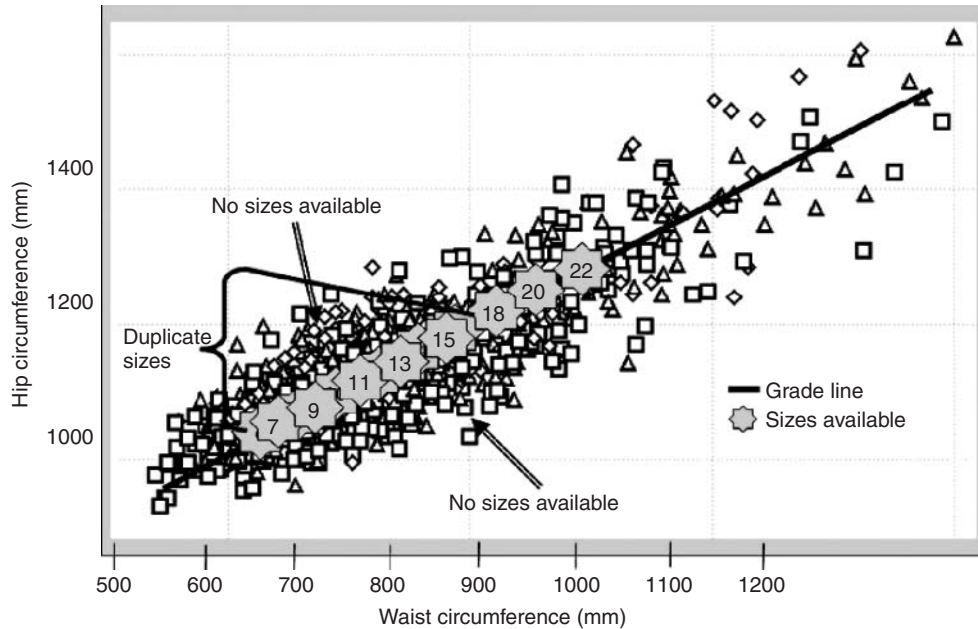


**Figure 13**   Before fit mapping apparel sizes.
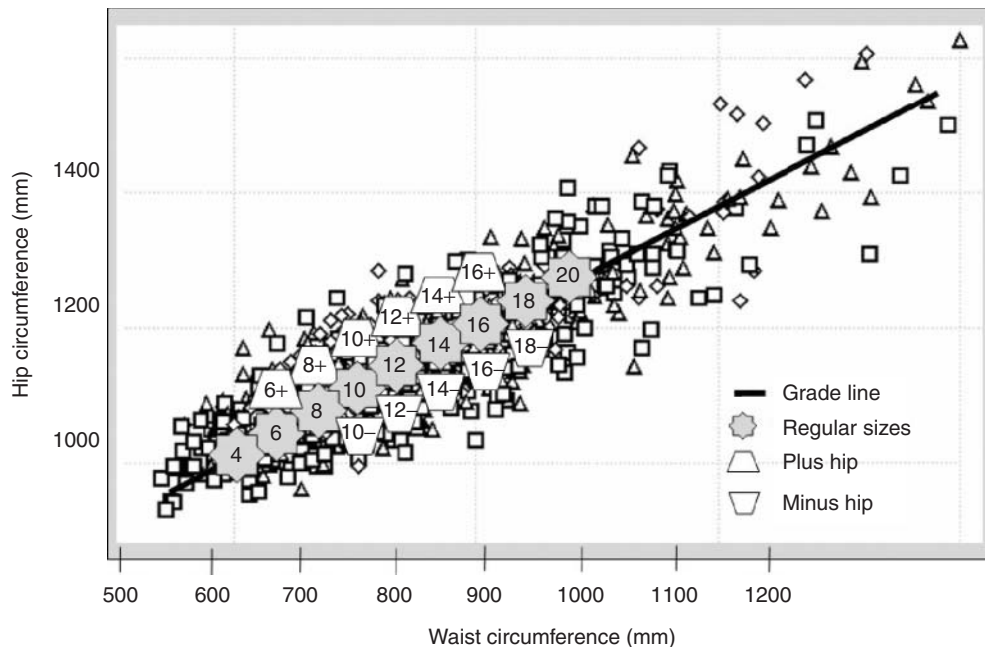


**Figure 14**   After fit-mapping sizes.

the same waist and are called plus hip (+), and the sizes below the line have a smaller hip for the same waist and are called minus hip (−). Before these sizes were added, women who fell in the plus-hip region had to wear a size with a very large waist in order to fit their hip. Then they had to have the entire waist-to-hip region altered. Women who fell in the minus-hip region previously had to get a garment that was way too large in the hips in order to get a fit for their waist. Adding sizes with the modified hip-to-waist proportion resulted in accommodating 99% of the women without needing alternations. The end result of adjusting the sizing based on fit mapping was to improve accommodation from 25 to 99%, with the same number of sizes (Figure 14).

## 4  THREE-DIMENSIONAL ANTHROPOMETRY

Three-dimensional anthropometry has been around since the advent of stereophotography. Originally stereopairs had to be viewed through a stereoviewer and digitized manually, and it was very time consuming. This process is described by Herron (1972). However, digital photography allowed us to automate the process, and this has dramatically affected our ability to design effectively. Automated three-dimensional scanning began to take off in the 1980s (Robinette, 1986). Now there are many tools available to use and analyze three-dimensional scan data, and the first civilian survey to provide whole-body scans of all subjects, CAESAR, was completed in 2002 (Blackwell et al., 2002; Harrison and Robinette, 2002; Robinette et al., 2002). We describe briefly here some of the benefits of the new technology.

### 4.1  Why Three-Dimensional Scans?

By far the biggest advantage of three-dimensional surface anthropometry is visualization of cases, particularly

the ability to visualize them with respect to the equipment or apparel they wear or use. When cases are selected, some assumptions are made about the measurements that are critical for the design. Three-dimensional scans of the subjects often reveal other important information that might otherwise have been overlooked. An example of this is illustrated in Figure 15. When designing airplane, stadium, or theater seats, two common assumptions are made: (1) that the minimum width of the seat should be based on hip breadth, sitting and (2) that the minimum width of the seat should be based on the large male. In Figure 15 we see the scans of two figures overlaid, a male with a 99th percentile hip breadth, sitting and a female with a 99th percentile hip breadth, sitting. The male figure is in dark gray and the female in light gray. It is immediately apparent that the female figure has broader hips than the male. Although she is shorter and has smaller shoulders, her hips are wider by more than 75 mm (almost 3 in.). Second, it is also clear that the shoulders and arms of the male figure extend out beyond the female hips. The breadth across the arms when seated comfortably is clearly a more appropriate measure for the spacing of seats.

For the design of a vehicle interior, measures such as buttock–knee length and eye height, sitting are often considered to be key. Figure 16 shows two women who



**Figure 15**  View of male and female with 99th percentile hip breadth, sitting.
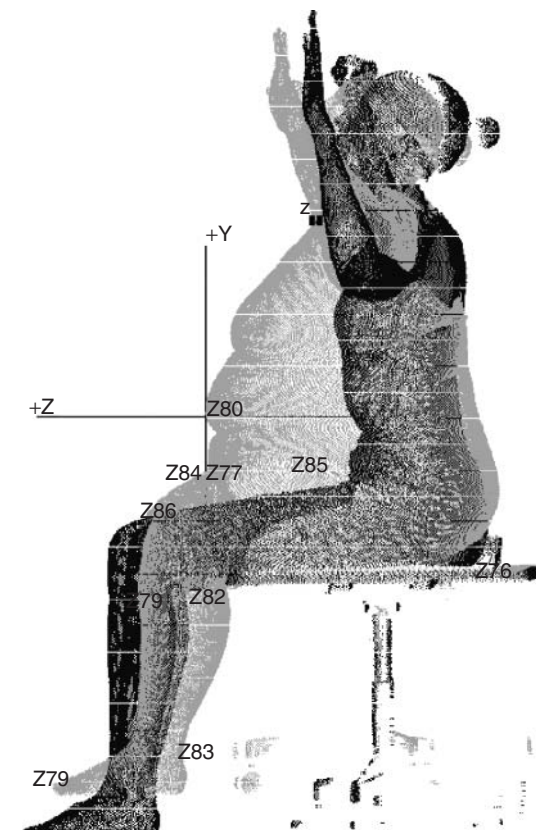


**Figure 16**  Two women with the same buttock–knee length and eye height, sitting.

have the same buttock–knee length and eye height, sitting. However, it is immediately clear from the image that because of the difference in their soft tissue distribution, they would have very different needs in terms of steering wheel placement. These things are much more difficult to comprehend by looking at tables of numbers. Three-dimensional anthropometry captures some measurements, such as contour change, three-dimensional landmark locations, or soft tissue distribution that cannot be captured adequately with traditional anthropometry. Finally, three-dimensional anthropometry offers the opportunity to measure the location of a person with respect to a product for use in identifying fit problems during fit mapping and even for creating custom fit apparel or equipment. For example, by scanning subjects with and without a flight helmet and examining the range of ear locations within the helmet, fit problems due to ear misplacement can be identified. This is illustrated in Figure 17.

Figure 17 shows four examples of using three-dimensional anthropometric measurement to visualize and quantify fit. This figure was created using scans of the subjects with and without the helmet and superimposing the two images in three dimensions using software called Integrate (Burnsides et al., 1996). The image at the upper left of Figure 17 shows the location of the ears of eight subjects in the helmet being tested. The red curved lines show the point at which the subjects complained of ear pain. The image at the lower left shows the locations of two different subjects in the same helmet as they actually wore it to fly, demonstrating different head orientations. The image at the upper right shows the 90 and 95% accommodation ellipses for the point on the ear called the *tragion* for those subjects who did not complain of ear pain. The image at the lower right shows the spread of the *tragion* points for those

subjects who did not complain of ear pain along with one of the subject's ears (subject 4). It can be seen that the points are not elliptical but seem to have a concave shape, indicating a rotational difference between ear locations. These four images together with the fit and comfort evaluations completed by the subjects enable an understanding of the geometry of ear fit in that helmet. Without the three-dimensional images, the fit and comfort scores are difficult to interpret.

The new challenge is to combine static three-dimensional models with human motion. The entertainment industry has been combining these two technologies, but their interest is in rapidly characterizing and sensationalizing the unreal rather than representing truth. Cheng and Robinette (2009) and Cheng et al. (2010) describe the challenge of characterizing true human variability dynamically and present some approaches to addressing the challenge.

## 5  SUMMARY

Whether a product is personal gear (such as clothing or safety equipment), the crew station of a vehicle, or the layout of an office workspace, accommodating the variation in shape and size of the future user population will have an impact on a product's ultimate success. This chapter describes and demonstrates the use of cases, fit mapping, and three-dimensional anthropometry to design effectively, simultaneously minimizing cost and maximizing accommodation. In the section on cases alternatives to the often misused percentiles are discussed, including the use of PCA. The section on fit mapping explains how to incorporate knowledge of the relationship between the human and the product. The best anthropometric data in the world are not
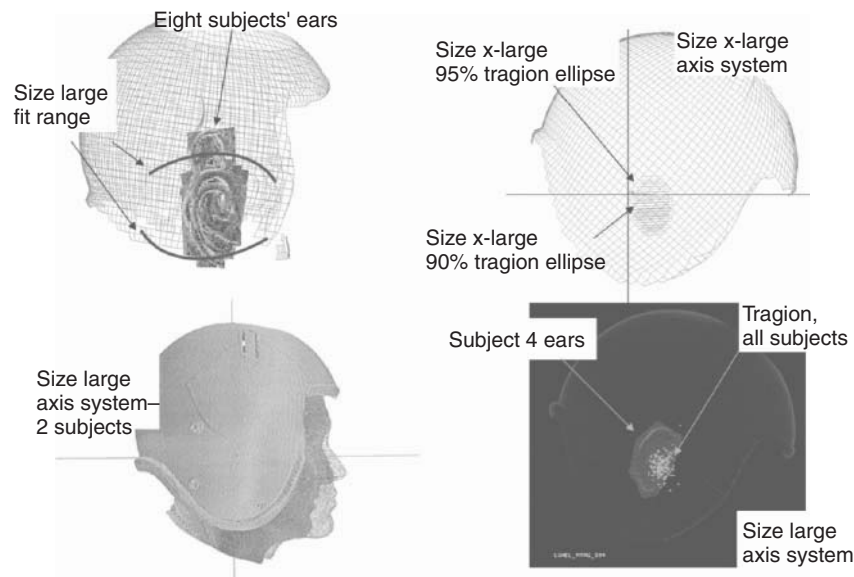


**Figure 17**  Three-dimensional scan visualizations to relate to fit-mapping data for ear fit within a helmet.

sufficient to create a good design if the relationship between the anthropometry and the product proportions that accommodate it is not known. Fit mapping is the study of this relationship. The fit-mapping process is described with examples to demonstrate its benefits.

Finally, for complex multidimensional design problems, three-dimensional imaging technology provides an opportunity to visualize and contrast the variation in a sample and to quantify the differences between locations of a product on subjects who are accommodated versus those who are not. The technology can also be used to capture shape or morphometric data, such as contour change, three-dimensional landmark locations, or soft tissue distribution that cannot be captured adequately with traditional anthropometry. Therefore, three-dimensional anthropometry offers comprehension of accommodation issues to a degree not possible previously.

## REFERENCES

Bittner, A. C., Glenn, F. A., Harris, R. M., Iavecchia, H. P., and Wherry, R. J. (1987), "CADRE: A Family of Mannikins for Workstation Design," in *Trends in Ergonomics/Human Factors*, Vol. 4, S. S. Asfour, Ed., North-Holland/Elsevier, Amsterdam, pp. 733–740.

Blackwell, S., Robinette, K., Daanen, H., Boehmer, M., Fleming, S., Kelly, S., Brill, T., Hoeferlin, D., and Burnsides, D. (2002), *Civilian American and European Surface Anthropometry Resource (CAESAR), Final Report*, Vol. II, *Descriptions*, AFRL-HE-WP-TR-2002-0173, U.S. Air Force Research Laboratory, Human Effectiveness Directorate, Crew System Interface Division, Wright-Patterson Air Force Base, OH.

Burnsides, D. B., Files, P., and Whitestone, J. J. (1996), "Integrate 1.25: A Prototype for Evaluating Three-Dimensional Visualization, Analysis and Manipulation Functionality (U)," Technical Report AL/CF-TR-1996-0095, Armstrong Laboratory, Wright-Patterson Air Force Base, OH.

Cheng, Z., and Robinette, K. (2009), "Static and Dynamic Human Shape Modeling," in *Lecture Notes in Computer Science*, Vol. 5620, pp. 3–12, Springer, New York.

Cheng, Z., Smith, J., Parakkat, J., Cheng, H., and Robinette, K. (2010), "Human Shape Modeling in Various Poses," in *Advances in Applied Digital Human Modeling*, CRC Press, Taylor & Francis Group, New York.

Daniels, G. S. (1952), "The Average Man," TN-WCRD 53–7 (AD 10 203), Wright Air Development Center, Wright-Patterson Air Force Base, OH.

Gordon, C. C., Churchill, T., Clauser, C. C., Bradtmiller, B., McConville, J. T., Tebbets, I., and Walker, R. (1989), "1988 Anthropometric Survey of U.S. Army Personnel: Methods and Summary Statistics," Technical Report NATICK/TR-89/044 (AD A225 094), U.S. Army Natick Research, Development and Engineering Center, Natick, MA.

Harrison, C., Robinette, K., and DeVilbiss, C. (2000), "Anthropometric Variable Selection for Laser Eye Protection Spectacles," in *SAFE Symposium Proceedings 2000, Creswell OR*.

Harrison, C. R., and Robinette, K. M. (2002), "CAESAR: Summary Statistics for the Adult Population (Ages 18–65) of the United States of America," AFRL-HE-WP-TR-2002-0170, U.S. Air Force Research Laboratory, Human Effectiveness Directorate, Crew System Interface Division, Wright-Patterson Air Force Base, OH.

Herron, R. E. (1972), "Biostereometric Measurement of Body Forms," *Yearbook of Physical Anthropology*, Vol. 16, pp. 80–121.

Krul, A. J., Daanen, H. A. M., and Choi, H. (2010), "Self-Reported and Measured Weight, Height and Body Mass Index (BMI) in Italy, the Netherlands and North America," in *European Journal of Public Health*, Vols. 1–6, published by Oxford University Press on behalf of the European Public Health Association, Oxford, United Kingdom.

Mellian, S. A., Ervin, C., and Robinette, K. M. (1990), "Sizing Evaluation of Navy Women's Uniforms," Technical Report 182, Navy Clothing and Textile Research Facility, Natick MA, and AL-TR-1991-0116 (AD A249 782), Armstrong Laboratory, Air Force Systems Command, Wright-Patterson Air Force Base, OH.

Robinette, K. M. (1986), "Three-Dimensional Anthropometry: Shaping the Future," in *Proceedings of the Human Factors Society 30th Annual Meeting*, Vol. 1, Human Factors Society, Santa Monica, CA, p. 205.

Robinette, K. M., and McConville, J. T. (1982), "An Alternative to Percentile Models," SAE Technical Paper 810217, in *1981 SAE Transactions*, Society of Automotive Engineers, Warrendale, PA, pp. 938–946.

Robinette, K. M., Mellian, S., and Ervin, C. (1990), "Development of Sizing Systems for Navy Women's Uniforms," Technical Report 183, Navy Clothing and Textile Research Facility, Natick MA, and AL-TR-1991-0117, Armstrong Laboratory, Air Force Systems Command, Wright-Patterson Air Force Base, OH.

Robinette, K., Blackwell, S., Daanen, H., Fleming, S., Boehmer, M., Brill, T., Hoeferlin, D., and Burnsides, D. (2002), *Civilian American and European Surface Anthropometry Resource (CAESAR), Final Report, Vol. I, Summary*, AFRL-HE-WP-TR-2002-0169, U.S. Air Force Research Laboratory, Human Effectiveness Directorate, Crew System Interface Division, Wright-Patterson Air Force Base, OH.

Searle, J. A., and Haslegrave, C. M. (1969), "Anthropometric Dummies for Crash Research," MIRA Bulletin 5, Monthly Summary of Automobile Engineering Literature, Motor Industry Research Association, Lindley, near Neuneaton, Warwickshire, England, pp. 25–30.

Zehner, G. F. (1996), "Cockpit Anthropometric Accommodation and the JPATS Program," *SAFE Journal*, Vol. 26, No. 3, pp. 19–24.

Zehner, G. F., Meindl, R. S., and Hudson, J. A. (1993), "A Multivariate Anthropometric Method for Crew Stations Design: Abridged," AL-TR-1992-0164, Armstrong Laboratory, Air Force Systems Command, Wright-Patterson Air Force Base, OH.

# CHAPTER 12

# BASIC BIOMECHANICS
# AND WORKSTATION DESIGN

**William S. Marras**
**Biodynamics Laboratory**
**The Ohio State University**
**Columbus, Ohio**

## 1 DEFINITIONS

Occupational biomechanics is an interdisciplinary field in which information from both the biological sciences and engineering mechanics is used to quantify the forces present on the body during work. Biomechanics assumes that *the body behaves according to the laws of Newtonian mechanics*. Kroemer has defined mechanics as "*the study of forces and their effects on masses*" (Kroemer, 1987, p. 170). The object of interest in occupational ergonomics is a quantitative assessment of mechanical loading occurring within the musculoskeletal system. The goal of such an assessment is to quantitatively describe the musculoskeletal loading that occurs during work so that one can derive an appreciation for the degree of risk associated with work-related tasks. This

high degree of precision and quantification is the characteristic that distinguishes occupational biomechanics analyses from other types of ergonomic analyses. Thus, with biomechanical techniques the ergonomics can address the issue of "how much exposure to the occupational risk factors is too much exposure?"

The workplace biomechanical approach is often called industrial or occupational biomechanics. Chaffin et al. (2006) defined occupational biomechanics as "the study of the physical interaction of workers with their tools, machines, and materials so as to enhance the worker's performance while minimizing the risk of musculoskeletal disorders." The current chapter addresses occupational biomechanical issues concepts as they apply to work design.

## 2   ROLE OF BIOMECHANICS IN ERGONOMICS

The approach to a biomechanical assessment is to characterize the human−work system situation through a mathematical representation or model. The model is intended to represent the various underlying biomechanical concepts through a series of rules or equations in a "system" that helps us understand how the human body is affected by the various main effects and interactions associated with risk factor exposure. One can think of a biomechanical systems model as the "glue" that holds our logic together when considering the various factors that would affect risk in a specific work situation.

The advantage of representing the worker in a biomechanical model is that the model permits one to quantitatively consider the *trade-offs* associated with workplace risk factors to various parts of the body in the design of a workplace. It is difficult to accommodate all parts of the body in an ideal biomechanical environment since improving the conditions for one body segment often make things worse for another part of the body. Therefore, the key to the proper application of biomechanical principles is to consider the appropriate biomechanical trade-offs associated with various parts of the body as a function of the work requirements and the various workplace design options and constraints. Ultimately, biomechanical analyses would be most effective in predicting workplace risk during the design stage before the physical construction of the workplace has begun.

This chapter will focus upon the information required to develop proper biomechanical reasoning when assessing physical demands of a workplace. The chapter will first present and explain a series of key biomechanical concepts that constitute the underpinning of biomechanical reasoning. Second, these concepts will be applied to the various parts of the body that are often affected during work. Once this reasoning is established, we will examine how the various biomechanical concepts must be considered collectively in terms of trade-off when designing a workplace from an ergonomic perspective under realistic conditions. The logic in this chapter will demonstrate that one *cannot* successfully practice ergonomics by simply memorizing a set of "ergonomic rules" (e.g., keep the wrist straight or don't bend from the waist when lifting) or applying a generic checklist to a workplace situation. These types of rule-based design strategies often result in suboptimizing the workplace ergonomic conditions or changing workplaces with no payoff.

## 3   BIOMECHANICAL CONCEPTS

### 3.1   Load−Tolerance

A fundamental concept in the application of occupational biomechanics to ergonomics is that one should design workplaces so that the load imposed upon a structure does not exceed the tolerance of the structure (Figure 1). Figure 1 illustrates the traditional concept of biomechanical risk in occupational biomechanics (McGill, 1997). This figure illustrates how a loading pattern is developed on a body structure that is repeated as
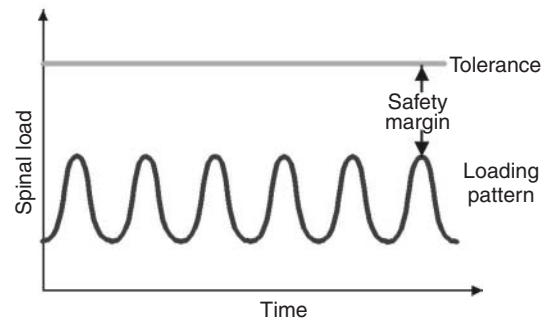


**Figure 1**   Traditional concept of biomechanical risk.

the work cycles recur during a job. Structure tolerance is also shown in this figure. When the magnitude of the load imposed on a structure is less than the tissue tolerance, then the task is considered safe and the magnitude of the difference between the load and the tolerance is considered the safety margin. Implicit in this figure is the idea that risk occurs when the imposed load exceeds the tissue tolerance. While tissue tolerance is defined as the ability of the tissue to withstand a load without damage, ergonomists are beginning to expand the concept of tolerance to include not only mechanical tolerance of the tissue but also the point at which the tissue exhibits an inflammatory reaction.

A recent trend in occupational tasks has been increased repetition while handling lighter loads. The conceptual load−tolerance model can also be adjusted to also account for this type of risk exposure. Figure 2 shows that occupational biomechanics logic can account for this trend by decreasing the tissue tolerance over time. Hence, occupational biomechanics models and logic are moving toward systems that consider manufacturing and work trends in the workplace and attempt to represent these observations (such as cumulative trauma disorders) in the model logic.

### 3.2   Acute versus Cumulative Trauma

In occupational settings two types of trauma can affect the human body and lead to musculoskeletal disorders in occupational settings. First, *acute* trauma can occur
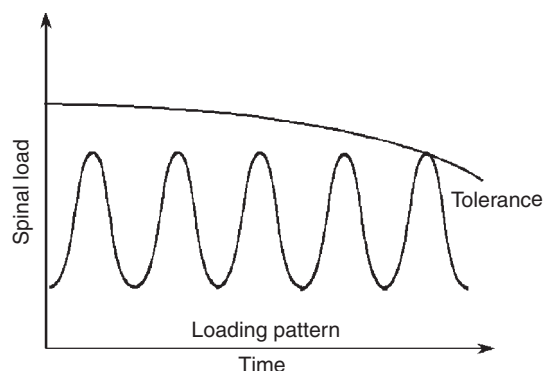


**Figure 2**   Realistic scenario of biomechanical risk.

when a single application of force is so large that it exceeds the tolerance of the body structure during an occupational task. Acute trauma is associated with large exertions of force that would be expected to occur infrequently, such as when a worker lifts an extremely heavy object. This situation would result in a peak load that exceeds the load–tolerance.

*Cumulative* trauma, on the other hand, refers to the repeated application of force to a structure that tends to wear down a structure, thus lowering its tolerance to the point where the tolerance is exceeded through a reduction of this tolerance limit (Figure 2). Cumulative trauma represents more of a "wear and tear" on the structure. This type of trauma is becoming more common in occupational settings as more repetitive jobs requiring lower force exertions become more prevalent in industry.

The cumulative trauma process can initiate a response resulting in a cycle that is extremely difficult to break. As shown in Figure 3, the cumulative trauma process begins by exposing the worker to manual exertions that are either frequent (repetitive) or prolonged. The repetitive application of force can affect either the tendons or the muscles of the body. If the tendons are affected, the tendons are subject to mechanical irritation as they are repeatedly exposed to high levels of tension. Groups of tendons may rub against each other. The physiological response to this mechanical irritation can result in inflammation and swelling of the tendon. The swelling will stimulate the nociceptors surrounding the structure and signal the central control mechanism

(brain) via pain perception that a problem exists. In response to this pain the body attempts to control the problem via two mechanisms. First, the muscles surrounding the irritated area will coactivate in an attempt to stabilize the joint and prevent motion of the tendons. Since motion will further stimulate the nociceptors and result in further pain, motion avoidance is indicative of the start of a cumulative trauma disorder and often indicated when workers shorten their motion cycle and move slower. Second, in an attempt to reduce the friction occurring within the tendon, the body can increase its production of lubricants (synovial fluid) within the tendon sheath. However, given the limited space available between the tendon and the tendon sheath, the increased production of synovial fluid often exacerbates the problem by further expanding the tendon sheath. This action further stimulates the surrounding nociceptors. This initiates a viscous cycle where the response of the tendon to the increased friction results in a reaction (inflammation and the increased production of synovial fluid) that exacerbates the problem (see Figure 3). Once this cycle is initiated, it is very difficult to stop and often anti-inflammatory agents are prescribed in order to break this cycle. The process results in chronic joint pain and a series of musculoskeletal reactions such as reduced strength, reduced tendon motion, and reduced mobility. Together, these reactions result in a functional disability.

Cumulative trauma can also affect the muscles. Muscles are overloaded when they become fatigued. Fatigue lowers the tolerance to stress and can result
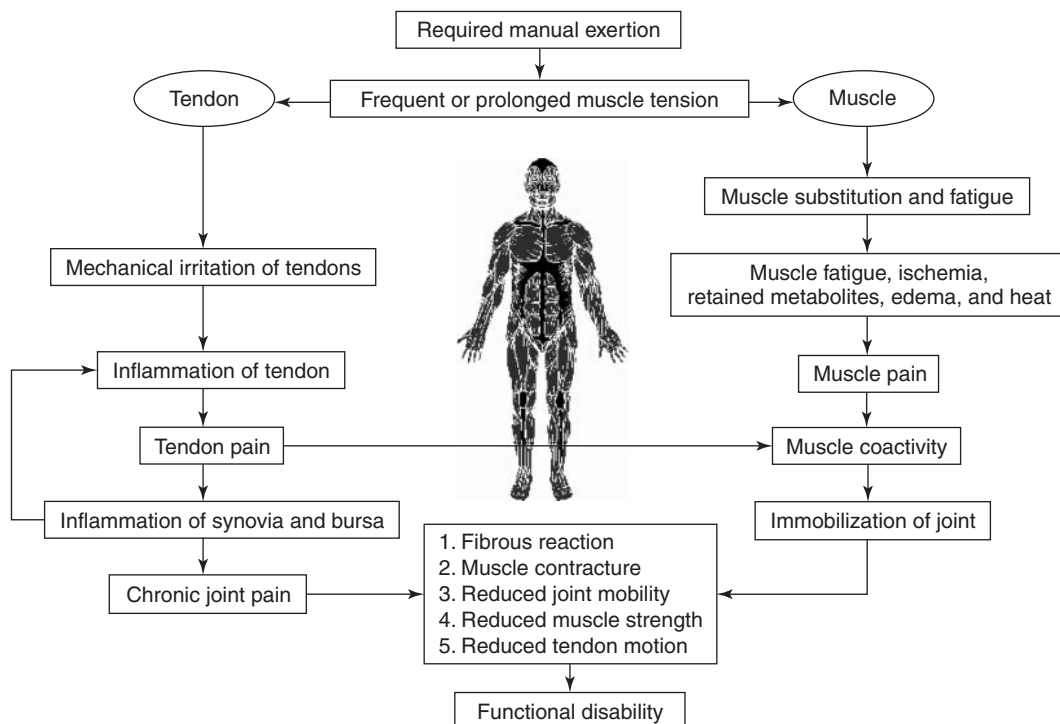


**Figure 3**  Sequence of events in cumulative trauma disorders.

in micro traumato the muscle fibers. This microtrauma typically means the muscle is partially torn, which causes capillaries to rupture and results in swelling, edema, or inflammation near the site of the tear. The inflammation can stimulate nociceptors and result in pain. Once again, the body reacts by cocontracting the surrounding musculature and minimizing the joint motion. However, since muscles do not rely on synovial fluid for their motion, there is no increased production of synovial fluid. However, the end result of this process is the same as that for tendons (i.e., reduced strength, reduced tendon motion, and reduced mobility). The ultimate consequence of this process is, once again, a functional disability.

Although the stimulus associated with the cumulative trauma process is somewhat similar between tendons and muscles there is a significant difference in the time required to heal from the damage to a tendon compared to a muscle. The mechanism of repair for both the tendons and muscles is dependent upon blood flow. Blood flow provides nutrients for repair as well as dissipates waste materials. However, the blood supply to a tendon is a fraction (typically about 5% in an adult) of that supplied to a muscle. Thus, given an equivalent strain to a muscle and a tendon, the muscle will heal rapidly (in about 10 days if not reinjured) whereas the tendon could take months (20 times longer) to accomplish the same level of repair. For this reason, ergonomists must be particularly vigilant in the assessment of workplaces that could pose a danger to the tendons of the body. This lengthy repair process also explains why many ergonomic processes place a high value on identifying potentially risky jobs before a lost-time incident occurs through mechanisms such as discomfort surveys.

### 3.3   Moments and Levers

Biomechanical loads are only partially defined by the magnitude of weight supported by the body. The position of the weight (or mass of the body segment) relative to the axis of rotation of the joint of interest defines the imposed load on the body and is referred to as a *moment*. A moment is defined as the product of force and distance. As an example, a mass of 50 N held at a horizontal distance of 75 cm (0.75 m) from the shoulder joint imposes a moment of 37.5 Nm (50 N × 0.75 m) on the shoulder joint, whereas the same weight held at a horizontal distance of 25 cm from the shoulder joint imposes a moment or load of only 12.5 Nm  (50 N × 0.25 m)  on the shoulder. Thus, the joint load is a function of where the load is held relative to the joint axis and the mass of the weight held. Hence, load is not simply a function of just weight.

As implied in the above example, moments are a function of the mechanical lever systems of the body. In biomechanics, the musculoskeletal system is represented by a system of levers and it is the lever systems that are used to describe the tissue loads with a biomechanical model. Three types of lever systems are common in the human body. First-class levers are those that have a fulcrum placed between the imposed load (on one end of the system) and an opposing force (*internal to the body*)

imposed on the opposite end of the system. The back or trunk is an example of a first-class lever. In this case, the spine serves as the fulcrum. As the human lifts, a moment (*load imposed external to the body*) is imposed anterior to the spine due to the object weight times the distance of the object from the spine. This moment is counterbalanced by the activity of the back muscles; however, they are located in such a way that they are at a mechanical disadvantage since the distance between the back muscles and the spine is much less than the distance between the object lifted and the spine.

A second-class lever system can be seen in the lower extremity. In the second-class lever situation the fulcrum is located at one end of the lever, the opposing force (internal to the body) is located at the other end of the system, and the applied load is in-between these two. The foot is a good example of this lever system. The ball of the foot acts as the fulcrum, the load is applied through the tibia or bone of the lower leg, and the restorative force is applied through the gastrocnemius or calf muscle. The muscle activates and causes the body to rotate about the fulcrum or ball of the foot and moves the body forward.

Finally, a third-class lever is one where the fulcrum is located at one end of the system, the applied load acts at the other end of the system, and the opposing (internal) force acts in between the two. An example of such a lever system in the human body is the elbow joint and is shown in Figure 4.

### 3.4   External versus Internal Loading

Based upon these lever systems, it is evident that two types of forces can impose loads on a tissue during work. *External* loads refer to those forces that are imposed on the body as a direct result of gravity acting upon an external object being manipulated by the worker. For example, Figure 4*a* shows a tool held in the worker's hand that is subject to the forces of gravity. This situation imposes a 44.5-N (10-lb) external load at a distance from the joint of 30.5 cm (12 in.) on the elbow joint. However, in order to maintain equilibrium, this external force must be counteracted by an *internal _force* generated by the muscles of the body. Figure 4*a* also shows that the internal load (muscle) acts at a distance relative to the elbow joint that is much closer to the fulcrum than the external load (tool). Thus, the internal force must be supplied at a biomechanical disadvantage (because of the smaller lever arm) and must be much larger (534 N, or 120 lb) than the external load (44.5 N, or 10 lb) in order to keep the musculoskeletal system in equilibrium. It is not unusual for the magnitude of the internal load to be much greater (often 10 times greater) than the external load. Thus, it is the internal loading that contributes most to cumulative trauma of the musculoskeletal system during work. The net sum of the external load and the internal load defines the total loading experienced at the joint. Therefore, when evaluating a workstation the ergonomist must not only consider the externally applied load but also be particularly sensitive to the magnitude of the internal forces that can load the musculoskeletal system.

**Internal load, F**

$F = 0.0254 \text{ m} = 44.5 \text{ N} = 0.305 \text{ m}$

$= \dfrac{44.5 \text{ N} = 0.305 \text{ m}}{0.0254 \text{ m}}$

$= 534.35 \text{ N (120 lb)}$

Internal load, F

External load

0.0254 m

0.305 m

44.5 N

**Internal load, F**

$F = 0.0127 \text{ m} = 89 \text{ N} = 0.1525 \text{ m}$

$= \dfrac{89 \text{ N} = 0.1525 \text{ m}}{0.0127 \text{ m}}$

$= 1068 \text{ N (240 lb)}$

Internal load, F

0.0127 m

External load

0.1525 m

89 N

(a)

(b)
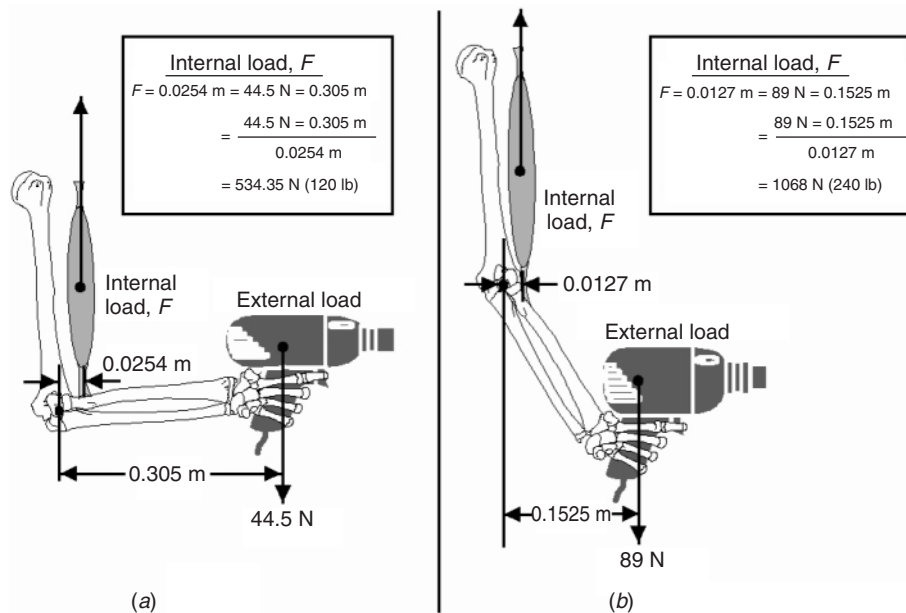
**Figure 4** Example of an anatomical third-class lever (a) demonstrating how the mechanical advantage changes as the elbow position changes (b).

### 3.5 Modifying Internal Loads

The previous section has emphasized the importance of understanding the relationship between the external loads imposed upon the body and the internal loads generated by the force-generating mechanisms within the body. The key to proper ergonomic design is based upon the principle of designing workplaces so that the internal loads are minimized. Internal forces can be thought of as both the component that loads the tissue as well as a structure that can be subject to overexertion. Thus, muscle strength or capacity can be considered as a tolerance measure. If the forces imposed on the muscles and tendons as of a result of the task exceed the strength (tolerance) of the muscle or tendon, a potential injury is possible. Generally, three components of the physical work environment (biomechanical arrangement of the musculoskeletal lever system, length−strength relationships, and temporal relationships) can be manipulated in order to facilitate this goal and serve as the basis for many ergonomic recommendations.

### 3.6 Biomechanical Arrangement of the Musculoskeletal Lever System

The posture imposed via the design of the workplace can affect the arrangement of the body's lever system and thus can affect the magnitude of the internal load required to support the external load. The arrangement of the lever system could influence the magnitude of the external moment imposed upon the body as well as dictate the magnitude of the internal forces and the subsequent risk of either acute or cumulative trauma. If one considers the biomechanical arrangement of the

elbow joint (shown in Figure 4a), it is evident that the mechanical advantage of the internal force generated by the biceps muscle and tendon is defined by a posture keeping one's arm bent at a 90° angle. If one palpates the tendon and inserts the index finger between the elbow joint center and the tendon, one can gain an appreciation for the internal moment arm distance. It is also possible to appreciate how this internal mechanical advantage can change with posture. With the index finger still inserted between the elbow joint and the tendon, if the elbow joint is extended, one can appreciate how the distance between the tendon and the joint center of rotation is significantly reduced. If the imposed moment about the elbow joint is held constant (shown in Figure 4b by a heavier tool) under these conditions, the mechanical advantage of the internal force generator is significantly reduced. Thus, the bicep muscle must generate greater force in order to support the external load. This greater force is transmitted through the tendon and can increase the risk of cumulative trauma. Hence, the positioning of the mechanical lever system (which can be accomplished through work design) can greatly affect the internal load transmission within the body. A task can be performed in a variety of ways, but some of these positions are much more costly in terms of loading of the musculoskeletal system than others.

### 3.7 Optimizing the Length−Strength Relationship

Another important relationship that influences the load on the musculoskeletal system is the length−strength relationship of the muscles. This relationship is shown
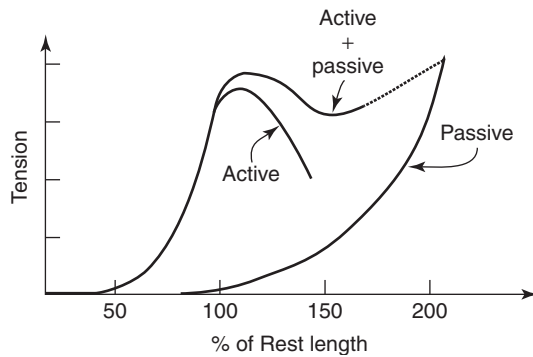
**Figure 5** Length–tension relationship for a human muscle. [Adapted from Basmajian, J. V., and De Luca, C. J. (1985), *Muscles Alive: Their Functions Revealed by Electromyography*, 5th ed., Williams and Wilkins, Baltimore, MD.]

in Figure 5. The active portion of this figure refers to active force–generating structures such as muscles. When muscles are at their resting length (generally seen in the fetal position) they have the greatest capacity to generate force. However, when the muscle length deviates from this resting position, the muscle's capacity to generate force is greatly reduced because the cross-bridges between the components of the muscle proteins become inefficient. When a muscle stretches or when a muscle attempts to generate force while at a short length, the ability to generate force is greatly diminished. As indicated in Figure 5, passive tissues in the muscle (and ligaments) can also generate tension when muscles are stretched. Thus, the length of a muscle during task performance can greatly influence the force available to perform work and can influence risk by altering the available internal force within the system. Therefore, what might be considered a moderate force for a muscle at the resting length can become the maximum force a muscle can produce when it is

in a stretched or contracted position, thus, increasing the risk of muscle strain. When this relationship is considered in combination with the mechanical load placed on the muscle and tendon (via the arrangement of the lever system), the position of the joint arrangement becomes a major factor in the design of the work environment. Typically, the length–strength relationship interacts synergistically with the lever system. Figure 6 indicates the effect of elbow position on the force generation capability of the elbow. The joint position can have a dramatic effect on force generation and can greatly affect the internal loading of the joint and the subsequent risk of cumulative trauma.

### 3.8  Impact of Velocity on Muscle Force

Motion can also influence the ability of a muscle to generate force and, therefore, load the biomechanical system. Motion can be a benefit to the biomechanical system if momentum is properly employed or it can increase the load on the system if the worker is not taking advantage of momentum. This relationship between muscle velocity and force generation is shown in Figure 7. The figure indicates that, in general, the faster the muscle is moving, the greater the reduction in force capability of the muscle. This reduction in muscle capacity can result in the muscle strain that may occur at a lower level of external loading and a subsequent increase in the risk of cumulative trauma. In addition, this effect is considered in dynamic ergonomic biomechanical models.

### 3.9  Temporal Relationships

### 3.9.1  Strength–Endurance

Strengt h must be considered as both an internal force and a tolerance. However, it is important to realize that strength is transient. A worker may generate a great amount of strength during a one-time exertion; however, if the worker is required to exert his or her strength either repeatedly or for a prolonged period of time, the amount of force that the worker can generate can be reduced
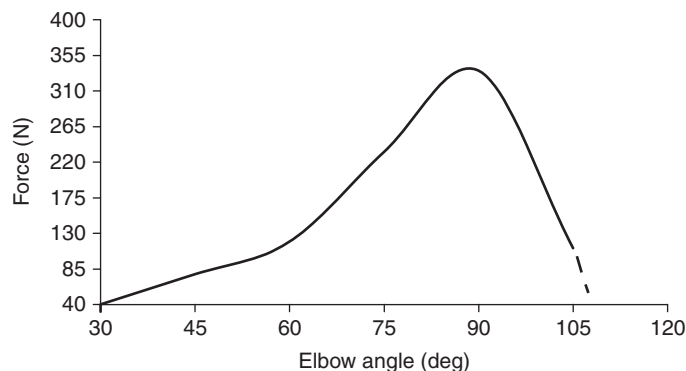


**Figure 6** Position–force diagram produced by flexion of the forearm in pronation. ''Angle'' refers to included angle between the longitudinal axes of the forearm and upper arm. The highest parts of the curve indicate the configurations where the biomechanical lever system is most effective. [Adapted from Chaffin, D. B., and Andersson, G. B. (1991), *Occupational Biomechanics*, JWiley, New York.]
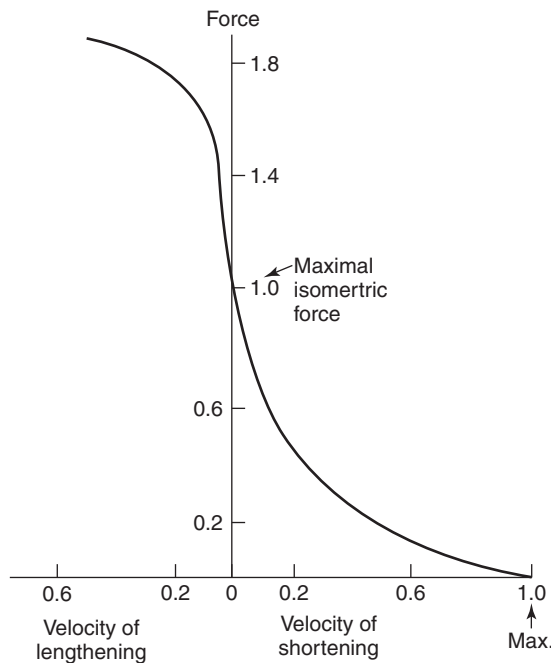
**Figure 7** Influence of velocity upon muscle force. [Adapted from Astrand and Rodahl (1977), *The Textbook of Work Physiology*, McGraw-Hill, New York.]

dramatically. Figure 8 demonstrates this relationship. The dotted line in this figure indicates the maximum force generation capacity of a static exertion over time.

Maximum force is only generated for a very brief period of time. As time advances, strength output decreases exponentially and levels off at about 20% of maximum after about 7 min. Similar trends occur during repeated dynamic conditions. If a task requires a large portion of a worker's strength, one must consider how long that portion of the strength must be exerted in order to ensure that the work does not strain the musculoskeletal system.

### 3.9.2 Rest Time

As discussed earlier, the risk of cumulative trauma increases when the capacity to exert force is exceeded by the force requirements of the job. Another factor that may influence strength capacity (and tolerance to muscle strain) is rest time. Rest time has a profound effect on a worker's ability to exert force. Figure 9 summarizes how energy for a muscular contraction is regenerated during work. Adenosine triphosphate (ATP) is required to produce a power producing muscular contraction. ATP changes into to adenosine diphosphate (ADP) once a muscular contraction has occurred; however, the ADP is not capable of producing a significant muscular contraction. The ADP must be converted to ATP in order to enable another muscular contraction. This conversion to ATP can occur with the addition of oxygen to the system. If oxygen is not available, then the system goes into oxygen debt and insufficient ATP is available for a muscular contraction. Figure 9 indicates that oxygen is a key ingredient in order to maintain a high level of muscular exertion. Oxygen is delivered to the target muscles via the blood. Under static exertions the blood flow is reduced and there is a subsequent reduction in the blood available to the muscle. This restriction of blood flow and subsequent oxygen deficit
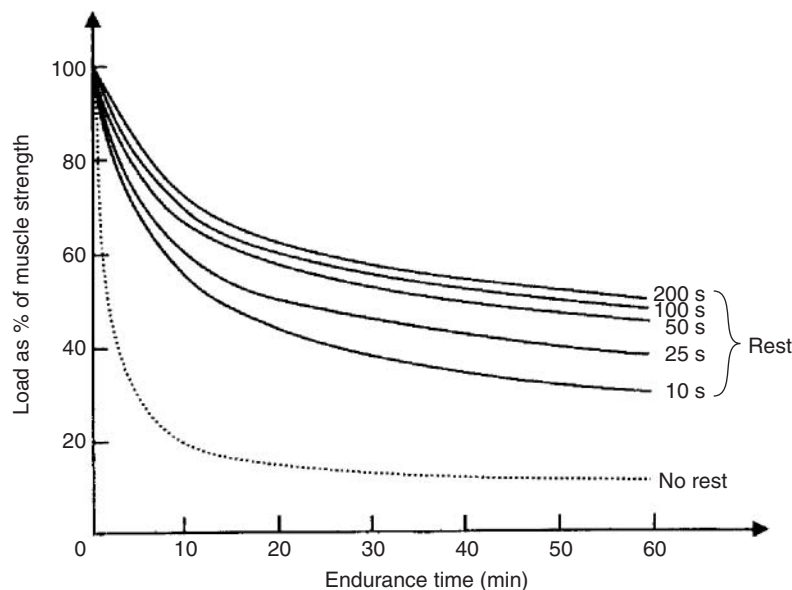


**Figure 8** Forearm flexor muscle endurance times in consecutive static contractions of 2.5 s duration with varied rest periods. [Adapted from Chaffin, D. B., and Andersson, G. B. (1991), *Occupational Biomechanics*, Wiley, New York.]
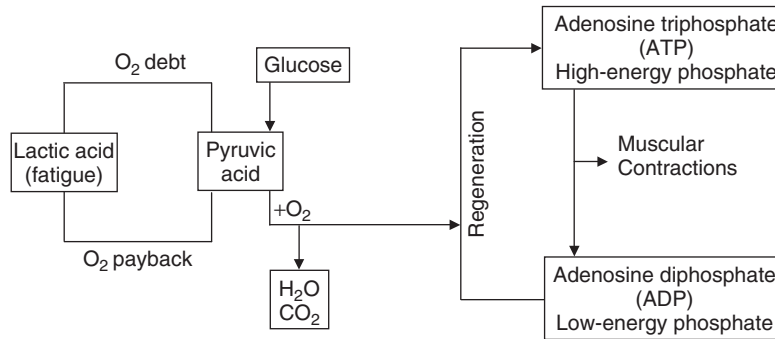
**Figure 9**   Body's energy system during work. [Adapted from Grandjean, E. (1982), *Fitting the Task to the Man: An Ergonomic Approach*, Taylor & Francis, London.]

are responsible for the rapid decrease in force generation over time, as shown in Figure 8. The solid lines in Figure 8 indicate how the force generation capacity of the muscles increases when different amounts of rest are permitted during a prolonged exertion. As more rest time is permitted, increases in force generation are achieved when more oxygen is delivered to the muscle and more ADP can be converted to ATP. This relationship indicates that any more than about 50 s of rest, under these conditions, does not result in a significant increase in force generation capacity of the muscle. Practically, this relationship indicates that in order to optimize the strength capacity of the worker and minimize the risk of muscle strain a schedule of frequent and brief rest periods would be more beneficial than lengthy, infrequent rest periods.

### 3.10   Load Tolerance

Biomechanical analyses must consider not only the loads imposed upon a structure but also the ability of the structure to withstand or tolerate a load during work. This section will briefly review the knowledge base associated with human structure tolerances.

#### 3.10.1   Muscle, Ligament, Tendon, and Bone Capacity

The precise tolerance characteristics of human tissues such as muscles, ligaments, tendons, and bones loaded under various working conditions are difficult to estimate. Tolerances of these structures vary greatly under similar loading conditions. In addition, tolerance depends upon many other factors, such as strain rate, age of the structure, frequency of loading, physiological influences, heredity, conditioning, as well as other, unknown factors. Furthermore, it is not possible to measure these tolerances under in vivo conditions. Therefore, most of the estimates of tissue tolerance have been derived from various animal and/or theoretical sources.

#### 3.10.2   Muscle and Tendon Strain

The muscle is the structure within the musculoskeletal system that has the lowest tolerance. The ultimate strength of a muscle has been estimated to be 32 MPa

(Hoy et al., 1990). In general, it is believed that the muscle will rupture prior to the (healthy) tendon (Nordin and Frankel, 1989) since tendon stress has been estimated at between 60 and 100 MPa (Hoy et al., 1990; Nordin and Frankel, 1989). It is commonly believed that there is a safety margin between the muscle failure point and the failure point of the tendon of about two- (Nordin and Frankel, 1989) to threefold (Hoy et al., 1990). Thus, tendon failure it generally thought to occur at around 60–100 MPa.

#### 3.10.3   Bone Tolerance

Bone tolerances have also been estimated in the literature (Ozkaya and Nordin, 1991). The ultimate stress of bone depends upon the direction of loading. Bone tolerance can range from 51 MPa in transverse tension to over 133 MPa in transverse compression and from 133 MPa in longitudinal loading tension to 193 MPa in longitudinal compression and 68 MPa in shear.

#### 3.10.4   Ligament Tolerance

In general, ultimate ligament stress has been estimated to be approximately 20 MPa. However, ligament properties vary greatly depending on their location within the body. Table 1 shows an overview of these properties as a function of their location. Note the much greater tolerances associated with greater body load bearing.

A strong temporal component to ligament recovery has also been identified. Solomonow found that ligaments require long periods of time to regain structural integrity during which compensatory muscle activities are observed (Solomonow, 2004; Solomonow et al., 1998, 1999, 2000, 2002; Stubbs et al, 1998; Gedalia et al, 1999; Wang et al., 2000). Recovery time has been observed to be several times the loading duration and can easily exceed the typical work–rest cycles observed in industry.

#### 3.10.5   Disc/End-Plate and Vertebrae Tolerance

The mechanism of cumulative trauma to the vertebral disc is thought to be associated with repeated trauma

**Table 1  Range of Ligament Tolerance Characteristics for Different Parts of Body**

| | Ligaments | Stiffness (Force-Deformation Curves) (N/mm) | Range of Modulus of Elasticity (Stress-Strain Linear Region) (MPa) | Cross-Sectional Area (mm$^2$) | Reference |
|---|---|---|---|---|---|
| **Upper Extremity** | Shoulder | 18.25–35.17 | 2.05–178.8 | 3.4–10.7 | Clavert et al., 2009<br>Fremerey et al., 2000<br>Jung et al., 2009<br>Moore et al., 2010<br>Ticker et al., 2006 |
| | Elbow | – | 15.90–20.67 | 8.90–17.01 | Regan et al., 1991 |
| | Forearm | 10.1–16.1 | 447.9–768.3 | 4.2–6.8 | Pfaeffle et al., 1996 |
| | Wrist | 10.5–18.2 | 29.34–46 | 7.5–8.4 | Viegas et al., 1999 |
| | Hand | 24.1–94.5 | – | – | Bettinger et al., 2000<br>Johnston et al., 2004 |
| **Spine** | Cervical | 6.4–32.6 | 2.64–12.8 | 11.1–48.9 | Yoganandan et al., 2000<br>Yoganandan et al., 2001 |
| | Lumbar | 11.5–33.9 | 19.6–120 | 1.6–114.0 | Nachemson and Evans, 1968<br>Pintar et al., 1992 |
| **Lower Extremity** | Hip | 10.4–100.7 | 76.1–285.8 | 13.1–107 | Hewitt et al., 2001<br>Hewitt et al., 2001 |
| | Knee | 214–270 | 322.6–367.4 | – | Jung et al., 2009<br>Butler et al., 1986<br>Quapp and Weiss, 1998<br>Woo et al., 1999 |
| | Ankle | 78–234 | – | – | Beumer et al., 2003<br>Hoefnagels et al., 2007 |
| | Foot | 66.3–189.7 | 5.5–7.4 | 28.2–68.6 | Kura et al., 2001 |

to the vertebral end plate. The end plate is a very thin (about 1-mm-thick) structure that facilitates nutrient flow from the vertebrae to the disc fibers (annulus fibrosis). The disc has no direct blood supply so it relies heavily on this nutrient flow for disc viability. Repeated microfracture of this vertebral end plate is thought to lead to the development of scar tissue which can impair the nutrient flow to the disc fibers. This, in turn, leads to atrophy of the fiber and fiber degeneration. Since the disc contains few nociceptors, the development of microfractures is typically unnoticed by the individual. Given this process, if one can determine the level at which the end plate experiences a microfracture, one can then minimize the effects of cumulative trauma and disc degeneration within the spine.

Several studies of in vitro disc end-plate tolerance have been reported in the literature. Figure 10 indicates the levels of end-plate compressive loading tolerance that have been used to establish safe lifting situations at the worksite [National Institute for Occupational Safety and Health (NIOSH), 1981]. This figure shows the compressive force mean (column value) as well as the compression force distribution (thin line and normal distribution curve) that would result in vertebral end-plate microfracture. The figure indicates that, for those under 40 years of age, end-plate microfracture

damage begins to occur at about 3432 N of compressive load on the spine. If the compressive load is increased to 6375 N, approximately 50% of those exposed will experience vertebral end-plate microfracture. Finally, when the compressive load on the spine reaches a value of 9317 N, almost all of those exposed to the loading will experience a vertebral end-plate microfracture. It is also obvious from this figure that the tolerance distribution shifts to lower levels with increasing age (Adams et al., 2000). In addition, it should be recognized that this tolerance is based upon compression of the vertebral end plate alone. Shear and torsional forces in combination with compressive loading would further lower the tolerance of the end plate.

This vertebral end-plate tolerance distribution has been widely used to set limits for spine loading and define risk. It should also be noted that others have identified different limits of vertebral end-plate tolerance. Jager et al. (1991) have reviewed the spine tolerance literature and suggested different compression value limits. Their spine tolerance summary is shown in Table 2. They have also been able to describe vertebral compressive strength based upon an analysis of 262 values collected from 120 samples. According to their data, the compressive strength of the lumbar spine can
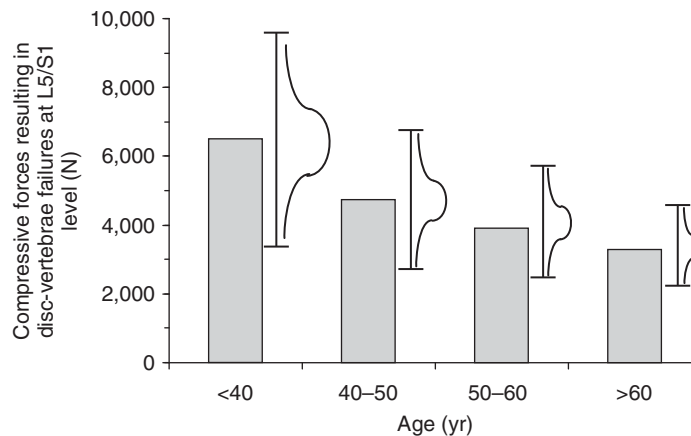
**Figure 10**   Mean and range of disc compression failures by age. [Adapted from National Institute for Occupational Safety and Health (NIOSH) (1981), ''Work Practices Guide for Manual Lifting,'' Publication No. 81-122, Department of Health and Human Services (DHHS), NIOSH, Washington, DC.]

**Table 2  Lumbar Spine Compressive Strength**

| Population | $n$ | Strength, kN | |
| --- | --- | --- | --- |
| | | Mean | s.d. |
| Females | 132 | 3.97 | 1.50 |
| Males | 174 | 5.81 | 2.58 |
| Total | 507 | 4.96 | 2.20 |

*Source*: Jager et al., 1991

be described according to a regression equation:

$$\text{Compressive Strength (kN)} = (7.26 + 1.88G) - 0.494 \\ + 0.468G) \times A + (0.042 + 0.106G) \times C - 0.145 \\ \times L - 0.749 \times S$$

where

$A$ = age in decade
$G$ = gender coded as 0 for female or 1 for male
$C$ = cross-sectional area of vertebrae, cm$^2$
$L$ = lumbar level unit, where 0 is the L5/S1 disc, 1 represents the L5 vertebrae, etc., through 10. which represents the T10/L1 disc
$S$ = structure of interest, where 0 is a disc and 1 is a vertebrae

This equation suggests that the decrease in strength within a lumbar level is about 0.15 kN of that of the adjacent vertebrae and that the strength of the vertebrae is about 0.8 kN lower than the strength of the discs (Jager et al., 1991). This equation can account for 62% of the variability among the samples.

It has also been suggested that spine tolerance limits vary as a function of frequency of loading (Brinkmann et al., 1988). Figure 11 indicates how spine tolerance varies as a function of spine load level and frequency of loading.

Finally, more recent investigations have shown that disc and end-plate tolerances vary greatly with flexion angle of the spine (Callahan and McGill, 2001; Gallagher et al., 2005). These studies have indicated that risk increases sharply at extreme spine flexion. This information suggests that tolerances to mechanical loading drop sharply at the end of the flexion range, especially under dynamic loading conditions.

### 3.10.6  Pain Tolerance

Over the past decade we have learned that there are numerous pathways to pain perception associated with musculoskeletal disorders (Khalsa, 2004; Cavanaugh et al, 1997; Cavanaugh, 1995). It is important to understand these pathways since these pathways may be able to be used as tissue tolerance limits as opposed to tissue damage limits. Hence, one might be able to consider the quantitative limits above which a pain pathway is initiated as a tolerance limit for ergonomic purposes. While none of these pathways have been defined quantitatively, they represent an appealing approach since they represent biologically plausible mechanisms that complement the view of injury association derived from the epidemiological literature.

Several categories of pain pathways are believed to exist that might be used as tolerance limits in the design of the workplace. These categories are (1) structural disruption, (2) tissue stimulation and proinflammatory response, (3) physiological limits, and (4) psychophysical acceptance. Each of these pathways is expected to respond differently to mechanical loading of the tissue and thus serve as tolerance limits. Although many of these limits have yet to be quantitatively defined, current biomechanical research is attempting to define these tolerances, and it is expected that one will be able to one day use these limits to identify the characteristics of a dose–response relationship.
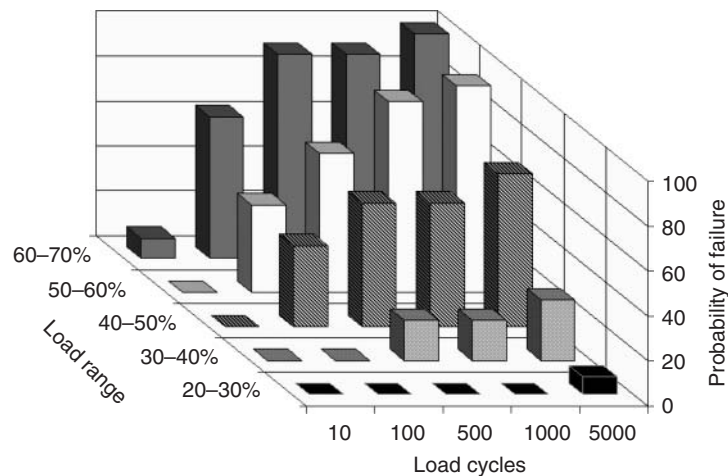
**Figure 11** Probability of a motion segment to be fractured in dependence on the load range and the number of load cycles. [Adapted from Brinckmann, P., Biggemann, M., and Hilweg, D. (1988), ''Fatigue Fracture of Human Lumbar Vertebrae,'' *Clinical Biomechanics*, Vol. 3, Suppl. 1, pp. S1–S23.]

## 4 APPLICATION OF BIOMECHANICAL PRINCIPLES TO REDUCING STRESS IN THE WORKPLACE

These basic concepts and principles of biomechanics can now be applied to workplace design situations. Different body parts, due to differences in structure, are affected by work design in different ways. This section will discuss, in general, how the established biomechanical principles relate to biomechanical loading of the parts of the body often affected by work.

### 4.1 Shoulder

Shoulder pain is believed to be one of the most under-recognized occupationally related musculoskeletal disorders. Shoulder disorders are increasingly being recognized as a major workplace problem by those organizations that have reporting systems sensitive enough to detect such trends. The shoulder is one of the more complex structures of the body with numerous muscles and ligaments crossing the shoulder joint girdle complex. Because of this biomechanical complexity, surgical repair can be problematic. During shoulder surgeries it is often necessary to damage much of the surrounding tissue in an attempt to reach the structure in need of repair. The target structure is often small (e.g., a joint capsule) and difficult to reach. Thus, often damage is done to surrounding tissues that may offset the benefit surgery. Hence, the best course of action is to ergonomically design workstations so that risk of initial injury is minimized.

Since the shoulder joint is biomechanically complex, much of our biomechanical knowledge is derived from empirical evidence. The shoulder represents a statically indeterminate system in that we can typically measure six external moments and forces acting about the point of rotation, yet there are far more internal forces (over 30 muscles and ligaments) that are capable of counteracting the external moments. Thus, quantitative estimates of shoulder joint loading are not common for ergonomic purposes.

When shoulder intensive work is considered, optimal workplace design is typically defined in terms of preferred posture during work. Shoulder *abduction*, defined as the elevation of the shoulder in the lateral direction, is often a problematic posture when work is performed overhead. Figure 12 indicates shoulder performance measures in terms of both available strength and perceived fatigue when the shoulder is held at varying degrees of abduction. The figure indicates that the shoulder can produce a considerable amount of force throughout shoulder abduction angles of between 30° and 90°. However, when comparing reported fatigue at these same abduction angles, it is apparent that fatigue increases rapidly as the shoulder is abducted above 30°. Thus, even though strength is not an issue at shoulder abduction angles up to 90°, fatigue becomes the limiting factor. Therefore, the only position of the shoulder that is acceptable from both a strength and fatigue standpoint is a shoulder abduction of at most 30°.

Shoulder *flexion* has been examined almost exclusively as a function of reported fatigue. Chaffin (1973) has shown that even slight shoulder flexion can influence fatigue of the shoulder musculature. Figures 13 and 14 indicate the effects of vertical and horizontal positioning of the work, respectively, during shoulder flexion while seated, upon fatigability of the shoulder musculature. Fatigue occurs more rapidly as the worker's arm becomes more elevated (Figure 13). This trend is most likely due to the fact that the muscles are deviated from the neutral position as the shoulder becomes more elevated, thus affecting the length–strength relationship (Figure 5) of the shoulder muscles. Figure 14 indicates that as the horizontal distance between the work and the body is increased the time to reach significant fatigue is decreased. This is due to the fact that as a load is held
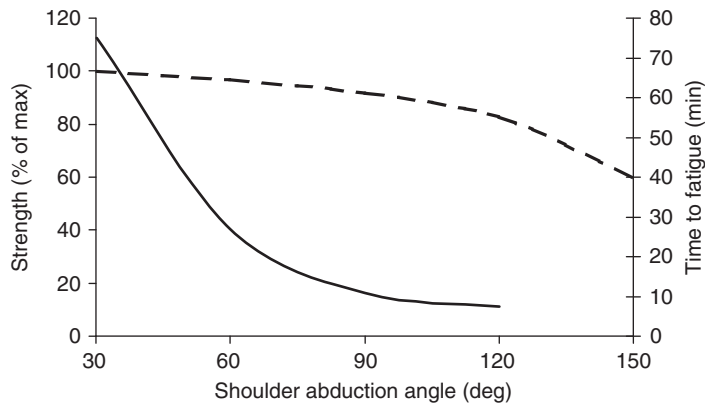
**Figure 12** Shoulder abduction strength and fatigue time as a function of shoulder abducted from the torso. [Adapted from Chaffin, D. B., and Andersson, G. B. (1991), *Occupational Biomechanics*, Wiley, New York.]
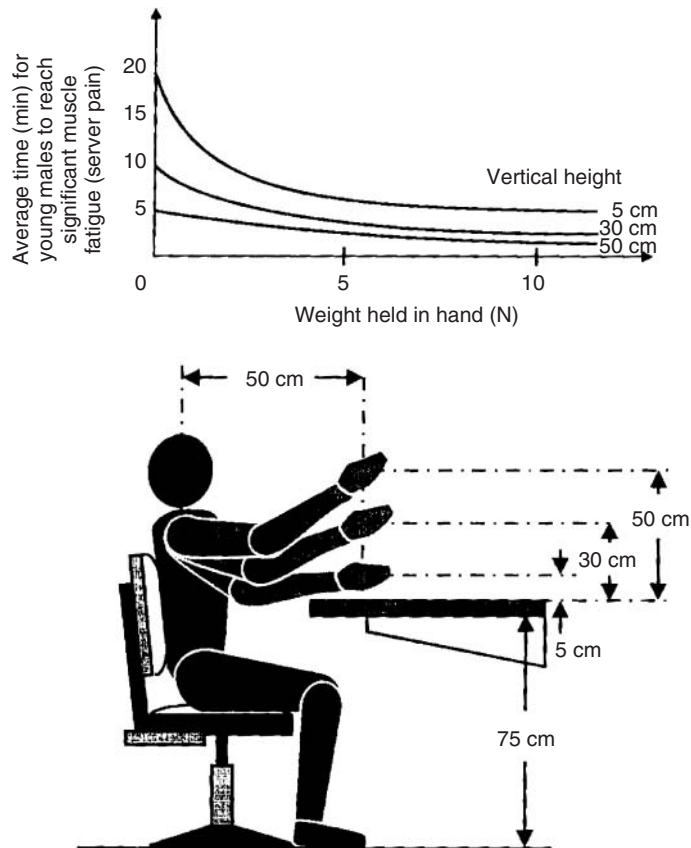


**Figure 13** Expected time to reach significant shoulder muscle fatigue for varied arm flexion postures. [Adapted from Chaffin, D. B., and Andersson, G. B. (1991), *Occupational Biomechanics*, Wiley, New York.]
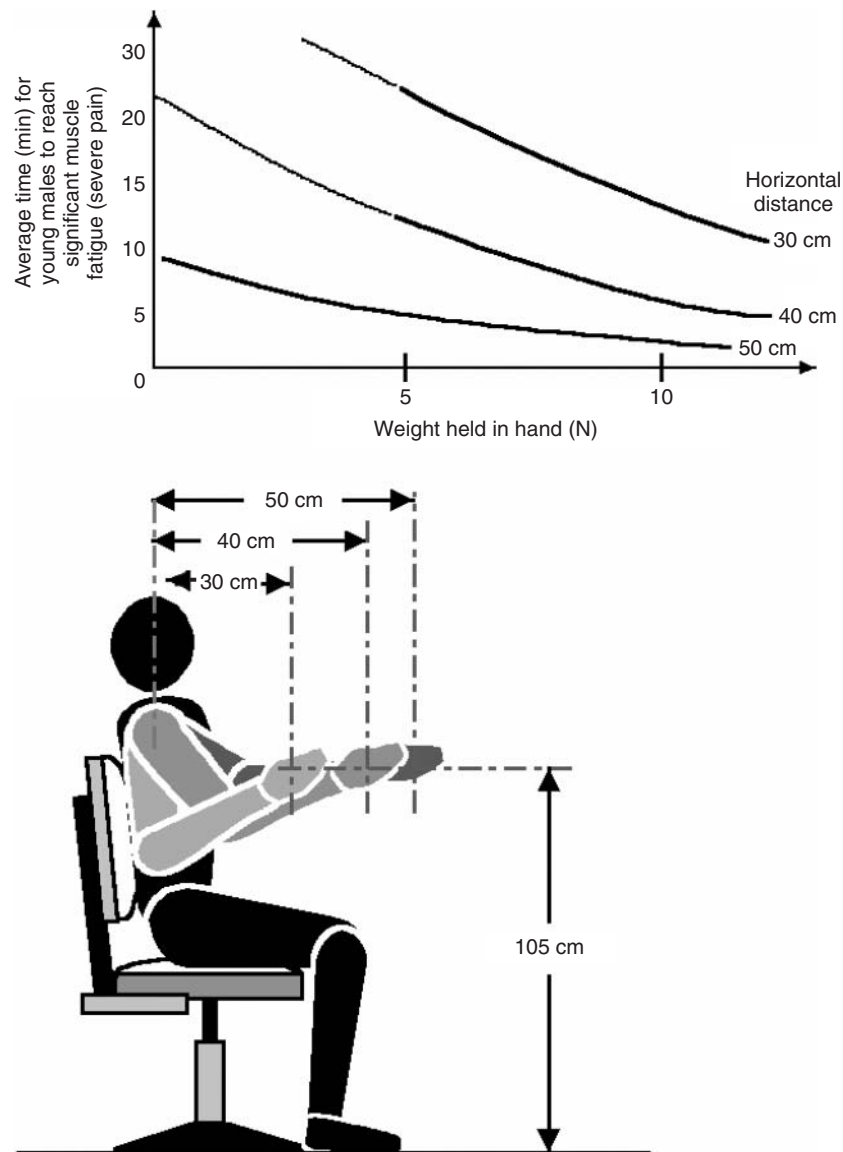
**Figure 14** Expected time to reach significant shoulder muscle fatigue for different forward arm reach postures. [Adapted from Chaffin, D. B., and Andersson, G. B. (1991), *Occupational Biomechanics*, Wiley, New York.]

further from the body more of the external moment (force × distance) must be supported by the shoulder. Thus, the shoulder muscles must produce a greater internal force when the load is held further from the body. With this increased force they fatigue quicker. Elbow supports can significantly increase the endurance time in these postures. In addition an elbow support changes the biomechanical situation by providing a fulcrum at the elbow. Thus, the axis of rotation becomes the elbow instead of the shoulder, and this makes the external moment much less. This not only increase the time one can maintain a posture but also significantly increases the external load one can hold in the hand (Figure 15).

### 4.2 Neck

Neck disorders may also be associated with sustained work postures. Generally, the more upright the posture of the head, the less muscle activity and neck strength are required to maintain the posture. Upright neck positions also have the advantage of reducing the extent of fatigue experienced in the neck (Figure 16). This figure indicates that when the head is tilted forward by $30°$ or more from the vertical position, the time to experience significant neck fatigue decreases rapidly. From a biomechanical standpoint, as the head is flexed, the center of mass of the head moves forward relative to
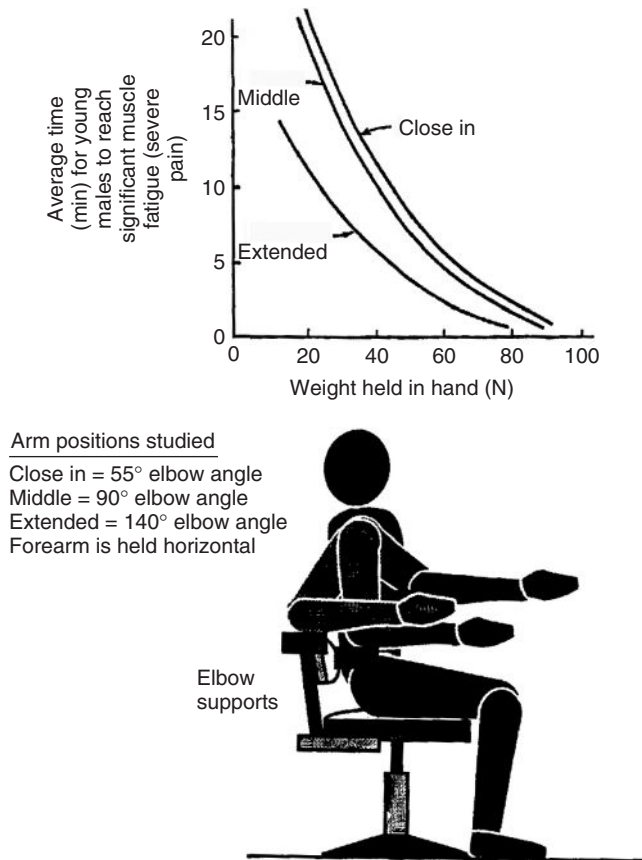
**Figure 15** Expected time to reach significant shoulder and arm muscle fatigue for different arm postures and loads with the elbow supported. The greater the reach, the shorter the endurance time. [Adapted from Chaffin, D. B., and Andersson, G. B. (1991), *Occupational Biomechanics*, Wiley, New York.]

the base of support of the head (spine). Therefore, as the head is moved forward, more of a moment is imposed about the spine and this necessitates increased activation of the neck musculature and greater probability of fatigue (since a static posture is maintained by the neck muscles). On the other hand, when the head is not flexed forward and is relatively upright, the neck can be positioned in such a way that minimal muscle activity is required of the neck muscles and thus fatigue is minimized.

### 4.3 Trade-Offs in Work Design

The key to proper ergonomic design of a workplace from a biomechanical standpoint is to consider the biomechanical trade-offs associated with a particular work situation. These trade-offs are necessary because it is often the case that a situation that is advantageous for one part of the body is disadvantageous for another part of the body. Thus, many biomechanical considerations in the ergonomic design of the workplace require one to consider the various trade-offs and rationales for various design options.

One of the most common trade-off situations encountered in ergonomic design is the trade-off between accommodating the shoulders and accommodating the neck. This trade-off is often resolved by considering the hierarchy of needs required by the task. Figure 17 illustrates this logic. It shows the recommended height of the work as a function of the type of work that is to be performed. Precision work requires a high level of visual acuity that is of utmost importance in order to accomplish the work task. If the work is performed at too low of a level, the head must be flexed in order to accommodate the visual requirements of the job. This situation could result in significant neck discomfort. Therefore, in this situation, visual accommodation is at the top of the hierarchy of task needs and the work is typically raised to a relatively high level (95–110 cm above the floor). This position accommodates the neck but creates a problem for the shoulders since they must be abducted when the work level is high. Thus, a trade-off must be considered. In this instance, ideal shoulder posture is sacrificed in order to accommodate the neck since the visual requirements of the job are great while the shoulder
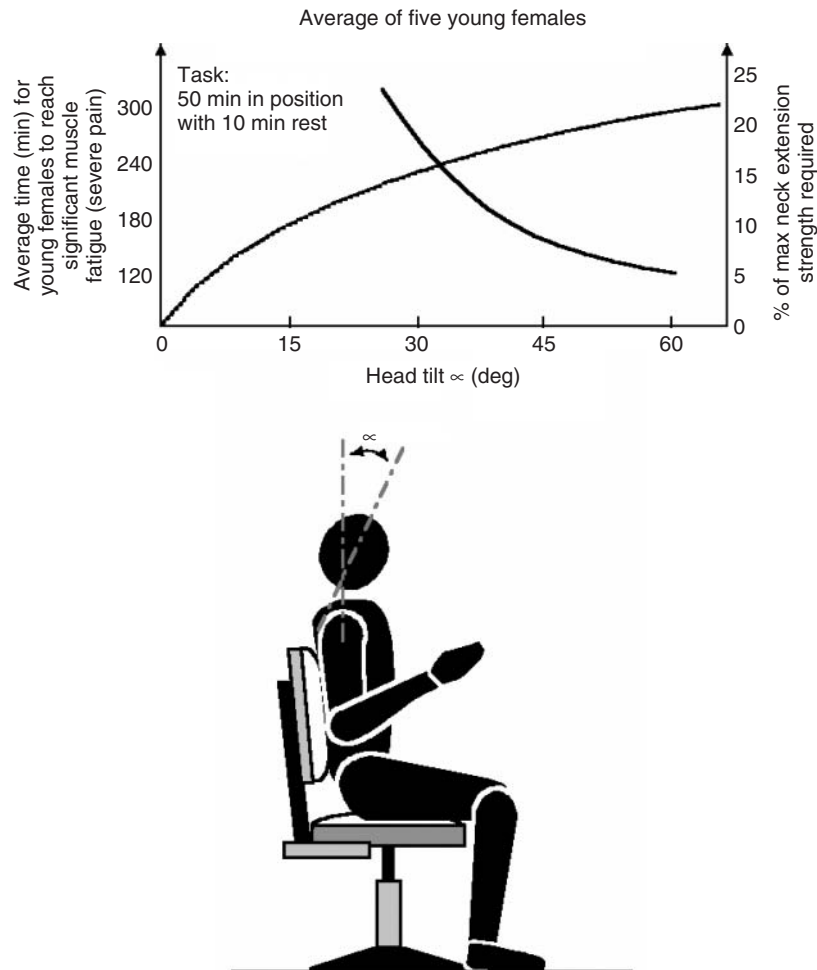
**Figure 16** Neck extensor fatigue and muscle strength required vs. head tilt angle. [Adapted from Chaffin, D. B., and Andersson, G. B. (1991), *Occupational Biomechanics*, Wiley, New York.]

strength required for precision work is low. Thus, visual accommodation is given a higher priority in the hierarchy of task needs. In addition, shoulder disorder risk can be minimized by providing wrist or elbow supports.

The other extreme of the working height situation involves heavy work. The greatest demand on the worker in heavy work is for a high degree of arm strength, whereas visual requirements in this type of work are typically minimal. Thus, the shoulder position is higher on the hierarchy of task needs in this situation. Therefore, in this situation ideal neck posture is typically sacrificed in favor of more favorable shoulder and arm postures. Hence, heavy work is performed at a height of 70–90 cm above floor level. With the work set at this height, the position the elbow angles are close to 90°, which maximizes strength (Figure 6), and the shoulders are close to 30° of abduction, which minimizes fatigue. In this situation, the neck is not in an optimal position,

but the logic dictates that the visual demands of a heavy task would not be substantial and, thus, the neck should not be flexed for prolonged periods of time.

A third work height situation involves light work. Light work is a mix of moderate visual demands with moderate strength requirements. In such a situation, work is a compromise between shoulder position and visual accommodation and neither the visual demands of the job nor the strength requirements dominate the hierarchy of job demands. Both are important considerations. The solution is to minimize the negative aspects of both the strength and neck posture situations by "splitting the difference" between extreme situations. Thus, the height of the work is set at a height between those of the precision work height level and the heavy-work height level. This situation leads to a situation where the work is performed at a level of between 85 and 95 cm off the floor under light-work conditions.
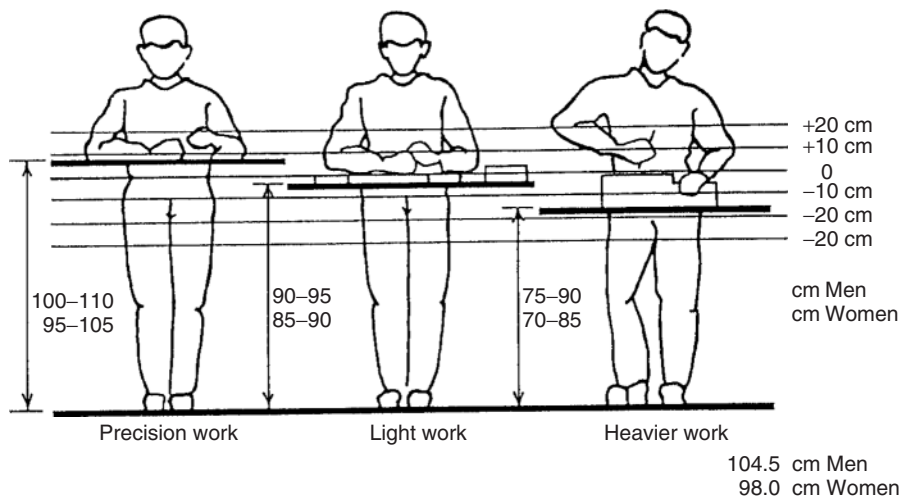
**Figure 17**   Recommended heights of bench for standing work. The reference line (+0) is the height of the elbows above the floor. [From Grandjean, E. (1982), *Fitting the Task to the Man: An Ergonomic Approach*, Taylor & Francis, London. With permission.]

## 4.4  The Back

Low-back disorders (LBDs) have been identified as one of the most common and significant musculoskeletal problems in the United States that results in substantial amounts of morbidity, disability, and economic loss (Hollbrook et al., 1984; Praemer et al., 1992; Guo et al., 1999). Low-back disorders are one of the most common reasons for missing work. Back disorders were responsible for the loss of more than 100 million lost workdays in 1988 with 22 million cases reported that year (Guo et al., 1999; Guo, 1993). Among those under 45 years of age, LBD is the leading cause of activity limitations and it can affect up to 47% of workers with physically demanding jobs (Andersson, 1997). The prevalence of LBD is also on the rise. It has been reported to have increased by 2700% since 1980 (Pope, 1993). Costs associated with LBD are also significant with health care expenditures incurred by individuals with back pain in the United States exceeding $90 billion per year in 1998 (Luo et al., 2004).

It is clear that the risk of LBD is associated with occupational tasks [National Research Council (NRC), 1999, 2001]. Thirty percent of occupation injuries in the United States are related to overexertion, lifting, throwing, holding, carrying, pushing, and/or pulling objects that weigh 50 lb or less. Around 20% of all workplace injuries and illnesses are back injuries which account for up to 40% of compensation costs. Estimates of occupational annual LBD prevalence vary from 1 to 15% depending upon occupation and, over a career, can seriously affect 56% of workers.

Manual materials handling (MMH) activities, specifically lifting, are most often associated with occupationally related LBD risk. It is estimated that lifting and MMH account for up to two-thirds of work-related back injuries (NRC, 2001). Biomechanical assessments target disc-related problems since disc problems are the most serious and costly type of back pain and have

a mechanical origin (Nachemson, 1975). The literature reports increased degeneration in the spines of cadaver specimens who had previously been exposed to physically heavy work (Videman et al., 1990). These findings suggest that occupationally related LBDs are closely associated with spine loading.

### 4.4.1  Significance of Moments

The most important component of occupationally related LBD risk is that of the external moments imposed about the spine (Marras et al., 1993, 1995). As with most biomechanical systems, loading is influenced greatly by the external moment imposed upon the system. However, because of the biomechanical disadvantage at which the torso muscles operate relative to the trunk fulcrum during lifting, very large loads can be generated by the muscles and imposed upon the spine. Figure 18 shows an idealized biomechanical arrangement of lever system. The back musculature is at a severe biomechanical disadvantage in many manual materials-handling situations. Supporting an external load of 222 N (about 50 lb) at a distance of 1 m from the spine imposes a 222-Nm external moment load about the spine. However, since the spine's supporting musculature is at a relatively close proximity relative to the external load, the trunk musculature must exert extremely large forces (4440 N, or 998 lb) to simply hold the external load in equilibrium. The internal loads can increase greatly if dynamic motion of the body is considered (since force is a product of mass and acceleration). Thus, this moment concept dominates risk interpretation in workplace design from a back protection standpoint. Thus, a fundamental issue is to keep the external load's moment arm at a minimum.

A recent study in distribution centers has shown that exposure to large sagittal bending moments when combined with greater lateral spine velocity and exposure to peak moments occurring late in the lift cycle are associated with a significant decrease in spine function over

Internal muscle force, F

$$F \cdot 5 \text{ cm} = 222 \text{ N} \cdot 1 \text{ m}$$

$$= \frac{222 \text{ N} \cdot \text{m}}{0.05 \text{ m}}$$

$$= 4440 \text{ N (998 lb)}$$
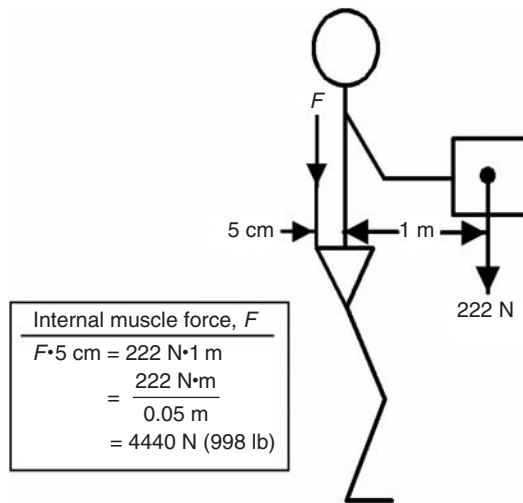
**Figure 18** Internal muscle force required to counterbalance an external load during lifting.

time (Marras et al., 2010). This effort emphasizes the importance of multiple dimensions of risk factors and the interaction of risk factors associated with moment exposure. It is clear from this work that exposure to high load moments in conjunction with dynamic loading factors greatly increases the chances of LBDs due to work tasks.

The concept of minimizing the external moment during lifting has major implications for lifting styles or the best "way" to lift. Since the externally applied moment during a lift significantly influences the internal loading, the lifting style is of far less concern compared to the magnitude of the applied moment. Some have suggested that proper lifting involves lifting "using the legs" or using the "stoop" lift method (bending from the waist). In addition, research has shown that spine load is a function of anthropometry as well as lifting style (Chaffin et al., 2006). Hence, biomechanical analyses (van Dieen et al., 1999; Park and Chaffin, 1974) have demonstrated that no one lift style is correct for all body types. For this reason, the NIOSH (1981) has concluded that lift style need not be a consideration when assessing risk due to materials handling. Some have suggested that the internal moment within the torso is optimized when lumbar lordosis is preserved during the lift (NIOSH, 1981; McGill et al., 2000; McGill, 2002; Anderson et al., 1985). However, from a practical, biomechanical standpoint, the primary indicator of spine loading and, thus, the correct lifting style is whatever style permits the worker to bring the center of mass of the load as close to the spine as possible.

### 4.4.2 Seated versus Standing Workplaces

Seated workplaces have become more prominent with modern work, especially with the aging of the workforce and the introduction of service-oriented and data-processing jobs. It has been documented that loads on the lumbar spine are greater when a worker is seated

compared to standing (Andersson et al., 1975). This is true since the posterior (bony) elements of the spine form an active load path when one is standing. However, when seated these elements are disengaged and more of the load passes through the intervertebral disc. Thus, work performed in a seated position puts the worker at greater risk of spine loading and, therefore, greater risk of damaging the disc. Given this mechanism of spine loading, it is important to consider the design features of a chair since it may be possible to influence disc loading through chair design. Figure 19 shows the results of a study involving pressure measurements taken within the intervetebral disc of individuals as the back angle of the chair and magnitude of lumbar support were varied (Andersson et al., 1975). It is infeasible to directly measure the forces in the spine in vivo. Therefore, disc pressure measures have traditionally been used as a rough approximation of loads imposed upon the spine. This figure indicates that both the seat back angle and lumbar support features have a significant impact on disc pressure. Disc pressure decreases as the backrest angle is increased. However, increasing the backrest angle in the workplace is often not practical since it can also move the worker farther away from the work and thereby increase external moment. Figure 19 also indicates that increasing lumbar support can significantly reduce disc pressure. This reduction in disc pressure is due to the fact that as lumbar curvature (lordosis) is reestablished (with lumbar support) the posterior elements play more of a role in providing an alternative load path, as is the case when standing in the upright position.

Less is known about risk to the low back relative to prolonged standing. The trunk muscles may experience low-level static exertion conditions and may be subject to the static overload through the muscle static fatigue process described in Figure 9. Muscle fatigue can result in lowered muscle force generation capacity and can, thus, initiate the cumulative trauma sequence of events (Figure 3). The fatigue and cumulative trauma sequence can be minimized through two actions. First, foot rails can provide a mechanism to allow relaxation of the large back muscles and thus increased blood flow to the muscle. This reduces the static load and subsequent fatigue in the muscle by the process described in Figure 9. When a leg is rested on the foot rest, the large back muscles are relaxed on one side of the body and the muscle can be supplied with oxygen. Alternating legs placed on the foot rest provides a mechanism to minimize back muscle fatigue throughout the day. Second, floor mats can decrease the fatigue in the back muscles provided that the mats have proper compression characteristics (Kim et al., 1994). Floor mats are believed to facilitate body sway, which enhances the pumping of blood through back muscles, thereby minimizing fatigue.

Knowledge of when standing workplaces are preferable to seated workplaces is dictated mainly by work performance criteria. In general, *standing* workplaces are preferred when (1) the task required a high degree of mobility (when reaching and monitoring in positions that exceed the employee's reach envelope or performing tasks at different heights or different locations), (2) precise manual control actions are not required, (3) leg room is not available (when leg room is not available the moment arm distance between the external load and
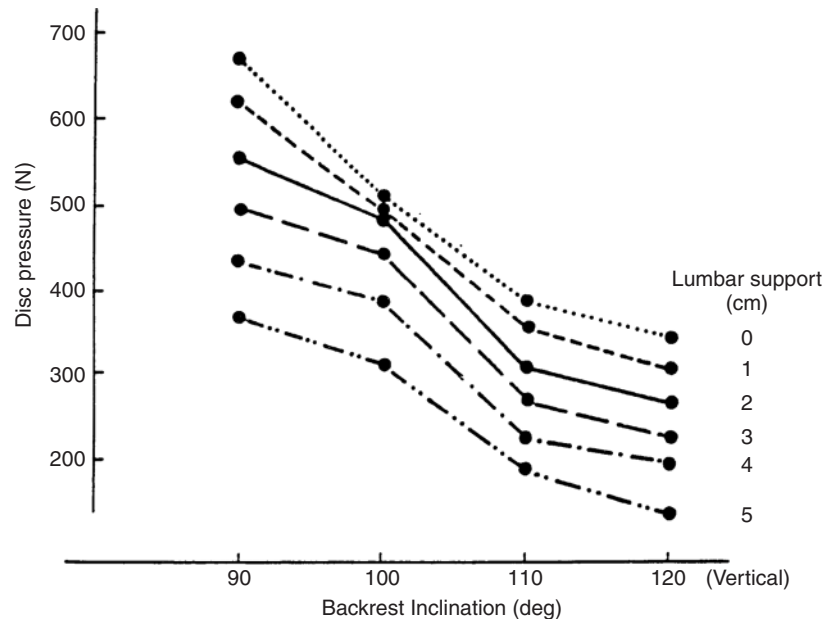
**Figure 19** Disc pressures measured with different backrest inclinations and different-size lumbar supports. [From Chaffin, D. B., and Andersson, G. B. (1991), *Occupational Biomechanics*, Wiley, New York. With permission.]

the back is increased and thus greater internal back muscle force and spinal load result), and (4) heavy weights are handled or large forces are applied. When jobs must accommodate both sitting and standing postures, it is important to ensure that the positions and orientations of the body, especially the upper extremity, are in the same location under both standing and sitting conditions.

### 4.5  Wrists

The Bureau of Labor Statistics reports that repetitive trauma had increased in prevalence from 18% of occupational illnesses in 1981 to 63% of occupational illnesses in 1993. Based upon these figures repetitive trauma has been described as a growing occupationally related problem. Although these numbers and statements appear alarming, one must realize that occupational illnesses represent only 6% of all occupational injuries and illnesses. Furthermore, the statistics for illness include illnesses unrelated to musculoskeletal disorders such as noise-induced hearing loss. Thus, the magnitude of the cumulative trauma problem should not be overstated. Nonetheless, there are specific industries (i.e., meat packing, poultry processing, etc.) where cumulative trauma to the wrist is a major problem and the problem has reached epidemic proportions within these industries.

#### 4.5.1  Wrist Anatomy and Loading

In order to understand the biomechanics of the wrist and how cumulative trauma occurs, one must appreciate the anatomy of the upper extremity. Figure 20 shows a simplified anatomical drawing of the wrist joint complex. The hand has few power-producing muscles in the hand itself. The thenar muscle, which activates
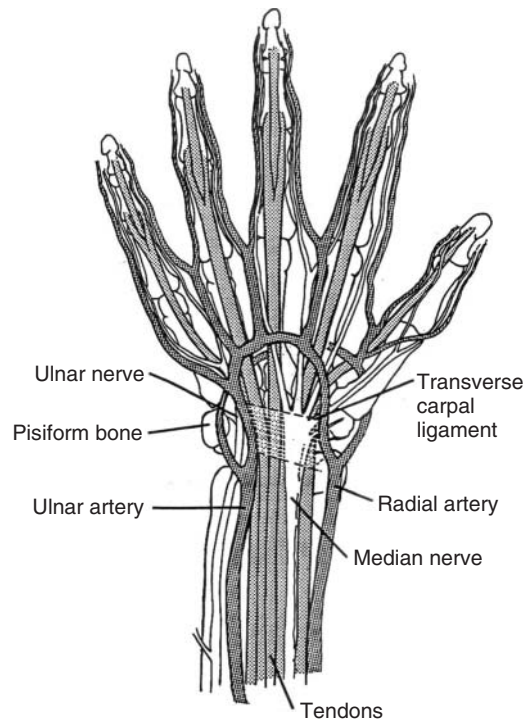


**Figure 20**  Important anatomical structures in the wrist.

the thumb, is one of the few power-producing muscles located in the hand. The vast majority of the power-producing muscles are located in the forearm. Force is

transmitted from these forearm muscles to the fingers through a series of tendons (tendons attach muscles to bone). The tendons originate at the muscles in the forearm, transverse the wrist (with many of them passing through the carpal canal), pass through the hand, and culminate at the fingers. These tendons are secured, or "strapped down," at various points along this path with ligaments that keep the tendons in close proximity to the bones, forming a pulley system around the joints. This results in a system (the hand) that is very small and compact yet capable of generating large amounts of force. However, the price the musculoskeletal system pays for this design is friction. The forearm muscles must transmit force over a long distance in order to supply internal forces to the fingers. Thus, a great deal of tendon travel must occur and this tendon travel can result in tendon friction under repetitive-motion conditions, thereby initiating the events outlined in Figure 3. The key to controlling wrist cumulative trauma is embedded in an understanding of those workplace factors that adversely affect the internal force generating (muscles) and transmitting (tendons) structures.

### 4.5.2 Biomechanical Risk Factors

A number of risk factors for upper extremity cumulative trauma disorders have been documented in the literature. Most of these risk factors have a biomechanical basis for their risk. First, deviated wrist postures reduce the volume of the carpal tunnel and, thus, increase tendon friction. In addition, grip strength is dramatically reduced once wrist posture is deviated from the neutral position. Figure 21 demonstrates the magnitude of grip strength decrement due to any deviation from the wrist's neutral position. The reduction in strength is caused by a change in the length–strength relationship (Figure 5) of the forearm muscles when the wrist is deviated from the neutral posture. Hence, the muscles must work at level lengths that are nonoptimal when the wrist is

bent. This reduced strength associated with deviated wrist positions can, therefore, more easily initiate the sequence of events associated with cumulative trauma (Figure 3). Therefore, deviated wrist postures not only increase tendon travel and friction but also increase the amount of muscle strength necessary to perform the gripping task.

Second, increasing the frequency or repetition of the work cycle has also been identified as a risk factor for cumulative trauma disorders (CTDs) (Silverstein et al., 1996, 1997). Studies have shown that increased frequency of wrist motions increases the risk of cumulative trauma disorder reporting. Repeated motions requiring a cycle time of less than 30 s are considered candidates for cumulative trauma. Increased frequency is believed to increase the friction within the tendons, thereby accelerating the cumulative trauma progression described in Figure 3.

Third, the force applied by the hands and fingers during a work cycle has been identified as a cumulative trauma risk factor. In general, the greater the force required by the work, the greater the risk of CTD. Greater hand forces result in greater tension within the tendons and greater tendon friction and tendon travel. Another factor related to force is that of wrist acceleration. Industrial surveillance studies report that repetitive jobs resulting in greater wrist acceleration are associated with greater CTD incident rates (Schoenmarklin et al., 1994; Marras and Schoenmarklin, 1993). Force is a product of mass and acceleration. Thus, jobs that increase the angular acceleration of the wrist joint result in greater tension and force transmitted through the tendons. Therefore, wrist acceleration can be another mechanism to impose force on the wrist structures.

Finally, as shown in Figure 20, the anatomy of the hand is such that the median nerve becomes very superficial at the palm. Direct impacts to the palm through pounding or striking an object (with the palm)
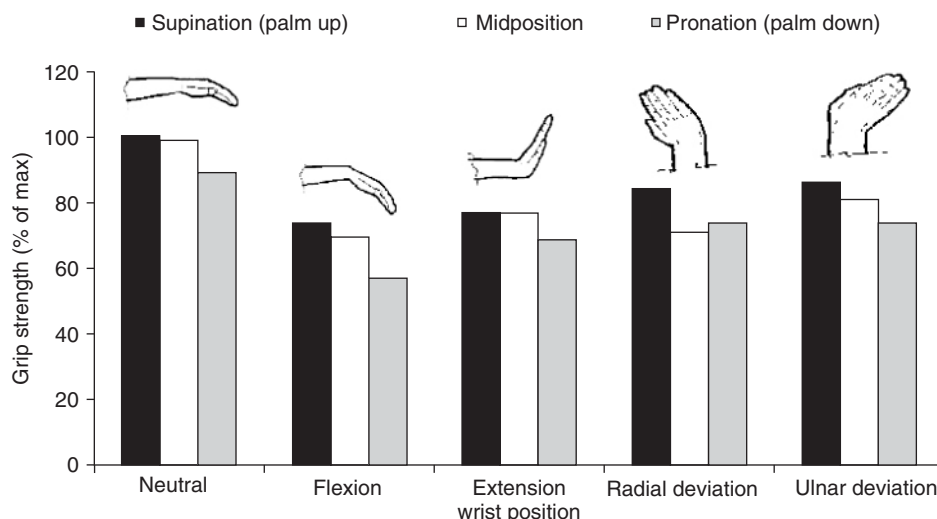


**Figure 21** Grip strength as a function of wrist and forearm position. [Adapted from Sanders, M. S., and McCormick, E. F. (1993), *Human Factors in Engineering and Design*, McGraw-Hill, New York.]
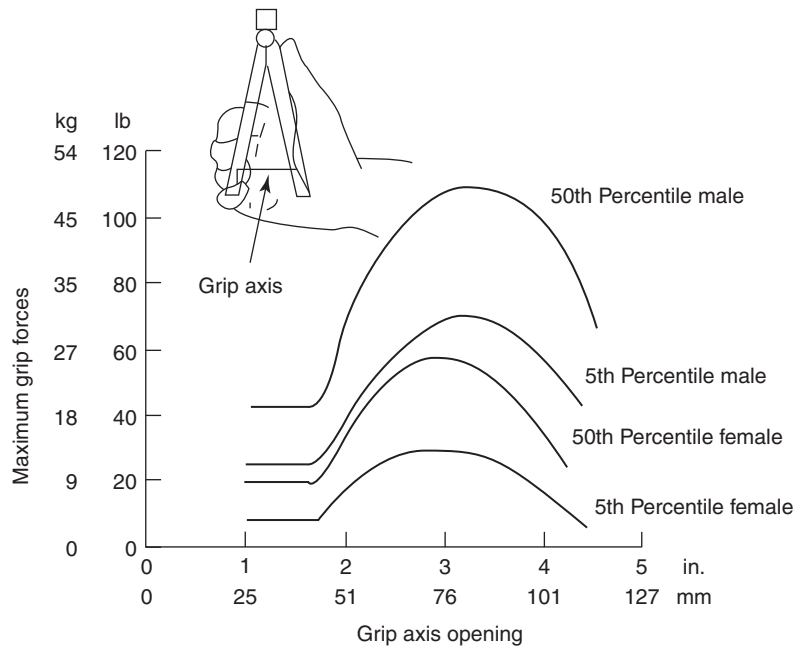
**Figure 22**  Grip strength as a function of grip opening and hand anthropometry. [Adapted from Sanders, M. S., and McCormick, E. J. (1993), *Human Factors in Engineering and Design*, McGraw-Hill, New York.]

can directly assault the median nerve and initiate symptoms of cumulative trauma even though the work may not be repetitive.

### 4.5.3  Grip Design

The design of a tool's gripping surface can impact the activity of the internal force transmission system (tendon travel and tension). Grip opening and shape have a major influence on the available grip strength. Figure 22 indicates how grip strength capacity changes as a function of the separation distance of the grip opening. This figure indicates that maximum grip strength occurs within a very narrow range of grip span. If the grip opening deviates from this ideal range by as little as an inch (a couple of centimeters), then grip strength is markedly reduced. This reduction in strength is, once again, due to the length–strength relationship of the forearm muscles. Also indicted in Figure 22 are the effects of hand size. The worker's hand anthropometry as well as hand preference can influence grip strength and risk. Therefore, proper design of tool handles is crucial in optimizing ergonomic workplace design.

Handle shape can also influence the strength of the wrist. Figure 23 shows how changes in the design of screwdriver handles can impact the maximum force that can be exerted on the tool. The biomechanical origin of these differences in strength capacity is believed to be related to the length–strength relationship of the forearm muscles as well as contact area with the tool. The handle designs resulting in diminished strength permit the wrist to twist or the grip to slip, resulting in a deviation from the ideal length–strength position in the forearm muscles.
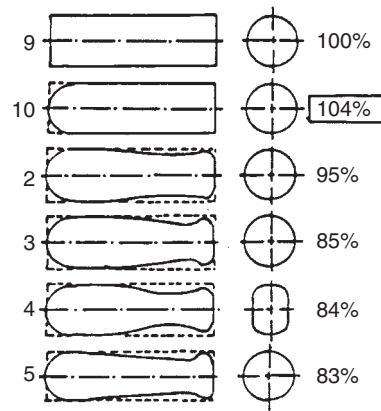


**Figure 23**  Maximum force which could be exerted on a screwdriver as a function of handle shape. [From Konz, S. A. (1983), *Work Design: Industrial Ergonomics*, 2nd ed., Grid Publishing, Columbus, OH. With permission.]

### 4.5.4  Gloves

The use of gloves can also significantly influence the generation of grip strength and may play a role in the development of cumulative trauma disorders. When gloves are worn during work, three factors must be considered. First, the grip strength that is generated is often reduced. Typically, a 10–20% reduction in grip strength is noted when gloves are worn. Gloves reduce the coefficient of friction between the hand and the tool, which in turn permit some slippage of the hand upon the

tool surface. This slippage may result in a deviation from the ideal muscle length and thus a reduction in available strength. The degree of slippage and the subsequent degree of strength loss depend upon how well the gloves fit the hand as well as the type of material used in the glove. Poorly fitting gloves are likely to result in greater strength loss. Figure 24 indicates how the glove material and the glove fit can influence grip force potential.

Second, while wearing gloves, even though the externally applied force (grip strength) is often reduced, the internal forces are often very large relative to a bare-hand condition. For a given grip force application, the muscle activity is significantly greater when using gloves compared to a bare-handed condition (Kovacs et al., 2002). Thus, the musculoskeletal system is less efficient when wearing a glove due to the fact that the hand typically slips within the glove, thereby altering the length–strength relationship of the muscle.

Third, the ability to perform a task is significantly affected when wearing gloves. Figure 25 shows the
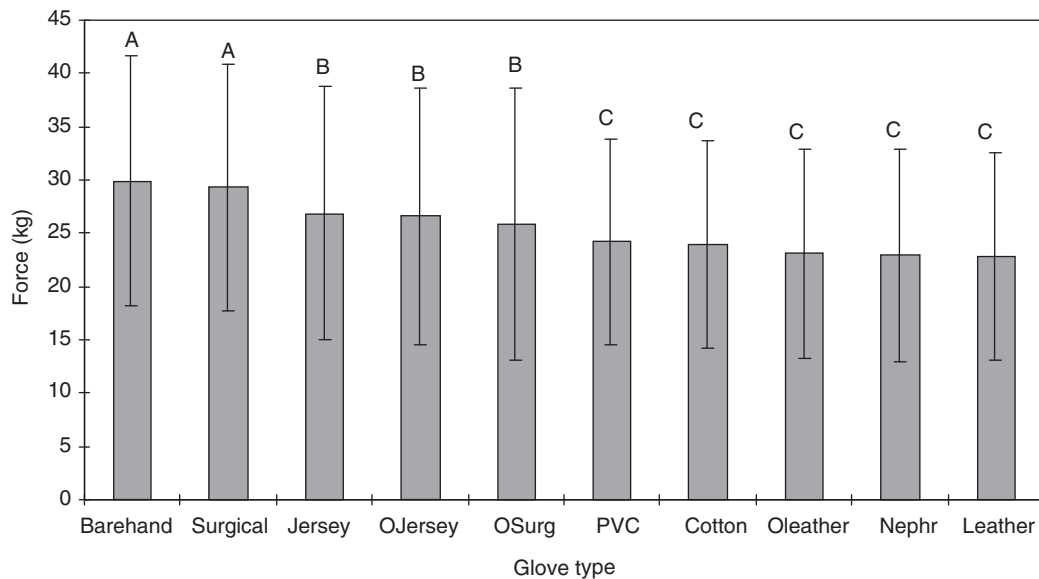


**Figure 24**  Peak grip force shown as a function of type of glove. Different letters above the columns indicate statistically significant differences.
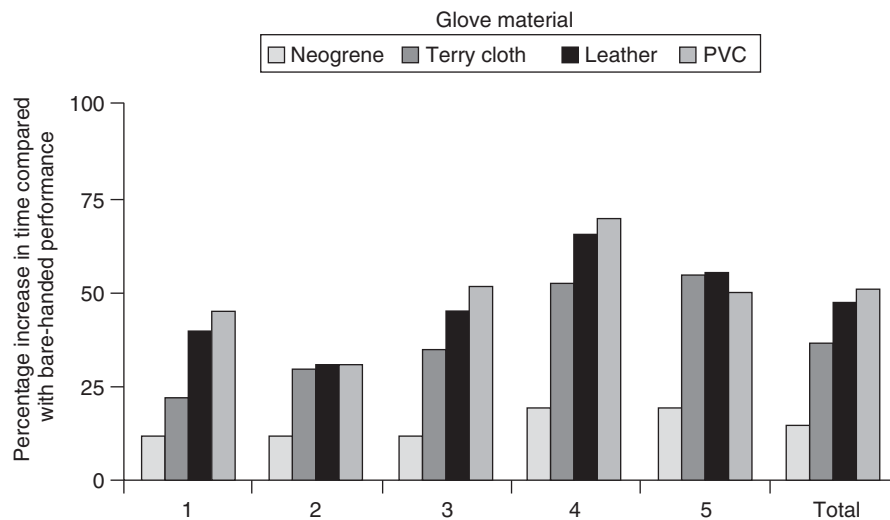


**Figure 25**  Performance (time to complete) on a maintenance-type task while wearing gloves constructed of five different materials. [From Sanders, M. S., and McCormick, E. J. (1993), *Human Factors in Engineering and Design*, McGraw-Hill, New York. With permission.]

increase in time required to perform work tasks when wearing gloves composed of different materials compared to performing the same task bare handed. The figure indicates that task performance can increase up to 70% when wearing certain types of gloves.

These effects have indicated that there are biomechanical costs associated with the use of gloves. Less strength capacity is available to the worker, more internal force is generated, less force output is available, and worker productivity is reduced when wearing gloves. These negative effects of glove use do not mean that gloves should never be worn at work. When hand protection is required, gloves should be considered as a potential solution. However, protection should only be provided to the parts of the hand that are at risk. For example, if the palm of the hand requires protection but not the fingers, fingerless gloves might provide an acceptable solution. If the fingers require protection but there is little risk to the palm of the hand, then grip tape wrapped around the fingers might be considered as a potential solution. Additionally, different styles, materials, and sizes of gloves will fit workers differently. Thus, gloves produced by different manufacturers and of different sizes should be available to the worker to minimize the negative effects mentioned above.

### 4.5.5 Design Guidelines

This discussion has indicated that there are many factors that can impact the biomechanics of the wrist and the subsequent risk of cumulative trauma disorders. Proper ergonomic design of a work task cannot be accomplished by simply providing the worker with an "ergonomically designed" tool. Ergonomics is associated with matching the workplace design to the worker's capabilities and it is not possible to design an "ergonomic tool" without considering the workplace design and task requirements simultaneously. What might be an "ergonomic" tool for one work condition may be improper for use while a worker is assuming another work posture. For example, an *in-line* tool may keep the wrist straight when inserting a bolt into a horizontal surface. However, if the bolts are to be inserted into a vertical surface, a *pistol grip* tool may be more appropriate. Using the in-line tool in this situation (inserting a bolt into a vertical surface) may cause the wrist to be deviated. This illustrates that there are no ergonomic tools, there are just ergonomic *situations*. A tool that is considered ergonomically correct in one situation may be totally incorrect in another work situation. Thus, workplace design should be performed with care and one should be alert to the trade-offs between different parts of the body that must be considered by taking into consideration the various biomechanical trade-offs.

Given these considerations, the following components of the workplace should be considered when designing a workplace to minimize cumulative trauma risk. First, keep the wrist in a neutral posture. A neutral posture is a relaxed wrist posture with a slight extension (which optimizes the forearm's length–strength relationship), not a rigid linear posture. Second, minimize tissue compression on the hand. Third, avoid tasks and actions that repeatedly impose force on the internal structures. Fourth, minimize required wrist accelerations and motions through the design of the work. Fifth, be sensitive to the impact of glove use, hand size, and left-handed workers.

## 5 BIOMECHANICAL MODELING AS A MEANS OF ASSESSING AND CONTOLLING RISK

Several models of joint and tissue loads have been created over the years with the intent of using model output for job risk analysis and control. Many modeling techniques have been employed that embrace various degrees of modeling sophistication. Based upon these assessments control measures have been developed to evaluate and control biomechanical loading of the body during work tasks. A model is nothing more than a way to organize one's logic when considering all the interacting risk factors discussed earlier. Since LBDs are often associated with spine-loading magnitude, most analysis methods have focused on risk to the back. Many biomechanical models have been developed and they all vary in the degree of complexity included in the analyses. Models range from very simple models with a great number of simplifying assumptions to very sophisticated models that monitor the precise motions of the body and recruitment of the muscles in their estimation of spinal tissue loads. While most of these models are focused upon risk assessment of the low back, several of the measures also include analyses of risk to other body parts.

### 5.1 NIOSH Lifting Guide and Revised Equation

The NIOSH has developed two assessment tools or guides to help determine the risk associated with manual materials-handling tasks. This guide was intended to be a simple representation of risk to the low back that could be employed by most people with little need for measurement or analysis equipment. The lifting guide was originally developed in 1981 (NIOSH, 1981) and applied to lifting situations where the lifts were performed in the sagittal plane and to motions that are slow and smooth. Two benchmarks or limits were defined by this guide. The first limit is called the *action limit* (AL) and represents a magnitude of weight in a given lifting situation which would impose a spine load corresponding to the beginning of LBD risk along a risk continuum. The AL was associated with the point at which people under 40 years of age just begin to experience a risk of vertebral end-plate microfracture (3400 N of compressive load) (See Figure 10). The guide estimates the force imposed upon the spine of a worker as a result of lifting a weight and compares the spine load to the AL. If the weight of the object results in a spine load that is below the AL, the job is considered safe. If the weight lifted by the worker is larger than the AL, there is some level of risk associated with the task. The general form of the AL formula is defined according to the equation

$$AL = k \, (HF)(VF)(DF)(FF) \qquad (1)$$

where

AL = action limit, kg or lb

$k$ = load constant (40 kg, or 90 lb), which is the greatest weight a subject can lift if all lifting conditions are optimal

HF = horizontal factor defined as horizontal distance from point bisecting ankles to center of gravity of load at lift origin, defined algebraically as 15/$H$ (metric units) or 6/$H$ (U.S. units)

VF = vertical factor or height of load at lift origin, defined algebraically as (0.004) $|V - 75|$ (metric) or $1 - (.01)|V - 30|$(U.S.).

DF = distance factor or vertical travel distance of load, defined algebraically as $0.7 + 7.5/D$ (metric) or $0.7 + 3/D$ (U.S.).

FF = frequency factor or lifting rate, defined algebraically as $1 - F/F_{max}$

$F$ = average frequency of lift, $F_{max}$ is shown in Table 3

This equation assumes that if the lifting conditions are ideal a worker could safely hold (and implies lift) the load constant $k$ (40 kg, or 90 lb). However, if the lifting conditions are not ideal, the allowable weight is discounted according to the four factors HF, VF, DF, and FF. These four discounting factors are shown in monogram form in Figures 26–29 and relate to many of the biomechanical principles discussed earlier. According to the relationships indicated in these figures, the HF, which is associated with the external moment, has the most dramatic effect on acceptable lifting conditions. Both VF and DF are associated with the back muscle's length–strength relationship. Finally, FF attempts to account for the cumulative effects of repetitive lifting.

The second benchmark associated with the 1981 lifting guide is the *maximum permissible limit* (MPL). The MPL represents the point at which significant risk, defined in part as a significant risk of vertebral end-plate microfracture (Figure 10), occurs. The MPL is associated with a compressive load on the spine of 6400 N, which corresponds to the point at which 50% of the people would be expected to suffer a vertebral end-plate microfracture. The MPL is a function of the AL and is defined as

$$MPL = 3(AL) \qquad (2)$$

The weight that the worker is expected to lift in a work situation is compared to the AL and MPL. If the magnitude of weight falls below the AL, the work is

**Table 3** $F_{max}$ **Table**

| | Average Vertical Location (cm) (in.) | |
|---|---|---|
| Period | Standing $V > 75$ (3) | Stooped $V \le 75$ (3) |
| 1 h | 18 | 15 |
| 8 h | 15 | 12 |

Reprinted from NIOSH, 1981.
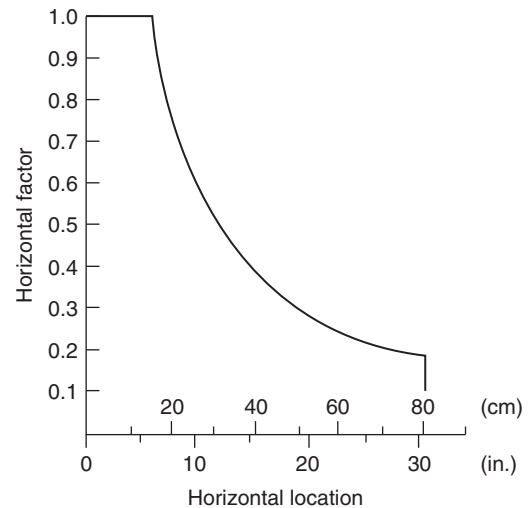


**Figure 26** Horizontal factor (HF) varies between the body interference limit and the limit of functional reach. [Adapted from National Institute for Occupational Safety and Health (NIOSH) (1981), ''Work Practices Guide for Manual Lifting,'' Publication No. 81-122, Department of Health and Human Services (DHHS), NIOSH, Washington, DC.]
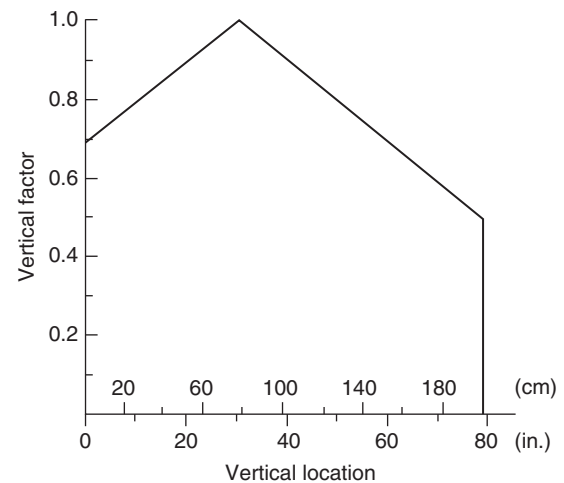


**Figure 27** Vertical factor (VF) varies both ways from knuckle height. [Adapted from National Institute for Occupational Safety and Health (NIOSH) (1981), ''Work Practices Guide for Manual Lifting,'' Publication No. 81-122, Department of Health and Human Services (DHHS), NIOSH, Cincinnati, OH.]

considered safe and no work adjustments are necessary. If the magnitude of the weight falls above the MPL, then the work is considered to represent a significant risk and engineering changes involving the adjustment of HF, VF, and/or DF are required to reduce the AL and MPL.
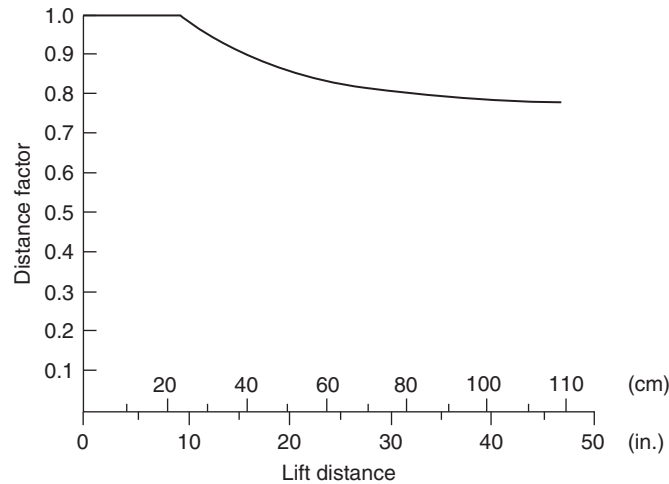
**Figure 28**  Distance factor (DF) varies between a minimum vertical distance moved of 25 cm (10 in.) to a maximum distance of 200 cm (80 in.). [Adapted from National Institute for Occupational Safety and Health (NIOSH) (1981), ''Work Practices Guide for Manual Lifting,'' Publication No. 81-122, Department of Health and Human Services (DHHS), NIOSH, Cincinnati, OH.]
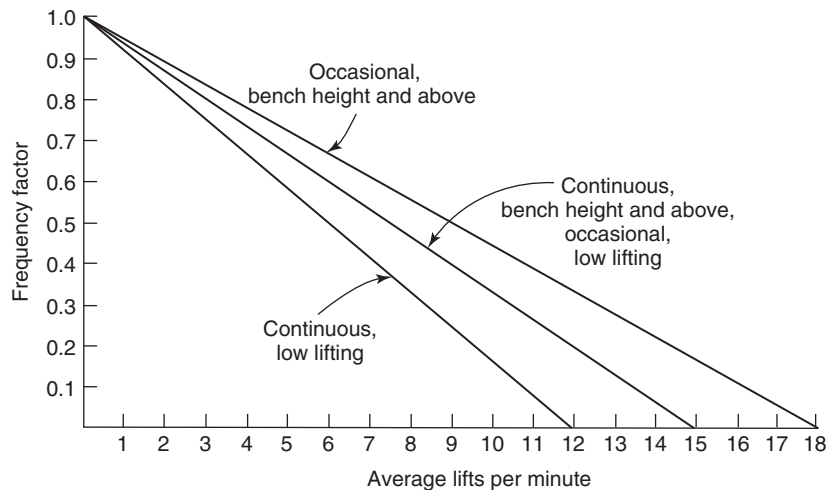


**Figure 29**  Frequency factor (FF) varies with lifts/minute and the $F_{max}$ curve. The $F_{max}$ depends upon lifting posture and lifting time. [Adapted from National Institute for Occupational Safety and Health (NIOSH) (1981), ''Work Practices Guide for Manual Lifting,'' Publication No. 81-122, Department of Health and Human Services (DHHS), NIOSH, Cincinnati, OH.]

If the weight falls between the AL and MPL, then either engineering changes or administrative changes, defined as selecting workers who are less likely to be injured or rotating workers, would be appropriate.

The AL and MPL were also indexed to relative to nonbiomechanical benchmarks. The NIOSH (1981) states that these limits also correspond to strength, energy expenditure, and psychophysical acceptance points.

The 1993 NIOSH revised lifting equation was introduce in order to address those lifting jobs that violate the sagittally symmetric lifting assumption of the original 1981 lifting guide (Waters et al., 1993). The concepts of

AL and MPL were replaced with a concept of a *lifting index* (LI) defined as

$$LI = \frac{L}{RWL} \tag{3}$$

where

$L$ = load weight or weight of object to be lifted
RWL = recommended weight limit for particular lifting situation
LI = lifting index used to estimate relative magnitude of physical stress for a particular job

If the LI is greater than 1.0, an increased risk of suffering a lifting-related LBD exists. The RWL is similar in concept to the NIOSH1(981) AL equation (equation 1) in that it contains factors that discount the allowable load according to the horizontal distance, vertical location of the load, vertical travel distance, and frequency of lift. However, the form of these discounting factors was adjusted. In addition, two discounting factors have been included. These additional factors include a lift asymmetry factor which accounts for asymmetric lifting conditions and a coupling factor that accounts for whether or not the load lifted has handles. The RWL is represented in equations (4) (metric units) and (5) (U.S. units):

$$RWL(kg) = 23(25/H)[1 - (0.003|V - 75|)][0.82$$
$$+ (4.5/D)](FM)[1 - (0.0032A)](CM)$$
$$(4)$$

$$RWL(lb) = 51(10/H)[1 - (0.0075|V - 30|)][0.82$$
$$+ (1.8/D)](FM)[1 - (0.0032A)](CM)$$
$$(5)$$

where

$H$ = horizontal location forward of midpoint between ankles at origin of lift; if significant control is required at destination, then $H$ should be measured at both origin and destination of lift
$V$ = vertical location at origin of lift
$D$ = vertical travel distance between origin and destination of lift

$FM$ = frequency multiplier shown in Table 4
$A$ = angle between midpoint of ankles and midpoint between hands at origin of lift
$CM$ = coupling multiplier ranked as good, fair, or poor and described in Table 5

In this revised equation the load constant has been significantly reduced relative to the 1981 equation. The discounting adjustments for load moment, muscle length–strength relationships, and cumulative loading are still integral parts of this equation. However, these adjustments relationships have been changed (compared to the 1981 guide) to reflect the most conservative value of the biomechanical, physiological, psychophysical, or strength data upon which they are based. Effectiveness studies report that the 1993 revised equation yields a more conservative (protective) prediction of work-related LBD risk (Marras et al., 1999).

**Table 5 Coupling Multiplier**

| Coupling Type | Coupling Multiplier | |
|---|---|---|
| | $V < 30$ in. (75 cm) | $V \geq 30$ in. (75 cm) |
| Good | 1.00 | 1.00 |
| Fair | 0.95 | 1.00 |
| Poor | 0.90 | 0.90 |

Reprinted from *Application Manual for Revised NIOSH Equation*, NIOSH, Cincinnati, OH, 1994.

**Table 4 Frequency Multiplier Table (FM)**

| Frequency Lifts, min ($F$)[b] | Work Duration[a] | | | | | |
|---|---|---|---|---|---|---|
| | $\leq 1$ h | | $> 1$ but $\leq 2$ h | | $> 2$ but $\leq 8$ h | |
| | $V < 30$ | $V \geq 30$ | $V < 30$ | $V \geq 30$ | $V < 30$ | $V \geq 30$ |
| $\geq 0.2$ | 1.00 | 1.00 | .95 | .95 | .85 | .85 |
| 0.5 | .97 | .97 | .92 | .92 | .81 | .81 |
| 1 | .94 | .94 | .88 | .88 | .75 | .75 |
| 2 | .91 | .91 | .84 | .84 | .65 | .65 |
| 3 | .88 | .88 | .79 | .79 | .55 | .55 |
| 4 | .84 | .84 | .72 | .72 | .45 | .45 |
| 5 | .80 | .80 | .60 | .60 | .35 | .35 |
| 6 | .75 | .75 | .50 | .50 | .27 | .27 |
| 7 | .70 | .70 | .42 | .42 | .22 | .22 |
| 8 | .60 | .60 | .35 | .35 | .18 | .18 |
| 9 | .52 | .52 | .30 | .30 | .00 | .15 |
| 10 | .45 | .45 | .26 | .26 | .00 | .13 |
| 11 | .41 | .41 | .00 | .23 | .00 | .00 |
| 12 | .37 | .37 | .00 | .21 | .00 | .00 |
| 13 | .00 | .34 | .00 | .00 | .00 | .00 |
| 14 | .00 | .31 | .00 | .00 | .00 | .00 |
| 15 | .00 | .28 | .00 | .00 | .00 | .00 |
| >15 | .00 | .00 | .00 | .00 | .00 | .00 |

Reprinted from *Applications Manual for the Revised NIOSH Lifting Equation*, NIOSH, Cincinnati, OH, 1994.
[a]*Values* of $V$ are in inches.
[b]*For* lifting less frequently than once per 5 min, set $F = 0.2$ lifts/min.

## 5.2  Static Single Equivalent Muscle Biomechanical Models

Biomechanically based spine models have been developed to assess occupationally related manual materials-handling tasks. These models assess the task based upon both spine-loading criteria and a strength assessment of task requirements. One of the early static assessment models was developed by Don Chaffin at the University of Michigan (Chaffin, 1969). The original two-dimensional (2D) model has been expanded to a three-dimensional (3D) static model (Chaffin et al, 2006; Chaffin and Muzaffer, 1991). In this model, the moments imposed upon the various joints of the body due to the object lifted are evaluated assuming that a static posture is representative of the instantaneous loading of the body. These models compare the imposed moments about each joint with the static strength capacity derived from a working population. The static strength capacity required of the major joint articulations used in this model have been documented in a database of over 3000 workers. In this manner the proportion of the population capable of performing a particular static exertion is estimated. The joint that limits the capacity to perform the task can be identified via this method. The model assumes that a single equivalent muscle (internal force) supports the external moment about each joint. By considering the contribution of the externally applied load and the internally generated single muscle equivalent,

spine compression at the lumbar discs is predicted. The predicted compression can then be compared to the tolerance limits for the vertebral end plate (Figure 10). Two important assumptions of this model are that (1) no significant motion occurs during the exertion since it is a static model (postures must be considered as a freeze frame in time) and (2) one "equivalent muscle" counter balances the external loads imposed upon the body (thus, coactivation of the muscle is not represented). Figure 30 shows the output screen for this computer model where the lifting posture, lifting distances, strength predictions, and spine compression are shown.

## 5.3  Multiple Muscle System Models

One significant simplifying assumption form a biomechanical standpoint in most static models is that one internal force counteracts the external moment. In reality a great deal of coactivity (simultaneous recruitment of multiple muscles) occurs in the trunk muscles during an exertion, and the more complex the exertion, the greater the coactivity. Hence, the trunk is truly a multiple-muscle system with many major muscle groups supporting and loading the spine (Schultz and Andersson, 1981). This arrangement can be seen in the cross section of the trunk shown in Figure 31. Significant coactivation also occurs in many of the major muscle groups in the trunk during realistic *dynamic* lifting (Marras and
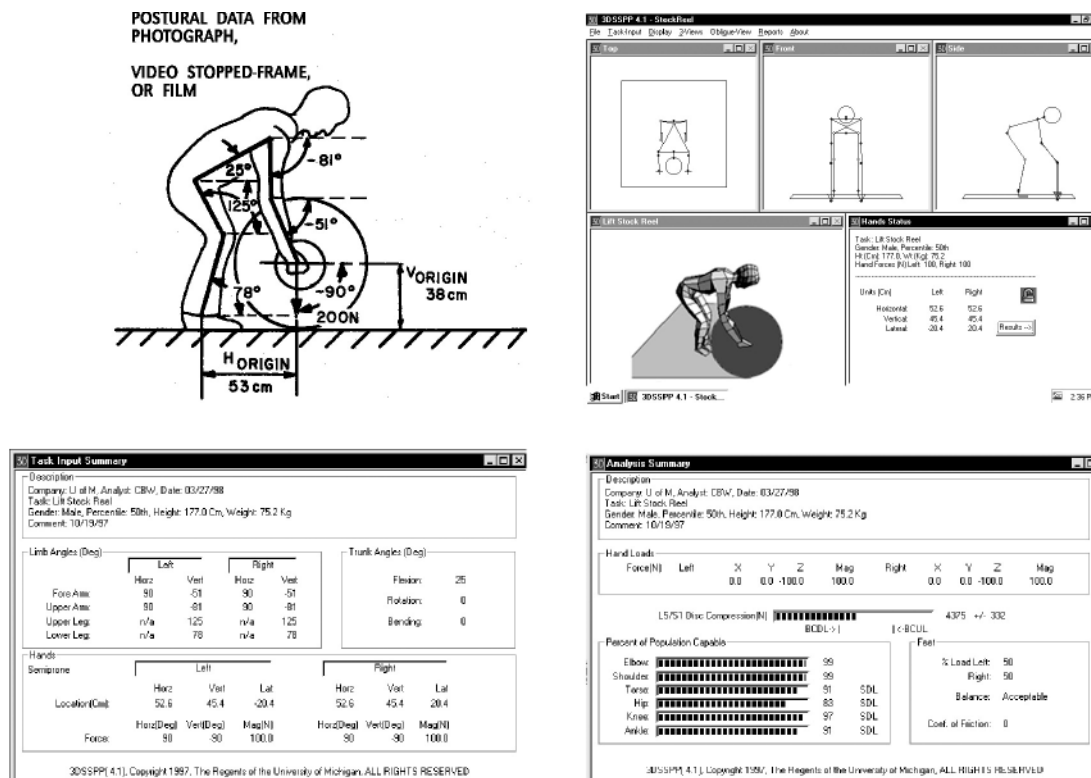


**Figure 30**  The 2D static strength prediction model. [Adapted from Chaffin, D. B., and Andersson, G. B. (1991), *Occupational Biomechanics*, Wiley, New York. With permission.]
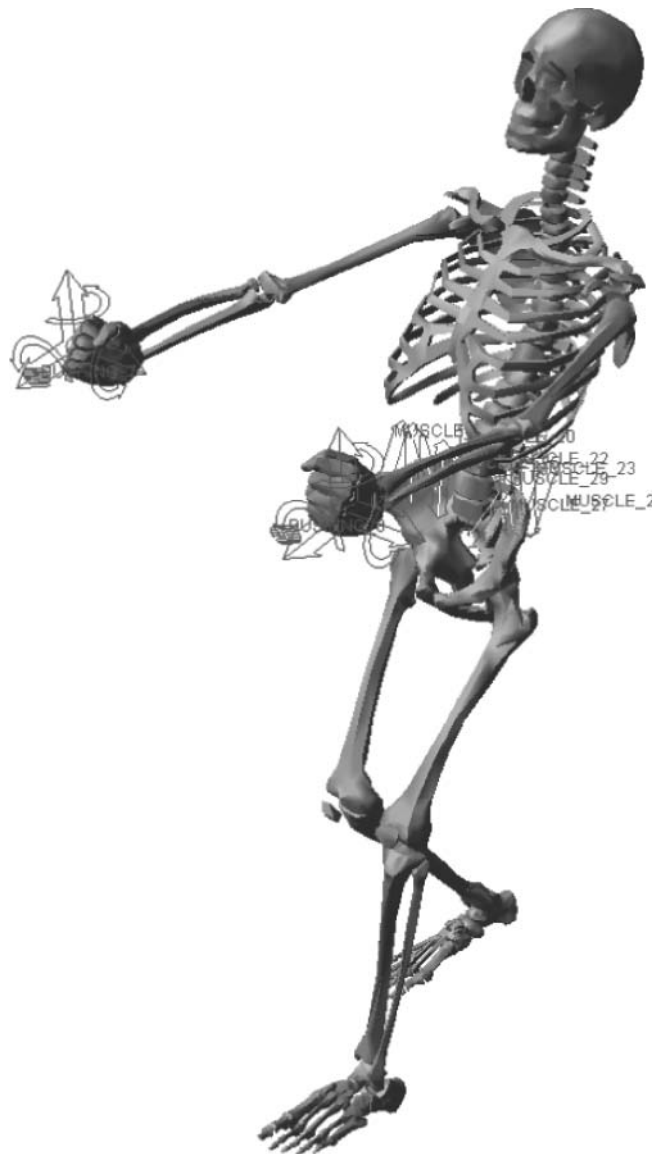
**Figure 31** Bomechanical model of a pushing-and-pulling task.

Mirka, 1993). Accounting for coactivation in these models is important because all the trunk muscles have the ability to load the spine since antagonist muscles can oppose each other during occupational tasks, thereby increasing the total load on the spine. Ignoring the coactivation of the trunk muscles during dynamic lifting can misrepresent spine loading by 45–70% (Granata and Marras, 1995a; Thelen et al., 1995). In order to more accurately estimate the loads on the lumbar spine, especially under complex, changing (dynamic) postures, multiple-muscle-system models of the trunk have been developed. However, predicting the activity of the muscles is the key to accurate low-back loading assessments.

## 5.4 Biologically Assisted Models of the Spine

One way of assessing the degree of activation of the trunk muscles during a task is to monitor the muscle force contribution by directly measuring the muscle activity within the human biological system and use this information as input to a biomechanical model. These biologically driven models typically monitor muscle activities via electromyography or EMG and use this information to directly account for muscle coactivity. EMG-assisted models take into account the individual recruitment patterns of the muscles during a specific lift for a specific individual. By directly monitoring muscle activity the EMG-assisted model is capable of

determining individual muscle force and the subsequent spine loading. These models have been developed and tested under bending and twisting dynamic motion conditions and have been validated (McGill and Norman, 1985, 1986; Marras and Reilly 1988; Reilly and Marras, 1989; Marras and Sommerich, 1991a,b; Granata and Marras, 1993,1995b). These models are the only biomechanical models that can predict the *multidimensional loads* on the lumbar spine under many three-dimensional complex dynamic lifting conditions.

Traditionally, models used for ergonomic purposes were only able to predict loads imposed on the lumbosacral junction (L5/S1). However, recently EMG-assisted models have been expanded so that they are able to predict loads on the entire lumbar spine (Knapik and Marras, 2009). This has become a particularly important development in order to assess risk during pushing and pulling since the load on the mid to upper lumbar vertebrae are much greater than those at L5/S1. These models have enabled us, for the first time, to realistically determine risk associated with pushing-and-pulling tasks in occupational environments (Marras et al., 2009a,b). While these models have become very sophisticated and accurate, their disadvantage is that they require significant instrumentation of the worker in order to generate accurate predictions of spinal loading. Figure 31 shows a graphical representation of a model used to assess a pushing-and-pulling task.

### 5.5   Finite-Element Spine Models

Finite-element models (FEMs) of the lumbar spine have been used for some time, particularly to assess the loading of the spine for clinical assessment purposes (Arjmand and Shirazi-Adl, 2005; Bowden et al., 2008; Goel and Gilbertson, 1995; Goel and Pope, 1995; Shirazi-Adl et al., 1986; Suwito et al., 1992; Zander et al., 2004). FEMs are valuable techniques to consider how loads imposed upon a structure cause deformation and damage to the structure. The idea behind FEMs is that the structure is represented by many small elements that are representative of the underlying structural strength of the material. When loads are imposed on the structure, the FEM will predict how the elements rearrange themselves and provide information about how the structure will fail.

FEMs appear to be valuable tools to assess load tolerance of a structure if they are properly modeled. These models have been used often to assess the impact of spinal instrumentation (Bowden et al., 2008; Goel and Pope, 1995; Bono et al., 2007; Dooris et al., 2001). However, these models do little to help assess how loads are imposed upon the spine during task performance due to muscle recruitment since they do not include the coactive influence of spine-loading muscles in their analyses.

### 5.6   Personalized Hybrid Spine Models

Recently hybrid EMG-driven/FEM models have been developed to take advantage of the strengths of both EMG-assisted models and FEMs. In these hybrid models EMG-assisted techniques are used to assess the forces imposed upon the spine during a task, whereas FEM techniques are used to determine the impact of these

forces on spine tolerance. Some of the latest improvements in these models have enabled the ability to import anatomic images of a specific individual into the model. This has facilitated the development of a "personalized" hybrid model of specific persons. With this enhancement one can assess spinal loads and consider the effects of degeneration upon spine loading. These models have been used to assess clinical interventions such as surgery (Marras et al., 2008) as well as occupational tasks (Marras, 2008). These personalized models represent the future of biomechanical assessments and should become more readily available for applications as computer processing power continues to improve. An example of one of these models is shown in Figure 32.

### 5.7   Stability-Driven Models of the Spine

Stability of a system refers to the ability of the system to return to a state of equilibrium after a perturbation to the system. This concept is key to predicting system "balance" and has been traditionally used to predict force experienced by joints such as the knee during sports. The idea is that when a simple joint (e.g., knee) is unstable the ligaments will become stretched or torn and damage will occur.

While this concept is generally accepted for the assessment of risk for simple joints, particularly during sporting activities, it has also been proposed that similar concepts may apply to the much more complex system of the low back (Panjabi, 1992a,b, 2003). However, muscle involvement in the spine is much more difficult to predict than that of a simple joint.

From a biomechanical standpoint, the concept of stability is important for two reasons. First, if instability



**Figure 32**  Example of a personalized hybrid EMG-driven/FEM of the lumbar spine.

can be predicted, then it might be able to pinpoint the tissues at risk due to occupational task performance. Second, instability might "drive" the muscular recruitment system, making it possible to predict muscle activity so that one might be able to understand when overexertion occurs. In addition, prediction of muscle recruitment patterns might eliminate the need for equipment-intensive EMG measurements during work, thereby making EMG-driven modeling easier. Several researchers have attempted to model the spine via stability-driven principles (Cholewicki et al., 2000, 2005; Cholewicki and VanVliet, 2002; Granata and England, 2006; Granata and Marras, 2000; Granata and Orishimo, 2001). Unfortunately, thus far these techniques have been successful only when applied to static conditions and have not yet been able to assess dynamic task activities. Thus, these approaches have not been able to represent realistic risk at the workplace. The principal problem with this approach is that the stability principle does not appear to predict trunk muscle recruitment accurately because of its inability to predict dynamic cocontraction of the trunk muscles. The torso's muscle recruitment system appears to be driven by an individual's own mental model that they develop over their lifetime (Marras, 2008; Erlandson and Fleming, 1974) and is unique to an individual.

## 5.8 Predicting Muscle Recruitment for Spine Model Use

It is obvious from the preceding discussion regarding biomechanical modeling that a critical requirement for accurate biomechanical modeling of spinal tissues is the ability to accurately assess the behavior of the power-producing muscles of the trunk. Since these muscles have short moment arms relative to the externally applied moments, their influence upon spine tissue loading is immense. This is why biologically (EMG-)assisted models are currently the most accurate and precise means to assess spine loading during occupational task performance. Unfortunately, EMG requires significant equipment and is sometimes impractical at the worksite and therefore could require task simulation in a laboratory environment.

In an effort to minimize the need for EMG collection several researchers have attempted to predict muscle activities during task performance. Several techniques have been attempted. Optimization techniques have been attempted for some time (Bean et al., 1988; Brown et al., 2005; Cholewicki and McGill, 1994; Hughes and Chaffin, 1995; Li et al., 2006; van Dieen and Kingma, 2005; Zhang et al., 1998). Some of these attempts have been able to predict muscle activity under steady-state static loading conditions. However, these conditions are not representative of realistic work situations and can dramatically underestimate spine tissue loading (Granata and Marras, 1995a. These optimization-based assessments have not been able to accurately predict loading under realistic dynamic task performance conditions.

Several efforts have attempted to employ neural network and fuzzy logic techniques to predict trunk muscle coactivation under occupational task performance conditions (Hou et al., 2007; Lee et al., 2000, 2003; Nussbaum and Chaffin, 1997). These efforts have employed

databases of EMG responses to various kinematic and kinetic spine-loading conditions to train neural network models of trunk muscle responses. Given a specific peak moment exposure and velocity of trunk motion, these models appear able to relatively accurately predict muscle behavior for a wide range of subjects. While these models show promise, it is unfortunate that they require a large volume of training data so that a variety of activities can be represented. It is not expected that such a large database will be available for comprehensive modeling of the muscle activities of the trunk.

## 5.9 Dynamic Motion Assessment at the Workplace

It is clear that that dynamic activity may significantly increase the risk of LBD, yet there are few assessment tools available to quickly and easily assess the biomechanical demands associated with workplace dynamics and the risk of LBD. In order to assess this biomechanical situation at the worksite, one must know the type of motion that increases biomechanical load and determine "how much motion exposure is too much motion exposure" from a biomechanical standpoint. These issues were the focus of several industrial studies performed over a six-year period in 68 industrial environments. Trunk motion and workplace conditions were assessed in workers exposed to high-risk of LBD jobs and compared to trunk motions and workplace conditions associated with low-risk jobs (Marras et al., 1993, 1995). A trunk goniometer (lumbar motion monitor, or LMM) has been used to document the trunk motion patterns of workers at the workplace and is shown in Figure 33. Based upon this study, a five-factor multiple logistic



**Figure 33**   The LMM.

regression model was developed that is capable of discriminating between task exposure that indicates the probability of high-risk group membership. These risk factors include (1) frequency of lifting, (2) load moment (load weight multiplied by the distance of the load from the spine), (3) average twisting velocity (measured by the LMM), (4) maximum sagittal flexion angle through the job cycle (measured by the LMM), and (5) maximum lateral velocity (measured by the LMM). This LMM risk assessment model is the only model capable of assessing the *risk* associated with three-dimensional trunk motion on the job. This model has a high degree of predictability (odds ratio 10.7) compared to previous attempts to assess work-related LBD risk. The advantage of such an assessment is that the evaluation provides information about risk that would take years to derive from historical accounts of incidence rates. The model has also been validated prospectively (Marras et al., 2000).

## 5.10 Threshold Limit Values

Threshold limit values (TLVs) have been recently introduced as a means for controlling biomechanical risk to the back in the workplace. TLVs have been introduced through the American Conference of Governmental Industrial Hygienists (ACGIH) and provide lifting weight limits as a function of lift origin "zones" and repetitions associated with occupational tasks. Lift origin zones are defined by the lift height off the ground and lift distance from the spine associated with the lift origin. Twelve zones are defined that relate to lifts within $+/-30°$ of asymmetry from the sagittal plane. These zones are represented in a series of figures with each figure corresponding to different lift frequency and time exposures. Within each zone weight-lifting limits are specified based upon the best information available from several sources which include (1) EMG-assisted biomechanical models, (2) the 1993 revised lifting
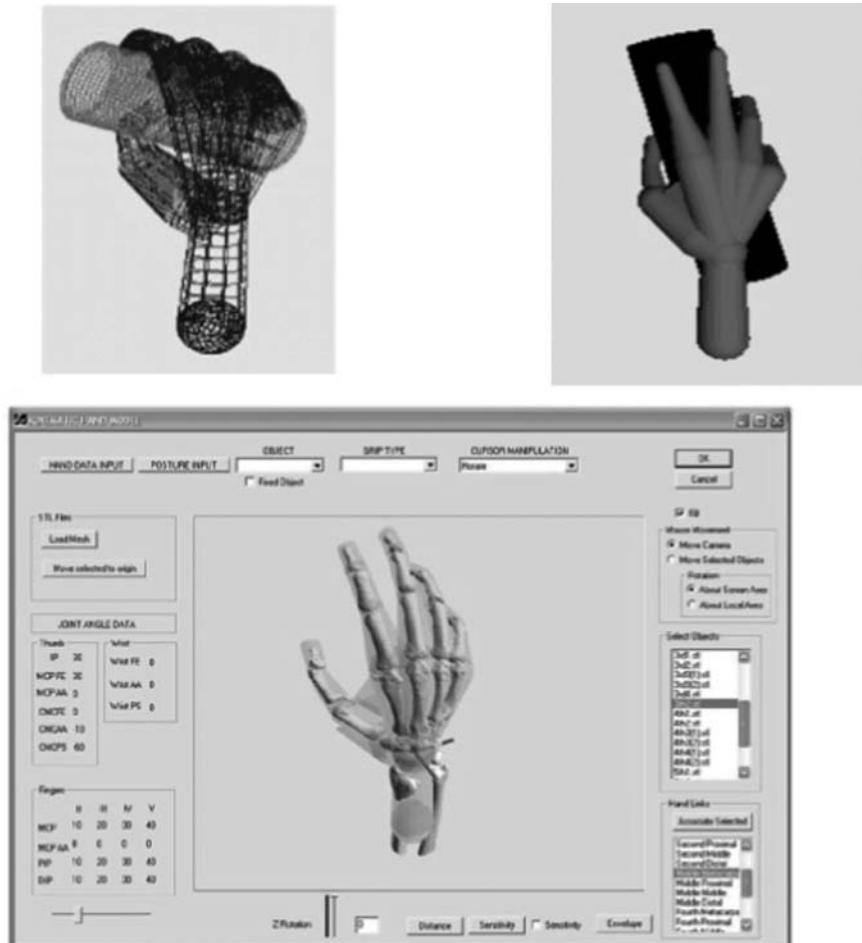


**Figure 34**  Upper extremity biomechanical model used for ergonomics assessments. (Courtesy of T. Armstrong.)

equation, and (3) the historical risk data associated with the LMM database. The weight lifted by the worker is compared to these limits. Weights exceeding the zone limit are considered hazards.

## 5.11 Upper Extremity Models

Recently, the Center for Ergonomics at the University of Michigan has developed a kinetic model of the upper extremity that is intended to be used to assess hand-intensive tasks (Armstrong et al., in press). This model consists of a link system that represents the joints of the hand and cone shapes are used to represent finger surfaces. The model estimates hand postures and finger movements. The model has been used to determine how workers grasp objects in the workplace and assesses how much space will be required for the hand and the required tendon forces and hand strength necessary to perform a task. The model has recently been used to evaluate hose insertion tasks. Figure 34 illustrates the graphical nature of this model.

## 6 SUMMARY

This chapter has shown that biomechanics provides a means to *quantitatively* consider the implications of workplace design. Biomechanical design considerations are important when a particular job is suspected of imposing large or repetitive forces on the structures of the body. It is particularly important to recognize that the internal structures of the body, such as muscles, are the primary generators of force within the joint and tendon structures. In order to evaluate the risk of injury due to a particular task, one must consider the contribution of both the external loads and internal loads upon a structure and how they relate to the tolerance of the structure. Armed with an understanding of some general biomechanical concepts (presented in this chapter) and how they apply to different parts of the body (affected by work), one can logically reason through the design considerations and trade-offs so that musculoskeletal disorders are minimized due to the design of the work.

## REFERENCES

Adams, M. A., Freeman, B. J., Morrison, H. P., Nelson, I. W., and Dolan, P. (2000), "Mechanical Initiation of Intervertebral Disc Degeneration," *Spine*, Vol. 25, p. 1625.

Anderson, C. K., Chaffin, D. B., Herrin, G. D., and Matthews, L. S. (1985), "A Biomechanical Model of the Lumbosacral Joint during Lifting Activities," *Journal of Biomechanics*, Vol. 18, p. 571.

Andersson, B. J., Ortengren, R., Nachemson, A. L., Elfstrom, G., and Broman, H. (1975), "The Sitting Posture: An Electromyographic and Discometric Study," *Orthopedic Clinics of North America*, Vol. 6, p. 105.

Andersson, G. B. (1997), in *The Adult Spine: Principles and Practice,* Vol. 1, J. W. Frymoyer, Ed., Lippincott-Raven Publishers, Philadelphia, pp. 93–141.

Arjmand, N., and Shirazi-Adl, A. (2005, December 1), "Biomechanics of Changes in Lumbar Posture in Static Lifting," *Spine*, Vol. 30, p. 2637.

Armstrong, T., et al. (in press), "Development of a Kinematic Hand Model for Study and Design of Hose Installation," *International Journal of Human Factors Modeling and Simulation*.

Bean, J. C., Chaffin, D. B., and Schultz, A. B. (1988), "Biomechanical Model Calculation of Muscle Contraction Forces: A Double Linear Programming Method," *Journal of Biomechanics*, Vol. 21, p. 59.

Bettinger, P. C., Smutz, W. P., Linscheid, R. L., Cooney III, W. P., and An, K. N. (2000, November), "Material Properties of the Trapezial and Trapeziometacarpal Ligaments," *Journal of Hand Surgery America*, Vol. 25, p. 1085.

Beumer, A., van Hemert, W. L., Swierstra, B. A., Jasper, L. E., and Belkoff, S. M. (2003, May), "A Biomechanical Evaluation of the Tibiofibular and Tibiotalar Ligaments of the Ankle," *Foot and Ankle International*, Vol. 24, p. 426.

Bono, C. M., et al., (2007, February 15), "Residual Sagittal Motion after Lumbar Fusion: A Finite Element Analysis with Implications on Radiographic Flexion-Extension Criteria," *Spine*, Vol. 32, p. 417.

Bowden, A. E., Guerin, H. L, Villarraga, M. L., Patwardhan, A. G., and Ochoa, J. A. (2008, June), "Quality of Motion Considerations in Numerical Analysis of Motion Restoring Implants of the Spine," *Clinical Biomechanics* (*Bristol, Avon*), Vol. 23, p. 536.

Brinkmann, P., Biggermann, M., and Hilweg, D. (1988), "Fatigue Fracture of Human Lumbar Vertebrae," *Clinical Biomechanics* (*Bristol, Avon*), Vol. 3, p. S1.

Brown, S. H., Howarth, S. J., and McGill, S. M. (2005, September), "Spine Stability and the Role of Many muscles," *Archives of Physical Medicine and Rehabilitation*, Vol. 86, p. 1890; author reply p. 1890.

Butler, D. L., Kay, M. D., and Stouffer, D. C. (1986), "Comparison of Material Properties in Fascicle-Bone Units from Human Patellar Tendon and Knee Ligaments," *Journal of Biomechanics*, Vol. 19, p. 425.

Callaghan, J. P., and McGill, S. M. (2001), "Intervertebral Disc Herniation: Studies on a Porcine Model Exposed to Highly Repetitive Flexion/Extension Motion with Compressive Force," *Clinical Biomechanics* (*Bristol, Avon*), Vol. 16, p. 28.

Cavanaugh, J. M. (1995), "Neural Mechanisms of Lumbar Pain," *Spine*, Vol. 20, p. 1804.

Cavanaugh, J. M., et al. (1997), "Mechanisms of Low Back Pain: A Neurophysiologic and Neuroanatomic Study," *Clinical Orthopaedics*, Vol. 335 p. 166.

Chaffin, D. B. (1969), "A Computerized Biomechanical Model: Development of and Use in Studying Gross Body Actions," *Jounal of Biomechanics*, Vol. 2, p. 429.

Chaffin, D. B. (1973, April), "Localized muscle Fatigue—Definiton and Measurement," *Journal of Occupational Medicine*, Vol. 15, p. 346.

Chaffin, D. B., and Muzaffer, E. (1991), "Three-Dimensional Biomechanical Static Strength Prediction Model Sensitivity to Postural and Anthropometric Innaccuracies," *IIE Transactions* Vol. 23, p. 215.

Chaffin, D. B., Andersson, G. B. J., and Martin, B. J. (2006), *Occupational Biomechanics*, 4th ed., Wiley, New York.

Cholewicki, J., and McGill, S. M. (1994), "EMG Assisted Optimization: A Hybrid Approach for Estimating Muscle Forces in an Indeterminate Biomechanical Model," *Journal of Biomechanics*, Vol. 27, p. 1287.

Cholewicki, J., and VanVliet, I. J. (2002), "Relative Contribution of Trunk Muscles to the Stability of the Lumbar Spine during Isometric Exertions," *Clinical Biomechanics* (*Bristol, Avon*), Vol. 17, p. 99.

Cholewicki, J., Simons, A. P., and Radebold, A. (2000), "Effects of External Trunk Loads on Lumbar Spine Stability," *Journal of Biomechanics*, Vol. 33, p. 1377.

Cholewicki, J., et al. (2005, December 1), "Delayed Trunk Muscle Reflex Responses Increase the Risk of Low Back Injuries," *Spine*, Vol. 30, p. 2614.

Clavert, P., Kempf, J. F., and Kahn, J. L. (2009, January-February), "Biomechanics of Open Bankart and Coracoid Abutment Procedures in a Human Cadaveric Shoulder Model," *Journal of Shoulder and Elbow Surgery*, Vol. 18, p. 69.

Dooris, A. P., Goel, V. K., Grosland, N. M., Gilbertson, L. G., and Wilder, D. G. (2001, March 15), "Load-Sharing between Anterior and Posterior Elements in a Lumbar Motion Segment Implanted with an Artificial Disc," *Spine*, Vol. 26, p. E122.

Erlandson, R. F., and Fleming, D. G. (1974), "Uncertainty Sets Associated with Saccadic Eye Movements—Basis of Satisfaction Control," *Vision Research*, Vol. 14, p. 481.

Fremerey, R., Bastian, L., and Siebert, W. E. (2000), "The Coracoacromial Ligament: Anatomical and Biomechanical Properties with Respect to Age and Rotator Cuff Disease," *Knee Surgery, Sports Traumatology, Arthroscopy*, Vol. 8, p. 309.

Gallagher, S., Marras, W. S., Litsky, A. S., and Burr, D. (2005, October 15), "Torso Flexion Loads and the Fatigue Failure of Human Lumbosacral Motion Segments," *Spine*, Vol. 30, p. 2265.

Gedalia, U., et al. (1999), "Biomechanics of Increased Exposure to Lumbar Injury Caused by Cyclic Loading. Part 2. Recovery of Reflexive Muscular Stability with Rest," *Spine*, Vol. 24, p. 2461.

Goel, V. K., and Gilbertson, L. G. (1995, August 1), "Applications of the Finite Element Method to Thoracolumbar Spinal Research—Past, Present, and Future," *Spine*, Vol. 20, p. 1719.

Goel, V. K., Pope, and M. H. (1995, December 15), "Biomechanics of Fusion and Stabilization," *Spine*, Vol. 20, p. 85S.

Granata, K. P., and England, S. A. (2006, May 1), "Stability of Dynamic Trunk Movement," *Spine*, Vol. 31, p. E271.

Granata, K. P., and Marras, W. S. (1993), "An EMG-Assisted Model of Loads on the Lumbar Spine during Asymmetric Trunk Extensions," *Journal of Biomechanics*, Vol. 26, p. 1429.

Granata, K. P., and Marras, W. S. (1995a), "The Influence of Trunk Muscle Coactivity on Dynamic Spinal Loads," *Spine*, Vol. 20, p. 913.

Granata, K. P., and Marras, W. S. (1995b), "An EMG-Assisted Model of Trunk Loading during Free-Dynamic Lifting," *Journal of Biomechanics*, Vol. 28, p.1 309.

Granata, K. P., and Marras, W. S. (2000), "Cost-Benefit of Muscle Cocontraction in Protecting against Spinal Instability," *Spine*, Vol. 25, p. 1398.

Granata, K. P., and Orishimo, K. F. (2001), "Response of Trunk Muscle Coactivation to Changes in Spinal Stability," *Journal of Biomechanics*, Vol. 34, p. 1117.

Guo, H. R. (1993), in *American Occupational Health Conference*, NIOSH Report.

Guo, H. R., Tanaka, S., Halperin, W. E., and Cameron, L. L. (1999, July), "Back Pain Prevalence in US Industry and Estimates of Lost Workdays," *American Journal of Public Health*, Vol. 89, p. 1029.

Hewitt, J., Guilak, F., Glisson, R., and Vail, T. P. (2001, May), "Regional Material Properties of the Human Hip Joint Capsule Ligaments," *Journal of Orthopaedic Research*, Vol. 19, p. 359.

Hewitt, J. D., Glisson, R. R., Guilak, F., and Vail, T. P. (2002, January), "The Mechanical Properties of the Human Hip Capsule Ligaments," *Journal of Arthroplasty*, Vol. 17, p. 82.

Hoefnagels, E. M., Waites, M. D., Wing, I. D., Belkoff, S. M., and Swierstra, B. A. (2007, May), "Biomechanical Comparison of the Interosseous Tibiofibular Ligament and the Anterior Tibiofibular Ligament," *Foot and Ankle International*, Vol. 28, p. 602.

Hollbrook, T. L., Grazier, K., Kelsey, J. L., and Stauffer, R. N. (1984), *The Frequency of Occurrence, Impact and Cost of Selected Musculoskeletal Conditions in the United States*, American Academy of Orthopaedic Surgeons, Chicago, IL, pp. 24–45.

Hou, Y., Zurada, J. M., Karwowski, W., Marras, W. S., and Davis, K. (2007, February), "Estimation of the Dynamic Spinal Forces Using a Recurrent Fuzzy Neural Network," *IEEE Transactions on Systems, Man, and Cybernetics, Part B, Cybernetics*, Vol. 37, p. 100.

Hoy, M. G., Zajac, F. E., and Gordon, M. E. (1990), "A Musculoskeletal Model of the Human Lower Extremity: The Effect of Muscle, Tendon, and Moment Arm on the Moment-Angle Relationship of Musculotendon Actuators at the Hip, Knee, and Ankle," *Journal of Biomechanics*, Vol. 23, p. 157.

Hughes, R. E., and Chaffin, D. B. (1995), "The Effect of Strict Muscle Stress Limits on Abdominal Muscle Force Predictions for Combined Torsion and Extension Loadings," *Journal of Biomechanics*, Vol. 28, p. 527.

Jager, M., Luttmann, A., and Laurig, W. (1991), "Lumbar Load during One-Hand Bricklaying," *International Journal of Industrial Ergonomics*, Vol. 8, p. 261.

Johnston, J. D., Small, C. F., Bouxsein, M. L., and Pichora, D. R. (2004, July), "Mechanical Properties of the Scapholunate Ligament Correlate with Bone Mineral Density Measurements of the Hand," *Journal of Orthopaedic Research*, Vol. 22, p. 867.

Jung, H. J., Fisher, M. B., and Woo, S. L. (2009), "Role of Biomechanics in the Understanding of Normal, Injured, and Healing Ligaments and Tendons," *Sports Medicine, Arthroscopy, Rehabilitatiion, Therapy and Technology*, Vol. 1, p. 9.

Khalsa, P. S. (2004, February), "Biomechanics of Musculoskeletal Pain: Dynamics of the Neuromatrix," *Journal of Electromyography and Kinesiology*, Vol. 14, p. 109.

Kim, J., Stuart-Buttle, C., and Marras, W. S. (1994), "The Effects of Mats on Back and Leg Fatigue," *Applied Ergonomics*, Vol. 25, p. 29.

Knapik, G. G., and Marras, W. S. (Jan, 2009, January), "Spine Loading at Different Lumbar Levels during Pushing and Pulling," *Ergonomics*, Vol. 52, p. 60.

Kovacs, K., Splittstoesser, R., Maronitis, A., and Marras, W. S. (2002, May-June), "Grip Force and Muscle Activity Differences Due to Glove Type," *AIHA Journal* (*Fairfax, Va*), Vol. 63, p. 269.

Kroemer, K. H. E. (1987), "Biomechanics of the Human Body," in *Handbook of Human Factors,* G. Salvendy, Ed., Wiley, New York.

Kura, H., Luo, Z. P., Kitaoka, H. B., Smutz, W. P., and An, K. N. (2001, February), "Mechanical Behavior of the Lisfranc and Dorsal Cuneometatarsal Ligaments: In Vitro Biomechanical Study," *Journal of Orthopaedic Trauma*, Vol. 15, p. 107.

Lee, W., Karwowski, W., and Marras, W. S. (2000), in *XIVth Triennial Congress of the International Ergonomics Association*, Vol. 5, Human Factors and Ergonomics Society, San Diego, CA, pp. 276–279.

Lee, W., Karwowski, W., Marras, W. S., and Rodrick, D. (2003, January 15), "A Neuro-Fuzzy Model for Estimating Electromyographical Activity of Trunk Muscles Due to Manual Lifting," *Ergonomics*, Vol. 46, p. 285.

Li, G., Pierce, J. E., and Herndon, J. H. (2006), "A Global Optimization Method for Prediction of Muscle Forces of Human Musculoskeletal System," *Journal of Biomechanics*, Vol. 39, p. 522.

Luo, X., Pietrobon, R., S, X. S., Liu, G. G., and Hey, L. (2004, January 1), "Estimates and Patterns of Direct Health Care Expenditures among Individuals with Back Pain in the United States," *Spine*, Vol. 29, p. 79.

Marras, W. S. (2008), *The Working Back: A Systems View*, Wiley-Interscience, Hoboken, NJ.

Marras, W. S., and Granata, K. P. (1995), "A Biomechanical Assessment and Model of Axial Twisting in the Thoracolumbar Spine," *Spine*, Vol. 20, p. 1440.

Marras, W. S., and Granata, K. P. (1997a), "Spine Loading during Trunk Lateral Bending Motions," *Journal of Biomechanics*, Vol. 30, p. 697.

Marras, W. S., and Granata, K. P. (1997b), "The Development of an EMG-Assisted Model To Assess Spine Loading during Whole-Body Free-Dynamic Lifting," *Journal of Electromyography and Kinesiology*, Vol. 7, p. 259.

Marras, W. S., and Mirka, G. A. (1993), "Electromyographic Studies of the Lumbar Trunk Musculature during the Generation of Low-Level Trunk Acceleration," *Journal of Orthopaedic Research*, Vol. 11, p. 811.

Marras, W. S., and Reilly, C. H. (1988), "Networks of Internal Trunk-Loading Activities under Controlled Trunk-Motion Conditions," *Spine*, Vol. 13, p. 661.

Marras, W. S., and Schoenmarklin, R. W. (1993, April), "Wrist Motions in Industry," *Ergonomics*, Vol. 36, p. 341.

Marras, W. S., and Sommerich, C. M. (1991a), "A Three-Dimensional Motion Model of Loads on the Lumbar Spine: I. Model Structure," *Human Factors*, Vol. 33, p. 123.

Marras, W. S., and Sommerich, C. M. (1991b), "A Three-Dimensional Motion Model of Loads on the Lumbar Spine: II. Model Validation," *Human Factors*, Vol. 33, p. 139.

Marras, W. S., et al. (1993), "The Role of Dynamic Three-Dimensional Trunk Motion in Occupationally- Related Low Back Disorders. The Effects of Workplace Factors, Trunk Position, and Trunk Motion Characteristics on Risk of Injury," *Spine*, Vol. 18, p. 617.

Marras, W. S., et al. (1995), "Biomechanical Risk Factors for Occupationally Related Low Back Disorders," *Ergonomics*, Vol. 38, p. 377.

Marras, W. S., Fine, L. J., Ferguson, S. A., and Waters, T. R. (1999), "The Effectiveness of Commonly Used Lifting Assessment Methods to Identify Industrial Jobs Associated with Elevated Risk of Low-Back Disorders," *Ergonomics*, Vol. 42, p. 229.

Marras, W. S., Allread, W. G., Burr, D. L., and Fathallah, F. A. (2000), "Prospective Validation of a Low-Back Disorder Risk Model and Assessment of Ergonomic Interventions Associated with Manual Materials Handling Tasks [In Process Citation]," *Ergonomics*, Vol. 43, p. 1866.

Marras, W. S., Davis, K. G., and Splittstoesser, R. E. (2001), "Spine Loading during Whole Body Free Dynamic Lifting," Final project report Grant #:R01 OH03289, The Ohio State University, Columbus, OH.

Marras, W. S., Knapik, G. G., and Gabriel, J. (2008), "The Development of a Personalized Hybrid EMG-Assisted/Finite Element Biomechanical Model to Access Surgical Options," in *Motion Preservation Surgery of the Spine: Advanced Techniques and Controversies*, J. J. Yue, R. Bertagnoli, P. C. McAfee, and H. S. An, Eds., Saunders, Philadelphia, pp. 687–694.

Marras, W. S., Knapik, G. G., and Ferguson, S. (2009a, February), "Loading Along the Lumbar Spine as Influenced by Speed, Control, Load Magnitude, and Handle Height during Pushing," *Clinical Biomechanics (Bristol, Avon)*, Vol. 24, p. 155.

Marras, W. S., Knapik, G. G., and Ferguson, S. (2009b, March), "Lumbar Spine Forces during Maneuvering of Ceiling-Based and Floor-Based Patient Transfer Devices," *Ergonomics* Vol. 52, p. 384.

Marras, W. S., Lavender, S. A., Ferguson, S. A., Splittstoesser, R. E., and Yang, G. (2010, April 15), "Quantitative Dynamic Measures of Physical Exposure Predict Low Back Functional Impairment," *Spine (Phila Pa 1976)*, Vol. 35, p. 914.

McGill, S. (2002), *Low Back Disorders: Evidence-Based Prevention and Rehabilitation*, Human Kinetics, Champaign, IL.

McGill, S. M. (1997), "The Biomechanics of Low Back Injury: Implications on Current Practice in Industry and the Clinic," *Journal of Biomechanics*, Vol. 30, p. 465.

McGill, S. M., and Norman, R. W. (1985), "Dynamically and Statically Determined Low Back Moments during Lifting," *Journal of Biomechanics*, Vol. 18, p. 877.

McGill, S. M., and Norman, R. W. (1986), "Partitioning of the L4-L5 Dynamic Moment into Disc, Ligamentous, and Muscular Components during Lifting [see comments]," *Spine*, Vol. 11, p. 666.

McGill, S. M., Hughson, R. L., and Parks, K. (2000), "Changes in Lumbar Lordosis Modify the Role of the Extensor Muscles [In Process Citation]," *Clinical Biomechanics (Bristol, Avon)*, Vol. 15, p. 777.

Moore, S. M., Ellis, B., Weiss, J. A., McMahon, P. J., and Debski, R. E. (2010, January), "The Glenohumeral Capsule Should Be Evaluated as a Sheet of Fibrous Tissue: A Validated Finite Element Model," *Annals of Biomedical Engineering*, Vol. 38, p. 66.

Nachemson, A., (1975, August), "Towards a Better Understanding of Low-Back Pain: A Review of the Mechanics of the Lumbar Disc," *Rheumatology and Rehabilitation*, Vol. 14, p. 129.

Nachemson, A. L., and Evans, J. H. (1968, August), "Some Mechanical Properties of the Third Human Lumbar Interlaminar Ligament (Ligamentum Flavum)," *Journal of Biomechanics*, Vol. 1, p. 211.

National Institute for Occupational Safety and Health (NIOSH) (1981), "Work Practices Guide for Manual Lifting," 81–122, Department of Health and Human Services (DHHS), NIOSH, Cincinnati, OH.

Nordin, M., and Frankel, V. (1989), *Basic Biomechanics of the Musculoskeletal System*, 2nd ed., Lea and Febiger, Philadelphia, PA.

National Research Council (NRC) (1999), "Work-Related Musculoskeletal Disorders: Report, Workshop Summary, and Workshop Papers," National Academy Press, Washington, DC.

National Research Council (NRC) (2001), "Musculoskeletal Disorders and the Workplace: Low Back and Upper Extremity," National Academy Press, Washington, DC.

Nussbaum, M. A., and Chaffin, D. B. (1997, March), "Pattern Classification Reveals Intersubject Group Differences in Lumbar Muscle Recruitment during Static Loading," *Clinical Biomechanics* (*Bristol, Avon*), Vol. 12, p. 97.

Ozkaya, N., and Nordin, M. (1991), *Fundamentals of Biomechanics, Equilibruim, Moation and Deformation*, Van Nostrand Reinhold, New York.

Panjabi, M. M. (1992a), "The Stabilizing System of the Spine. Part II. Neutral Zone and Instability Hypothesis," *Journal of Spinal Disorders*, Vol. 5, p. 390.

Panjabi, M. M. (1992b, December), "The Stabilizing System of the Spine. Part I. Function, Dysfunction, Adaptation, and Enhancement," *Journal of Spinal Disorders*, Vol. 5, p. 383.

Panjabi, M. M. (2003, August), "Clinical Spinal Instability and Low Back Pain," *Journal of Electromyography and Kinesiology*, Vol. 13, p. 371.

Park, K., and Chaffin, D. (1974), "A Biomechanical Evaluation of Two Methods of Manual Load Lifting," *AIIE Transactions*, Vol. 6, p. 105.

Pfaeffle, H. J., et al. (1996, September), "Tensile Properties of the Interosseous Membrane of the Human Forearm," *Journal of Orthopaedic Research*, Vol. 14, p. 842.

Pintar, F. A., Yoganandan, N., Myers, T., Elhagediab, A., and Sances, Jr., A. (1992, November), "Biomechanical Properties of Human Lumbar Spine Ligaments," *Journal of Biomechanics*, Vol. 25, p. 1351.

Pope, M. H. (1993), in *International Society of Biomechanics XIVth Congress*, Paris.

Praemer, A., Furner, S., and Rice. D. P. (1992), "Musculoskeletal Conditions in the United States," American Academy of Orthopaedic Surgeons, Park Ridge, IL, pp. 23–33.

Quapp, K. M., and Weiss, J. A. (1998, December), "Material Characterization of Human Medial Collateral Ligament," *Journal of Biomechanical Engineering*, Vol. 120, p. 757.

Regan, W. D., Korinek, S. L., Morrey, B. F., and An, K. N. (1991, October), "Biomechanical Study of Ligaments around the Elbow Joint," *Clinical Orthopaedics and Related Research*, Vol. 271, p. 170.

Reilly, C. H., and Marras, W. S. (1989), "Simulift: A Simulation Model of Human Trunk Motion," *Spine,* Vol. 14, p. 5.

Schoenmarklin, R. W., Marras, W. S., and Leurgans, S. E. (1994, September), "Industrial Wrist Motions and Incidence of Hand/Wrist Cumulative Trauma Disorders," *Ergonomics*, Vol. 37, p. 1449.

Schultz, A. B., and Andersson, G. B. (1981), "Analysis of Loads on the Lumbar Spine," *Spine*, Vol. 6, p. 76.

Shirazi-Adl , A., Ahmed, A. M., and Shrivastava, S. C. (1986), "A Finite Element Study of a Lumbar Motion Segment Subjected to Pure Sagittal Plane Moments," *Journal of Biomechanics*, Vol. 19, p. 331.

Silverstein, B. A., Stetson, D. S., Keyserling, W. M., and Fine, L. J. (1997), "Work-Related Musculoskeletal Disorders: Comparison of Data Sources for Surveillance," *American Journal of Industrial Medicine*, Vol. 31, p. 600.

Silverstein, M. A., Silverstein, B. A., and Franklin, G. M. (1996), "Evidence for Work-Related Musculoskeletal Disorders: A Scientific Counterargument," *Journal of Occupational and Environmental Medicine*, Vol. 38, p. 477.

Solomonow, M. (2004, February), "Ligaments: A Source of Work-Related Musculoskeletal Disorders," *Journal of Electromyography and Kinesiology*, Vol. 14, p. 49.

Solomonow, M., Zhou, B. H., Harris, M., Lu, Y., and Baratta, R. V. (1998), "The Ligamento-Muscular Stabilizing System of the Spine," *Spine*, Vol. 23, p. 2552.

Solomonow, M., Zhou, B. H., Baratta, R. V., Lu, Y., and Harris, M. (1999), "Biomechanics of Increased Exposure to Lumbar Injury Caused by Cyclic Loading: Part 1. Loss of Reflexive Muscular Stabilization," *Spine*, Vol. 24, p. 2426.

Solomonow, M., et al (2000), "Biexponential Recovery Model of Lumbar Viscoelastic Laxity and Reflexive Muscular Activity after Prolonged Cyclic Loading," *Clinical Biomechanics* (*Bristol, Avon*), Vol. 15, p. 167.

Solomonow, M., Zhou, B., Baratta, R. V., Zhu, M., and Lu, Y. (2002), "Neuromuscular Disorders Associated with Static Lumbar Flexion: A Feline Model," *Journal of Electromyography and Kinesiology*, Vol. 12, p. 81.

Stubbs, M., et al. (1998), "Ligamento-Muscular Protective Reflex in the Lumbar Spine of the Feline," *Journal of Electromyography and Kinesiology*, Vol. 8, p. 197.

Suwito, W., et al. (1992, March), "Geometric and Material Property Study of the Human Lumbar Spine Using the Finite Element Method," *Journal of Spinal Disordorders*, Vol. 5, p. 50.

Thelen, D. G., Schultz, A. B., and Ashton-Miller, J. A. (1995), "Co-contraction of Lumbar Muscles during the Development of Time-Varying Triaxial Moments," *Journal of Orthopaedic Research*, Vol. 13, p. 390.

Ticker, J. B., et al. (2006, November-December), "The Inferior Glenohumeral Ligament: A Correlative Investigation," *Journal of Shoulder and Elbow Surgery*, Vol. 15, p. 665.

van Dieen, J. H., and Kingma, I. (2005, March 15), "Effects of Antagonistic Co-contraction on Differences between Electromyography Based and Optimization Based Estimates of Spinal Forces," *Ergonomics*, Vol. 48, p. 411.

van Dieen, J. H., Hoozemans, M. J., and Toussaint, H. M. (1999), "Stoop or Squat: A Review of Biomechanical Studies on Lifting Technique," *Clinical Biomechanics* (*Bristol, Avon*), Vol. 14, p. 685.

Videman , T., Nurminen, M., and Troup, J. D. (1990), "1990 Volvo Award in Clinical Sciences. Lumbar Spinal Pathology in Cadaveric Material in Relation to History of Back Pain, Occupation, and Physical Loading," *Spine*, Vol. 15, p. 728.

Viegas, S. F., Yamaguchi, S., Boyd, N. L., and Patterson, R. M. (1999, May), "The Dorsal Ligaments of the Wrist: Anatomy, Mechanical Properties, and Function," *Journal of Hand Surgery America*, Vol. 24, p. 456.

Wang, J. L., Parnianpour, M., Shirazi-Adl, A., and Engin, A. E. (2000), "Viscoelastic Finite-Element Analysis of a Lumbar Motion Segment in Combined Compression and Sagittal Fexion. Effect of Loading Rate," *Spine*, Vol. 25, p. 310.

Waters, R. L., and Morris, J. M. (1973, July), "An In Vitro Study of Normal and Scoliotic Interspinous Ligaments," *Journal of Biomechanics*, Vol. 6, p. 343.

Waters, T. R., Putz-Anderson, V., Garg, A., and Fine, L. J. (1993), "Revised NIOSH Equation for the Design and Evaluation of Manual Lifting Tasks," *Ergonomics*, Vol. 36, p. 749.

Woo, S. L., Debski, R. E., Withrow, J. D., and Janaushek, M. A. (1999, July-August), "Biomechanics of Knee Ligaments," *American Journal of Sports Medicine*, Vol. 27, p. 533.

Yoganandan, N., Kumaresan, S., and Pintar, F. A. (2000, December), "Geometric and Mechanical Properties of Human Cervical Spine Ligaments," *Journal of Biomechanical Engineering*, Vol. 122, p. 623.

Yoganandan, N., Kumaresan, S., and Pintar, F. A. (2001, January), "Biomechanics of the Cervical Spine Part 2. Cervical Spine Soft Tissue Responses and Biomechanical Modeling," *Clinical Biomechanics* (*Bristol, Avon*), Vol. 16, p. 1.

Zander, T., Rohlmann, A., and Bergmann, G. (2004, January-February), "Analysis of Simulated Single Ligament Transection on the Mechanical Behaviour of a Lumbar Functional Spinal Unit," *Biomedical Technology* (*Berlin*), Vol. 49, p. 27.

Zhang, X., Kuo, A. D., and Chaffin, D. B. (1998), "Optimization-Based Differential Kinematic Modeling Exhibits a Velocity-Control Strategy for Dynamic Posture Determination in Seated Reaching Movements," *Journal of Biomechanics*, Vol. 31, p. 1035.

# PART 3
# DESIGN OF TASKS AND JOBS

# CHAPTER 13

# TASK ANALYSIS: WHY, WHAT, AND HOW

**Erik Hollnagel**
**University of Southern Denmark**
**Odense, Denmark**

## 1 THE NEED TO KNOW

The purpose of task analysis, as it is commonly practiced, is to describe tasks and more particularly to identify and characterize the fundamental characteristics of a specific activity or set of activities. According to the *Shorter Oxford Dictionary*, a *task* is "any piece of work that has to be done," which is generally taken to mean one or more functions or activities that must be carried out to achieve a specific goal. Task analysis can therefore be defined as the study of what people, individually and collectively, are required to do in order to achieve a given goal or objective.

Since a task by definition is a directed activity because it has a purpose or an objective, there is little or no methodological merit in speaking simply of activities and tasks without taking into account both their goals and the context in which they occur. Task analysis can therefore basically be defined as the study of what people, individually and collectively, are required to do to achieve a specific goal, or, simply put, as *who* does *what* and *why*.

**Who**    *Who* refers to the people who carry out a task. In the case where this is a single person, task analysis is a description of individual work. It is, however, far more common that people have to work together, in pairs, as a team, or in an organization, in which case task analysis is a description of the collective effort of what the team does. Whereas a description of individual work can focus on the activities, a description of collective work must also include how the collaboration is accomplished, that is, the organization and coordination of what the individuals do. Moving from the realm of human work to artifacts or agents (such as robots), task

analysis becomes the analysis of functions (e.g., movements) that an artifact must carry out to achieve a goal.

In industrialized societies, tasks are in most cases accomplished by people using some kind of technological artifact or system: in other words, a human–machine system. Task analysis is therefore often focused on the *what* of the human–machine system as such should do: for instance, as task analysis for human–computer interaction (e.g., Diaper and Stanton, 2003). More generally, humans and machines working together can be described as *cognitive systems* or *joint cognitive systems* (Hollnagel and Woods, 2005). Indeed, at the time of writing (2010) the main issues are no longer human work with technology, or human interaction with technology, but the *coagency* of multiple functions, providers, stakeholders, and so on. Human work with technology is no longer a question of human–technology interaction but rather a question of how complex sociotechnical systems function. The human–machine dyad and the focus on human–machine interaction are relics from the early days of human factors and are no longer adequate—if they have not already become irrelevant.

The built-in assumptions about the nature of who carries out the task have important consequences for task analysis, as will be clear from the following. The use of the pronoun *who* should not be taken to mean that task analysis is only about what humans do, although that was the original objective. In contemporary terms it would probably be more appropriate to refer to the *system* that carries out the task.

**What**    *What* refers to the contents of the task and is usually described in terms of the activities that constitute the task. Task analysis started by focusing on physical

tasks (i.e., manifest work) but has since the 1970s enlarged its scope to include cognitive or mental tasks. The content of the task thus comprises a systematic description of the activities or functions that make up the task, either in terms of observable actions (e.g., grasping, holding, moving, assembling) or in terms of the usually unobservable functions that may lie behind these actions, commonly referred to as *cognition* or *cognitive functions*.

**Why**    Finally, *why* refers to the purpose or goal of the task: for instance, the specific system state or condition that is to be achieved. A goal may be something that is objective and physically measurable (a product) but also something that is subjective: for instance, a psychological state or objective, such as "having done a good job." The task analysis literature has usually eschewed the subjective and affective aspects of tasks and goals, although they clearly are essential for understanding human performance as well as for designing artifacts and work environments.

Task analysis is supposed to provide concrete answers to the *practical* questions of how things should be done or are done. When dealing with work, and more generally with how people use sociotechnical artifacts to do their work, it is necessary to know both what activities (functions) are required to accomplish a specified objective and how people habitually go about doing them, particularly since the latter is usually different—and sometimes significantly different—from the former. Such knowledge is necessary to design, implement, and manage sociotechnical systems, and task analysis looks specifically at how work takes place and how it can be facilitated. Task analysis therefore has applications that go well beyond interface and interaction design and may be used to address issues such as training, performance assessment, event reporting and analysis, function allocation and automation, procedure writing, maintenance planning, risk assessment, staffing and job organization, personnel selection, and work management.

The term *task analysis* is commonly used as a generic label. A survey of task analysis methods shows that they represent many different meanings of the term (Kirwan and Ainsworth, 1992). A little closer inspection, however, reveals that they fall into a few main categories:

- The analysis and description of tasks or working situations that do not yet exist or are based on hypothetical events
- The description and analysis of observations of how work is carried out or of event reports (e.g., accident investigations)
- The representation of either of the above, in the sense of the notation used to capture the results (of interest due to the increasing use of computers to support task analysis)
- The various ways of further analysis or refinement of data about tasks (from either of the foregoing sources)

- The modes of presentation of results and the various ways of documenting the outcomes

Methods of task analysis should in principle be distinguished from methods of task description. A task description produces a generalized account or summary of activities as they have been carried out. It is based on empirical data or observations rather than on design data and specifications. A classical example is link analysis or even hierarchical task analysis (Annett et al., 1971). Properly speaking, task description or performance analysis deals with actions rather than with tasks. This distinction, by the way, is comparable to the French ergonomic tradition where the described task is seen as different from the effective task. The described task (*tâche prévue* or simply *tâche*) is the intended task or what the organization assigns to the person, what the person should do. The effective task (*tâche effective* or *activité*) is the actual task or the person's response to the prescribed task, what the person actually does (Daniellou, 2005). Understanding the task accordingly requires an answer to the question of *what* the person does, while understanding the activity requires an answer regarding *how* the person performs the task. (In practice, it is also necessary to consider *when* and *where* the task is carried out.) An important difference between tasks and activities is that the latter are dynamic and may change depending on the circumstances, such as fluctuations in demands and resources, changing physical working conditions, the occurrence of unexpected events, and so on. The distinction between the task described and the effective task can be applied to both individual and collective tasks (Leplat, 1991).

## 1.1   Role of Task Analysis in Human Factors

Task analysis has over the years developed into a stable set of methods that constitute an essential part of human factors and ergonomics as applied disciplines. The focus of human factors engineering or ergonomics is humans at work, more particularly the human use of technology in work, although it sometimes may look more like technology's use of humans. The aim of *human factors* (which in the following is used as a common denominator for human factors engineering and ergonomics) is to apply knowledge about human behavior, abilities, and limitations to design tools, machines, tasks, and work environments to be as productive, safe, healthy, and effective as possible. From the beginning, ergonomics was defined broadly as the science of work (Jastrzebowski, 1857). At that time work was predominantly manual work, and tools were relatively few and simple. Human factors, which originally was called human factors engineering, came into existence around the mid-1940s as a way of solving problems brought on by emerging technologies such as computerization and automation. Ergonomics and human factors thus started from different perspectives but are now practically synonymous. At the present time, literally every type of work involves the use of technology, and the difference between ergonomics and human factors engineering is rather nominal.

Throughout most of history, people have depended on tools or artifacts to do their work, such as the

painter's brush or the blacksmith's hammer. As long as users were artisans rather than workers and work was an individual rather than a collective endeavor, the need of prior planning, and therefore of anything resembling task analysis, was limited or even nonexistent. The demand for task analysis arose when the use of technology became more widespread, especially when tools changed from being simple to being complex. More generally, the need of a formal task analysis arises when one or more of the following three conditions are met:

1. The accomplishment of a goal requires more effort than one person can provide or depends on a combination of skills that goes beyond what a single individual can be expected to master. In such cases task analysis serves to break down a complex and collective activity to descriptions of a number of simpler and more elementary activities. For example, building a ship, in contrast to building a dinghy or a simple raft, requires the collaboration of many individuals and the coordination of many different types of work. In such cases people have to collaborate and must therefore adjust their own work to match the progress and demands of others. Task analysis is needed to identify the task components that correspond to what a person can achieve or provide over a reasonable period of time as well as to propose a way to combine and schedule the components to an overall whole.

2. Tasks become so complex that one person can no longer control or comprehend them. This may happen when the task becomes so large or takes so long that a single person is unable to complete it (i.e., the transition from individual to collective tasks). It may also happen when the execution of the task depends on the use of technological artifacts and where the use of the artifact becomes a task in its own right (cf. below). This is the case, for instance, when the artifacts can function in an independent or semiautonomous way (i.e., they begin partially to regulate themselves rather than passively carry out an explicit function under the user's control).

3. A similar argument goes when technology itself—machines—becomes so complex that the situation changes from simply being one of *using* the technology to one learning how to *understand*, *master*, or *control* the technology. In other words, being in control of the technology becomes a goal in itself, as a means to achieve the original goal. Examples are driving a car in contrast to riding an ordinary bicycle, using a food processor instead of a knife, using a computer (as in writing this chapter) rather than paper and pencil, and so on. In these cases, and of course also in cases of far more complex work, use of the technology is no longer straightforward but requires preparation and prior thought either by the person who does the task or work or by those who prepare tasks or tools for others. Task analysis can in these cases be used to describe situations where the task itself is very complex because it involves interaction and dependencies with other people. It can similarly be used to describe situations where use of the technology is no longer straightforward but requires mastery of the system to such a degree that not everyone can apply it directly as intended and designed.

In summary, task analysis became necessary when work changed from something that could be done by an unaided individual to something requiring the collective efforts of either people or joint cognitive systems. Although collective work has existed since the beginning of history, its presence became more conspicuous after the Industrial Revolution about 250 years ago. In this new situation the work of the individual became a mere part of the work of the collective, and individual control of work was consequently lost. The worker became part of a larger context, a cog in complex social machinery that defined the demands and constraints to work. One important effect of that was that people no longer could work at a pace suitable for them and pause whenever needed but instead had to comply with the pace set by others—and increasingly the pace set by machines.

## 1.2  Artifacts and Tools

It is common to talk about humans and machines, or humans and technology, and to use expressions such as *human−machine systems* or, even better, *human−technology systems*. In the context of tasks and task analysis, the term *technological artifact*, or simply *artifact*, will be used to denote that which is being applied to achieve a goal. Although it is common to treat computers and information technology as primary constituents of the work environment, it should be remembered that not all tools are or include computers, and task analysis is therefore far more than human−computer interaction. That something is an artifact means that it has been constructed or designed by someone, hence that it expresses or embodies a specific intention or purpose. In contrast to that, a natural object does not have an intended use, but is the outcome of evolution—or happenstance—rather than design. Examples of natural objects are stones used to hammer or break something and sticks used to poke for something.

A natural object may be seen as being instrumental to achieve something, hence used for that purpose. In the terminology of Gibson (1979), the natural object is perceived as having an actionable property (an affordance), which means that it is seen as being useful for a specific purpose. An artifact is designed with a specific purpose (or set of uses) in mind and should ideally offer a similar perceived affordance. To the extent that this is the case, the design has been successful. Task analysis (i.e., describing and understanding in advance the uses of artifacts) is obviously one of the ways in which that can be achieved. Although there is no shortage of examples of failure to achieve this

noble goal, it is of course in the best interest of the designer—and the producer—to keep trying.

When a person designs or constructs something for himself or herself an artifact or a composite activity, there is no need to ask what the person is capable of, what the artifact should be used for, or how it should be used. But the need is there in the case of a single but complex artifact where the use requires a series of coordinated or ordered actions. It is also there in the case of more complex, organized work processes where the activities or tasks of an individual must fit into a larger whole. Indeed, just as the designer of a complex artifact considers its components and how they must work together for the artifact to be able to provide its function, so must the work process designer consider the characteristics of people and how they must collaborate to deliver the desired end product or result. It is, indeed, no coincidence that the first task analyses were made for organized work processes rather than for the single users working with artifacts or machines.

From an analytical perspective, the person's knowledge of what he or she can do can be seen as corresponding to the designer's assumptions about the user, while the person's knowledge of how the artifact should be used can be seen as corresponding to the user's assumptions about the artifact, including the designer's intentions. As long as the artifact or the work processes are built around the person, there is little need to make any of these assumptions explicit or indeed to produce a formal description of them: The user and the designer are effectively the same person. There is also little need of prior thought or prior analysis since the development is an integral part of work rather than an activity that is separated in time and space. But when the artifact is designed by one person to be used by someone else, the designer needs to be very careful and explicit in making assumptions and to consider carefully what the future user may be able to do and will do. In other words, it is necessary in these cases to analyze how the artifact will be used or to perform a task analysis.

## 2   TASK TYPES AND TASK BREAKDOWN

Task analysis is in the main a collection of methods which describe (or prescribe) how the analysis will be performed, preferably by describing each step of the analysis as well as how they are organized. Each method should also describe the *stop rule* or criterion (i.e., define the principles needed to determine when the analysis has come to an end, for instance, that the level of elementary tasks has been reached).

An important part of the method is to name and identify the main constituents of a task and how they are organized. As described later in this chapter, task analysis has through its development embraced several different principles of task organization, of which the main ones are the sequential principle, the hierarchical principle, and the functional dependency principle. To do so, the method must obviously refer to a classification scheme or set of categories that can be used to describe and represent the essential aspects of a task.

The hallmark of a good method is that the classification scheme is applied consistently and uniformly, thereby limiting the opportunities for subjective interpretations and variations. Task analysis should depend not on personal experience and skills but on generalized public knowledge and common sense. A method is also important as a way of documenting how the analysis has been done and of describing the knowledge that was used to achieve the results. It helps to ensure that the analysis is carried out in a systematic fashion so that it, if needed, can be repeated, hopefully leading to the same results. This reduces the variability between analysts and hence improves the reliability.

The outcome of the task analysis accounts for the organization or structure of constituent tasks. A critical issue is the identification or determination of elementary activities or task components. The task analysis serves among other things to explain how something should be done for a user who does not know what to do or who may be unable to remember it (in the situation). There is therefore no need to describe things or tasks that the user definitely knows. The problem is, however, how that can be determined.

Task analysis has from the very start tried to demarcate the basic components. It would clearly be useful if it was possible to find a set of basic tasks—or activity atoms—that could be applied in all contexts. This is akin to finding a set of elementary processes or functions from which a complex behavior can be built. Such endeavors are widespread in the behavioral and cognitive sciences, although the success rate usually is quite limited. The main reason is that the level of an elementary task depends on the person as well as on the domain. Even if a common denominator could be found, it would probably be at a level of detail as to have little practical value (e.g., for training or scheduling).

### 2.1   Changing Views of Elementary Tasks

Although the search for all-purpose task components is bound to fail, it is nevertheless instructive to take a brief look at three different attempts to do so. Probably the first—and probably also the most ambitious—attempt was made by Frank Bunker Gilbreth, one of the pioneers of task analysis. The categorization, first reported about 1919, evolved from the observation by trained motion-and-time specialists of human movement, specifically of the fundamental motions of the hands of a worker. Gilbreth found that it was possible to distinguish among the following 17 types of motion: search, select, grasp, reach, move, hold, release, position, pre-position, inspect, assemble, disassemble, use, unavoidable delay, wait (avoidable delay), plan, and rest (to overcome fatigue). (The basic motions are known as *therbligs*, using an anagram of the developer's name.)

A more contemporary version is a list of typical process control tasks suggested by Rouse (1981). This comprises 11 functions, which are in alphabetical order: communicating, coordinating tasks, executing procedures, maintaining, planning, problem solving, recognizing, recording, regulating, scanning, and steering. In contrast to the *therbligs*, it is possible to organize these functions in several ways: for instance, in relation

to an input–output model of information processing, in relation to a control model, and in relation to a decision-making model. The functions proposed by Rouse are characteristically on a higher level of abstraction than the *therbligs* and refer to cognitive functions, or cognitive tasks, rather than to physical movements.

A final example is the GOMS model proposed by Card et al. (1983). The purpose of GOMS, which is an acronym that stands for "goals, operators, methods, and selection rules," was to provide a system for modeling and describing human task performance. Operators, one of the four components of GOMS, denote the set of atomic-level operations from which a user can compose a solution to a goal, while methods represent sequences of operators grouped together to accomplish a single goal. For example, the manual operators of GOMS are: Keystroke key_name, Type_in string_of_characters, Click mouse_button, Double_click mouse_button, Hold_down mouse button, Release mouse_button, Point_to target_object, and Home_to destination. These operators refer not to physical tasks, such as the *therbligs*, but rather to mediating activities for mental or cognitive tasks.

The definition of elementary tasks in scientific management could comfortably refer to what people did, hence to what could be reported by independent observers. The problem with defining elementary cognitive or mental tasks is that no such independent verification is possible. Although GOMS was successful in defining elementary tasks on the keystroke level, it was more difficult to do the same for the cognitive or mental aspects (i.e., the methods and selection rules). The physical reality of elementary tasks such as *grasp, reach, move*, and *hold* has no parallel when it comes to cognitive functions. The problems in identifying elementary mental tasks are not due to a lack of trying. This has indeed been a favorite topic of psychology from Donders (1969; orig. 1868) to Simon (1972). The problems come about because the "smallest" unit is defined by the theory being used rather than by intersubjective reality. In practice, this means that elementary tasks must be defined relative to the domain at the time of the analysis (i.e., in terms of the context rather than as absolutes or context-free components).

## 3  BRIEF HISTORY OF TASK ANALYSIS

Task analysis has a relatively short history starting around the beginning of the twentieth century. The first major publications were Gilbreth (1911) and Taylor (1911), which introduced the principles of scientific management. The developments that followed reflected both the changing view of human nature, for instance, in McGregor's (1960) theory X and theory Y, and the changes in psychological schools, specifically the models of the human mind. Of the three examples of a classification system mentioned above, Gilbreth (1911) represents the scientific management view, Rouse (1981) represents the supervisory control view (human–machine interaction), and Card et al. (1983) represents the information-processing

view (human–computer interaction). These views can be seen as alternative ways of describing the same reality: namely, human work and human activities. One standpoint is that human nature has not changed significantly for thousands of years and that different descriptions of the human mind and of work therefore only represent changes in the available models and concepts. Although this undoubtedly is true, it is also a fact that the nature of work has changed due to developments in technology. Gilbreth's description in terms of physical movements would therefore be as inapplicable to today's work as a description of cognitive functions would have been in 1911.

### 3.1  Sequential Task Analysis

The dawn of task analysis is usually linked to the proposal of a system of scientific management (Taylor, 1911). This approach was based on the notion that tasks should be specified and designed in minute detail and that workers should receive precise instructions about how their tasks should be carried out. To do so, it was necessary that tasks could be analyzed unequivocally or "scientifically," if possible in quantitative terms, so that it could be determined how each task step should be done in the most efficient way and how the task steps should be distributed among the people involved.

One of the classical studies is Taylor's (1911) analysis of the handling of pig iron, where the work was done by men with no "tools" other than their hands. A pig-iron handler would stoop down, pick up a pig weighing about 92 pounds, walk up an inclined plank, and drop it on the end of a railroad car. Taylor and his associates found that a gang of pig-iron handlers was loading on the average about $12\frac{1}{2}$ long tons per man per day. The aim of the study was to find ways in which to raise this output to 47 tons a day, not by making the men work harder but by reducing the number of unnecessary movements. This was achieved both by careful motion-and-time studies and by a system of incentives that would benefit workers as well as management.

Scientific management was based on four elements or principles, which were used in studies of work.

1. The development of the science of work with *rigid* rules for each motion of every person and the perfection and *standardization* of all implements and working conditions

2. The careful *selection* and subsequent *training* of workers into first-class people and the elimination of all people who refuse to or are unable to adopt the best methods

3. Bringing the first-class workers and the science of working together through the constant help and watchfulness of management and through paying each person a large daily bonus for working fast and doing what he or she is told to do

4. An almost equal division of the work and responsibility between workers and management

Of these four elements, the first (the development of the science of work) is the most interesting and

spectacular. It was essentially an analysis of a task into its components, using, for example, the list of *therbligs* mentioned above. In the case of manual work this was entirely feasible, since the task could be described as a single sequence of more detailed actions or motions. The motion-and-time study method was, however, unable to cope with the growing complexity of tasks that followed developments in electronics, control theory, and computing during the 1940s and 1950s. Due to the increasing capabilities of machines, people were asked—and tasked—to engage in multiple activities at the same time, either because individual tasks became more complex or because simpler tasks were combined into larger units. An important consequence of this was that tasks changed from being a sequence of activities referring to a single goal to an organized set of activities referring to a hierarchy of goals. The use of machines and technology also became more prevalent, so that simple manual work such as pig-iron handling was taken over by machines, which in turn were operated or controlled by workers.

Since the use of technology has made work environments more complex, relatively few tasks today are sequential tasks. Examples of sequential tasks are therefore most easily found in the world of cooking. Recipes are typically short and describe the steps as a simple sequence of actions, although novice cooks sometimes find that recipes are underspecified. As an example of a sequential task analysis, Figure 1 shows the process for baking Madeleines.

### 3.2 From Sequential to Hierarchical Task Organization

The technological development meant that the nature of work changed from being predominantly manual and became more dependent on mental capabilities (comprehension, monitoring, planning). After a while, human factors engineering, or classical ergonomics,
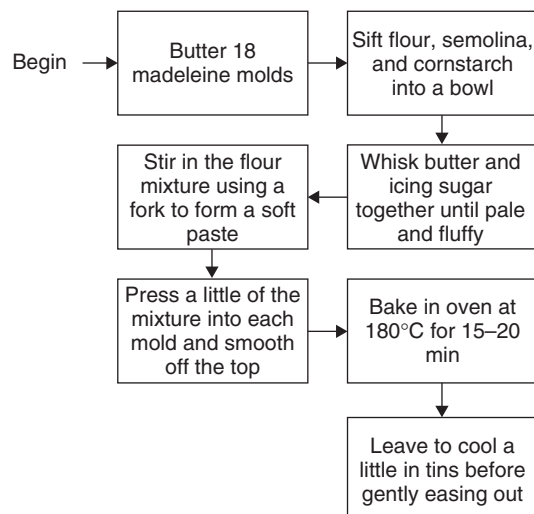
recognized that traditional methods of breaking the task down into small pieces, where each could be performed by a person, were no longer adequate. Since the nature of work had changed, the human capacity for processing information became decisive for the capacity of the human–machine system. This capacity could not be extended beyond its "natural" upper limit, and it soon became clear that the human capacity for learning and adaptation was insufficient to meet technological demands.

To capture the more complex task organization, Miller (1953) developed a method for human–machine task analysis in which main task functions could be decomposed into subtasks. Each subtask could then be described in detail, for instance, by focusing on information display requirements and control actions. This led to the following relatively simple and informal procedure for task analysis:

1. Specify the human–machine system criterion output.
2. Determine the system functions.
3. Trace each system function to the machine input or control established for the operator to activate.
4. For each respective function, determine what information is displayed by the machine to the operator whereby he or she is directed to appropriate control activation (or monitoring) for that function.
5. Determine what indications of response adequacy in the control of each function will be fed back to the operator.
6. Determine what information will be available and necessary to the operator from the human–machine "environment."
7. Determine what functions of the system must be modulated by the operator at or about the same time, or in close sequence, or in cycles.
8. In reviewing the analysis, be sure that each stimulus is linked to a response and that each response is linked to a stimulus.

The tasks were behavior groups associated with combinations of functions that the operator should carry out. These were labeled according to the subpurpose they fulfilled within the system. Point 8 reflects the then-current psychological thinking, which was that of stimulus-and-response couplings. The operator was, in other words, seen as a transducer or a machine that was coupled to the "real" machine. For the human–machine system to work, it was necessary that the operator interpret the machine's output in the proper way and that he or she respond with the correct input. The purpose of task analysis was to determine what the operator had to do to enable the machine to function as efficiently as possible.

#### 3.2.1 Task–Subtask Relation

The task–subtask decomposition was a significant change from sequential task analysis and was necessitated by the growing complexity of work. The



**Figure 1** Sequential task description.

development was undoubtedly influenced by the emerging practice—and later science—of computer programming, where one of the major innovations was the subroutine. Arguably the most famous example of a task–subtask relation is the TOTE (test–operate–test–exit), which was proposed as a building block of human behavior (Miller et al., 1960). This introduced into the psychological vocabulary the concept of a plan, which is logically necessary to organize combinations of tasks and subtasks. Whereas a subroutine can be composed of motions and physical actions, and hence in principle can be found even in scientific management, a plan is obviously a cognitive or mental component. The very introduction of the task–subtask relation, and of plans, therefore changed task analysis from describing only what happened in the physical world to describing what happened in the minds of the people who carried out the work.

Miller's task–subtask analysis method clearly implied the existence of a hierarchy of tasks and subtasks, although this was never a prominent feature of the method. As the technological environments developed further, the organization of tasks and subtasks became increasingly important for task analysis, culminating with the development of hierarchical task analysis (HTA) (Annett and Duncan, 1967; Annett et al., 1971). Since its introduction HTA has become the standard method for task analysis and task description and is widely used in a variety of contexts, including interface design.

The process of HTA is to decompose tasks into subtasks and to repeat this process until a level of elementary tasks has been reached. Each subtask or operation is specified by its goal, the conditions under which the goal becomes relevant or "active," the actions required to attain the goal, and the criteria that mark the attainment of the goal. The relationship between a set of subtasks and the superordinate task is governed by plans expressed as, for instance, procedures, selection rules, or time-sharing principles. A simple example of HTA is a description of how to get money from a bank account using an ATM (see Figure 2). In this description, there is an upper level of tasks (marked, 1, 2, 3), which describe the order of the main segments, and

a lower level of subtasks (marked 1.1, 1.2, etc.), which provide the details. It is clearly possible to break down each of the subtasks into further detail, for instance, by describing the steps comprised by *1.2 Enter PIN code*. This raises the nontrivial question of when the HTA should stop (i.e., what the elementary subtasks or task components are; cf. below).

The overall aim of HTA is to describe a task in sufficient detail, where the required level of resolution depends on the specific purposes (e.g., interaction design, training requirements, interface design, risk analysis). HTA can be seen as a systematic search strategy adaptable for use in a variety of different contexts and purposes within the field of human factors (Shepherd, 1998). In practice, performing HTA comprises the following steps. (Note, by the way, that this is a sequential description of hierarchical task analysis!)

1. Decide the purpose of the analysis.
2. Get agreement between stakeholders on the definition of task goals and criterion measures.
3. Identify sources of task information and select means of data acquisition.
4. Acquire data and draft a decomposition table or diagram.
5. Recheck the validity of the decomposition with the stakeholders.
6. Identify significant operations in light of the purpose of the analysis.
7. Generate and, if possible, test hypotheses concerning factors affecting learning and performance.

Whereas Miller's description of human–machine task analysis concentrated on how to analyze the required interactions between humans and machines, HTA extended the scope to consider the context of the analysis, in particular which purpose it served. Although this was a welcome and weighty development, it left the actual HTA somewhat underspecified. Indeed, in the description above it is only the fourth step that is the actual task analysis.
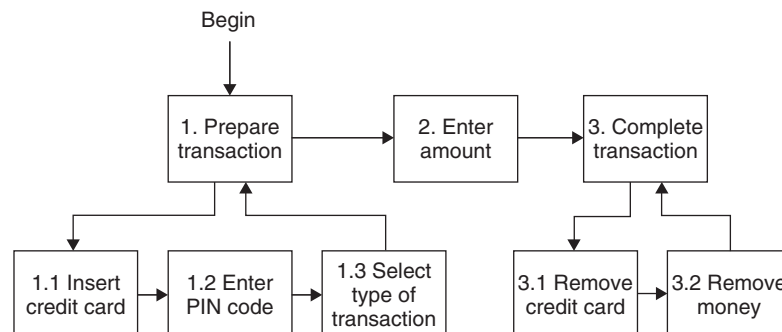


**Figure 2** Hierarchical task description.

### 3.2.2  Tasks and Cognitive Tasks

In addition to the change that led from sequential motion-and-time descriptions to hierarchical task organization, a further change occurred in the late 1980s to emphasize the cognitive nature of tasks. The need to consider the organization of tasks was partly a consequence of changing from a sequential to a hierarchical description, as argued above. The changes in the nature of work also meant that "thinking" tasks became more important than "doing." The need to understand the cognitive activities of the human–machine system, first identified by Hollnagel and Woods (1983), soon developed a widespread interest in cognitive task analysis, defined as the extension of traditional task analysis techniques to yield information about the knowledge, thought processes, and goal structures that underlie observable task performance (e.g., Schraagen et al., 2000). As such, it represents a change in emphasis from overt to covert activities. An example is the task analysis principles described by Miller et al. (1960), which refer to mental actions as much as to motor behavior. Since many tasks require a considerable amount of mental functions and effort, in particular in retrieving and understanding the information available and in planning and preparing what to do (including monitoring of what happens), much of what is essential for successful performance is covert. Whereas classical task analysis relies very much on observable actions or activities, the need to find out what goes on in other peoples' minds requires other approaches.

One consequence of the necessary extension of task analysis from physical to cognitive tasks was the realization that both the physical and the cognitive tasks were affected by the way the work situation was designed. Every artifact we design has consequences for how it is used. This goes for technological artifacts (gadgets, devices, machines, interfaces, complex processes) as well as social artifacts (rules, rituals, procedures, social structures and organizations). The consequences can be seen in the direct and concrete (physical) interaction with the artifact (predominantly manual work) as well as in how the use of the artifact is planned and organized (predominantly cognitive work). Introducing a new "tool" therefore affects not only how work is done but also how it is conceived of and organized. Yet interface design and instruction manuals and procedures typically describe how an artifact should be used but not how we should plan or organize the use of it even though the latter may be affected as much—or even more—than the former. The extension of task analysis to cognitive task analysis should therefore be matched by a corresponding extension of task design to cognitive task design (Hollnagel, 2003).

### 3.2.3  Elementary Task

All task analysis methods require an answer to what the elementary task is. As long as task analysis was occupied mainly with physical work, the question could be resolved in a pragmatic manner. But when task analysis changed to include the cognitive aspects of work, the answer became more contentious. This is obvious from the simple example of a HTA shown in Figure 2. For a person living in a developed or industrialized society, entering a PIN code can be assumed to be an elementary task. It can nevertheless be broken down into further detail by, for example, a motion-and-time study or a GOMS-type interaction analysis. The determination of what an elementary task is clearly cannot be done separately from assumptions about who the users are, what the conditions of use (or work) are, and what the purpose of the task analysis is. If the purpose is to develop a procedure or a set of instructions such as the instructions that appear on the screen of an ATM, there may be no need to go further than "enter PIN code" or possibly "enter PIN code and press ACCEPT." Given the population of users, it is reasonable for the system designer to take for granted that they will know how to do this. If, however, the purpose is to design the physical interface itself or to perform a risk analysis, it will be necessary to continue the analysis at least one more step. GOMS is a good example of this, as would be the development of instructions for a robot to use an ATM.

In the contexts of work, assumptions about elementary tasks can be satisfied by ensuring that users have the requisite skills (e.g., through training and instruction). A task analysis may indeed be performed with the explicit purpose of defining training requirements. Designers can therefore, in a sense, afford themselves the luxury of dictating what an elementary task is as long as the requirements can be fulfilled by training. In the context of artifacts with a more widespread use, typically in the public service domain, greater care must be taken in making assumptions about an elementary task, since users in these situations often are "accidental" (Marsden and Hollnagel, 1996).

### 3.3  Functional Dependency and Goals–Means Task Analysis

Both sequential and hierarchical task analyses are structural in the sense that they describe the order in which the prescribed activities are to be carried out. A hierarchy is by definition the description of how something is ordered, and the very representation of a hierarchy (as in Figure 2) emphasizes the structure. As an alternative, it is possible to analyze and describe tasks from a functional point of view (i.e., in terms of how tasks relate to or depend on each other). This changes the emphasis from how tasks and activities are ordered to what the tasks and activities are supposed to achieve.

Whereas task analysis in practice stems from the beginning of the twentieth century, the principle of functional decomposition can be traced back at least to Aristotle (Book III of the *Nicomachean Ethics).* This is not really surprising, since the focus of a functional task analysis is the reasoning about tasks rather than the way in which they are carried out (i.e., the physical performance). Whereas the physical nature of tasks has changed throughout history, and especially after the beginning of the Industrial Revolution, thinking about how to do things is largely independent of how things are actually done.

In relation to task analysis, functional dependency means thinking about tasks in terms of goals and means. The strength of a goals–means, or means–ends, decomposition principle is that it is ubiquitous, important, and powerful (Miller et al., 1960, p. 189). It has therefore been used widely, most famously as the basis for the General Problem Solver (Newell and Simon, 1961).

The starting point of a functional task analysis is a goal or an end, defined as a specified condition or state of the system. A description of the goal usually includes or implies the criteria of achievement or acceptability (i.e., the conditions that determine when the goal has been reached). To achieve the goal, certain means are required. These are typically one or more activities that need to be carried out (i.e., a task). Yet most tasks are possible only if specific conditions are fulfilled. For instance, you can work on your laptop only if you have access to an external power source or if the batteries are charged sufficiently. When these conditions are met, the task can be carried out. If not, bringing about these preconditions becomes a new goal, denoted a subgoal. In this way goals are decomposed recursively, thereby defining a set of goal–subgoal dependencies that also serves to structure or organize the associated tasks.

An illustration of the functions or tasks needed to start up an industrial boiler is shown in Figure 3

(see Lind and Larsen, 1995). The diagram illustrates how the top goal, "St1 established," requires that a number of conditions have been established, where each of these in turn can be described as subgoals. Although the overall structure is a hierarchical ordering of goals and means, it differs from a HTA because the components of the diagram are goals rather than tasks. The goals–means decomposition can be used as a basis for identifying the tasks that are necessary to start the boiler, but this may not necessarily fit into the same representation.

## 4 PRACTICE OF TASK ANALYSIS

As already mentioned, task analysis can be used for a variety of purposes. Although the direct interaction between humans and computers got the lion's share of attention in the 1990s, task analysis is necessary for practically any aspect of a human–machine system's functioning. Task analysis textbooks, such as Kirwan and Ainsworth (1992), provide detailed information and excellent descriptions of the many varieties of task analysis. More recent works, such as Hollnagel (2003), extend the scope from task analysis to task design, emphasizing the constructive use of task knowledge. Regardless of which method an investigator decides to
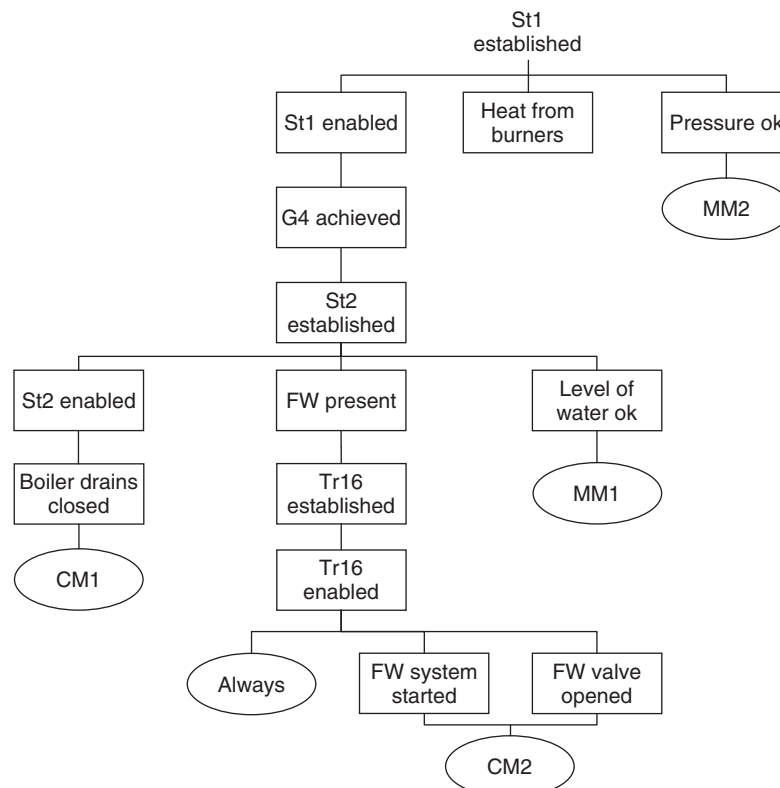
**Figure 3** Goals–means task description.

use, there are a number of general aspects that deserve consideration.

## 4.1  Task Data Collection Techniques

The first challenge in task analysis is to know where relevant data can be found and to collect them. The behavioral sciences have developed many ways of doing this, such as activity sampling, critical incident technique, field observations, questionnaire, structured interview, and verbal protocols. In many cases, data collection can be supported by various technologies, such as audio and video recording, measurements of movements, and so on, although the ease of mechanical data collection often is offset by the efforts needed to analyze the data.

As task analysis extended its scope from physical work to include cognitive functions, methods were needed to get data about the unobservable parts of a task. The main techniques used to overcome this were "think-aloud" protocols and introspection (i.e., extrapolating from one's own experience to what others may do). The issue of thinking aloud has been hotly debated, as has the issue of introspection (Nisbett and Wilson, 1977). Other structured techniques rely on controlled tasks, questionnaires, and so on. Yet in the end the problem is that of making inferences from some set of observable data to what goes on behind, in the sense of what is sufficient to explain the observations. This raises interesting issues of methods for data collection to support task analysis and leads to an increasing reliance on models of the tasks. As long as task analysis is based on observation of actions or performance, it is possible to establish some kind of objectivity or intersubjective agreement or verification. As more and more of the data refer to the unobservable, the dependence on interpretations, and hence on models, increases.

## 4.2  Task Description Techniques

When the data have been collected, the next challenge is to represent them in a suitable fashion. It is important that a task analysis represent the information about the task in a manner that can easily be comprehended. For some purposes, the outcome of a task analysis may simply be rendered as a written description of the tasks and how they are organized. In most cases this is supplemented by some kind of graphical representation or diagram, since this makes it considerably easier to grasp the overall relations. Examples are the diagrams shown in Figures 1–3. Other staple solutions are charting and networking techniques, decomposition methods, HTA, link analysis, operational sequence diagrams (OSDs), and timeline analyses.

## 4.3  Task Simulation Methods

For a number of other purposes, such as those that have to do with design, it is useful if the task can be represented in other ways, specifically as some kind of description or model that can be manipulated. The benefit is clearly that putative changes to the task can be implemented in the model and the consequences can be explored. This has led to the development of a range of methods that rely on some kind of symbolic model of the task or activity, going from the production rule systems to task networks (e.g., Petri nets). This development often goes hand in hand with user models (i.e., symbolic representation of users that can be used to simulate responses to what happens in the work environment). In principle, such models can carry out the task as specified by the task description, but the strength of the results depends critically on the validity of the model assumptions. Other solutions, which do not require the use of computers, are mock-ups, walk-throughs, and talk-throughs.

## 4.4  Task Behavior Assessment Methods

Task analyses are in many cases used as a starting point to look at a specific aspect of the task execution, usually risk or consequences for system safety. One specific type of assessment looks at the possibility for humans to carry out a task incorrectly (i.e., the issue of "human error" and human reliability). Approaches to human reliability analysis that are based on structural task descriptions are generally oversimplified not only because humans are not machines but also because there is an essential difference between described and effective tasks or between "work as imagined" and "work as done." Task descriptions in the form of event trees or as procedural prototype models represent an idealized sequence or hierarchy of steps. Tasks as they are carried out or as they are perceived by the person are more often series of activities whose scope and sequence are adjusted to meet the demands—perceived or real—of the current situation (Hollnagel, 2010a). It can be argued that task descriptions used for risk and reliability analyses on the whole are inadequate and unable to capture the real nature of human work. The decomposition principle has encouraged—or even enforced—a specific form of task description (the event tree), and this formalism has been self-sustaining. It has, however, led human reliability analysis into a cul-de-sac.

## 4.5  Future of Task Analysis

We started this chapter by pointing out that task analysis is the study of *who* does *what* and *why*, where the *who* should be broadened to include individual work, collective work, and joint cognitive systems. The future of task analysis is bright in the sense that there will always be a practical need to know how things should be done. The question is whether task analysis as it is currently practiced is capable of meeting this need in the long run. There are several reasons why the reply need not be unequivocally positive:

1. Task analysis has from the beginning been concerned mostly with individuals, whether as single workers or single users, despite the fact that most work involves multiple users (collaboration, distributed work) in complex systems (Hutchins, 1995). Although the importance of distributed cognition and collective work is generally acknowledged, only few methods are capable of analyzing that, over and above

**Table 1 Tractable and Intractable Systems**

| | Tractable System | Intractable System |
|---|---|---|
| Number of details | Descriptions are simple with few details | Descriptions are elaborate with many details |
| Comprehensibility | Principles of functioning are known | Principles of functioning are partly unknown |
| Stability | System does not change while being described | System changes before description is completed |
| Relationship to other systems | Independence | Interdependence |
| Organization of tasks and activities | Stable, work is highly regular and tasks can be prescribed | Unstable, work must be adjusted to match the conditions, tasks cannot be prescribed |

representing explicit interactions such as in link analysis and OSDs.

2. Many task analysis methods are adequate for describing single lines of activity. Unfortunately, most work involves multiple threads and timelines. Although HTA represents a hierarchy of tasks, each subtask or activity is carried out on its own. There is little possibility of describing two or more simultaneous tasks, even though that is often what people have to cope with in reality. Another shortcoming is the difficulty of representing temporal relations other than simple durations of activities.

3. There is a significant difference between described and effective tasks. Work in practice is characterized by ongoing adaptations and improvisations rather than the straightforward carrying out of a procedure or an instruction. The reasons for this are that demands and resources rarely correspond to what was anticipated when the task was developed and the actual situation may differ considerably from that which is assumed by the task description, thereby rendering the latter unworkable.

The problem in a nutshell is that task analysis was developed to deal with linear work environments, where effects were proportional to causes and where orderliness and regularity on the whole could be assured. Sociotechnical systems have, however, since the 1980s become steadily more complex due to rampant technological and societal developments. The scope of task analysis must therefore be extended in several directions. A "vertical" extension is needed to cover the entire system, from technology to organization. A "horizontal" extension is needed to increase the scope to include both design and maintenance. A second horizontal extension is needed to include both upstream and downstream processes. The latter in particular means that previously separate functions no longer can be treated as separate. There are important dependencies to what went before (upstream) and what comes after (downstream).

Today's task analysis must therefore address systems that are larger and more complex than the systems of yesteryear. Because there are many more details to consider, some modes of operation may be incompletely known, there are tight couplings among functions, and systems may change faster than they can be described,

the net result is that many systems today are underspecified or intractable. For these systems it is clearly not possible to prescribe tasks and actions in every detail. This means that performance must be variable or flexible rather than rigid. In fact, the less completely the system is described, the more performance variability is needed.

It is useful to make a distinction between tractable and intractable systems (Hollnagel, 2010b). Tractable systems can be completely described or specified, while intractable systems cannot. The differences between the two types of systems are summarized in Table 1.

Most established safety methods have been developed on the assumption that systems are tractable. As this assumption is no longer universally valid, it is necessary to develop methods to deal with intractable systems and irregular work environments. One way of doing that is to focus on which functions are required to achieve a goal and how they are organized relative to the current situation (e.g., existing resources and demands). This can be seen as a natural continuation of the development that has taken us from sequential task analysis via hierarchical task analysis to functional dependency and goals–means analysis. Doing so will have a major impact not only on how work situations are studied and analyzed but also on how the efficiency and safety of work can be ensured.

## REFERENCES

Annett, J., and Duncan, K. D. (1967), "Task Analysis and Training Design," *Occupational Psychology*, Vol. 41, pp. 211–221.

Annett, J., Duncan, K. D., Stammers, R. B., and Gray, M. J. (1971), "Task Analysis," Training Information Paper 6, Her Majesty's Stationery Office, London.

Card, S., Moran, T., and Newell, A. (1983), *The Psychology of Human–Computer Interaction*, Lawrence Erlbaum Associates, Mahwah, NJ.

Daniellou, F. (2005), "The French-Speaking Ergonomists' Approach to Work Activity: Cross-Influences of Field Intervention and Conceptual Models," *Theoretical Issues in Ergonomics Science*, Vol. 6, No. 5, pp. 405–427.

Diaper, D., and Stanton, N., Eds. (2003), *The Handbook of Task Analysis for Human–Computer Interaction*, Lawrence Erlbaum Associates, Mahwah, NJ.

Donders, F. C. (1969), "On the Speed of Mental Processes," *Acta Psychologica*, Vol. 30, pp. 412–431; translated from "Over de snelheid van psychische processen: Onderzoekingen gedaan in het Physiologisch Laboratorium der Utrechtsche Hoogeschool, 1868–1869," *Tweede Reeks*, Vol. II, pp. 92–120.

Gibson, J. J. (1979), *The Ecological Approach to Visual Perception*, Lawrence Erlbaum Associates, Mahwah, NJ.

Gilbreth, F. B. (1911), *Motion Study*, Van Nostrand, Princeton, NJ.

Hollnagel, E., Ed. (2003), *Handbook of Cognitive Task Design*, Lawrence Erlbaum Associates, Mahwah, NJ.

Hollnagel, E. (2010a). *The ETTO Principle: Efficiency-Thoroughness Trade-off. Why Things That Go Right Sometimes Go Wrong*, Ashgate, Farnham, UK.

Hollnagel, E., Ed. (2010b), *Safer Complex Industrial Environments*, CRC Press, Boca Raton, FL.

Hollnagel, E., and Woods, D. D. (1983), "Cognitive Systems Engineering: New Wine in New Bottles," *International Journal of Man–Machine Studies*, Vol. 18, pp. 583–600.

Hollnagel, E., and Woods, D. D. (2005), *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering*, CRC Press, Boca Raton, FL.

Hutchins, E. (1995), *Cognition in the Wild*, MIT Press, Cambridge, MA.

Jastrzebowski, W. (1857), "Rys ergonomiji czyli Nauki o Pracy, opartej naprawdach poczerpnietych z Nauki Przyrody" [An Outline of Ergonomics or the Science of Work Based on the Truths Drawn from the Science of Nature], *Przyoda i Przemysl*, Vol. 29, pp. 227–231.

Kirwan, B., and Ainsworth, L. K., Eds. (1992), *A Guide to Task Analysis*, Taylor & Francis, London.

Leplat, J. (1991), "Organization of Activity in Collective Tasks," in *Distributed Decision Making: Cognitive Models for Cooperative Work*, J. Rasmussen, B. Brehmer, and J. Leplat, Eds., Wiley, Chichester, West Sussex, England.

Lind, M., and Larsen, M. N. (1995), "Planning Support and the Intentionality of Dynamic Environments," in *Expertise and Technology: Cognition and Human–Computer Interaction*, J. M. Hoc, P. C. Cacciabue, and E. Hollnagel, Eds., Lawrence Erlbaum Associates, Mahwah, NJ.

Marsden, P., and Hollnagel, E. (1996), "Human Interaction with Technology: The Accidental User," *Acta Psychologica*, Vol. 91, pp. 345–358.

McGregor, D. (1960), *The Human Side of Enterprise*, McGraw-Hill, New York.

Miller, G. A., Galanter, E., and Pribram, K. H. (1960), *Plans and the Structure of Behavior*, Holt, Rinehart & Winston, New York.

Miller, R. B. (1953), "A Method for Man–Machine Task Analysis," Technical Report 53–137, Wright Air Force Development Center, Dayton, OH.

Newell, A., and Simon, H. A. (1961), "GPS: A Program That Simulates Human Problem-Solving," in *Proceedings of a Conference on Learning Automata*, Technische Hochschule, Karlsruhe, Germany, April 11–14.

Nisbett, R. E., and Wilson, T. D. (1977), "Telling More Than We Can Know: Verbal Reports on Processes," *Psychological Review*, Vol. 74, pp. 231–259.

Rouse, W. B. (1981), "Human–Computer Interaction in the Control of Dynamic Systems," *ACM Computing Survey*, Vol. 13, No. 1, pp. 71–99.

Schraagen, J. M., Chipman, S. F., and Shalin, V. L., Eds. (2000), *Cognitive Task Analysis*, Lawrence Erlbaum Associates, Mahwah, NJ.

Shepherd, A. (1998), "HTA as a Framework for Task Analysis," *Ergonomics*, Vol. 41, No. 11, pp. 1537–1552.

Simon, H. A. (1972), *The Sciences of the Artificial*, MIT Press, Cambridge, MA.

Taylor, F. W. (1911), *The Principles of Scientific Management*, Harper, New York.

# CHAPTER **14**

# TASK DESIGN AND MOTIVATION

**Holger Luczak**
**Research Institute for Rationalization and Operations Management**
**Aachen, Germany**


**Tanja Kabel and Torsten Licht**
**RWTH Aachen University**
**Aachen, Germany**

## 1  MEANING AND IMPACT OF WORK

Work has been and still is an object of study in many scientific disciplines. Work science, pedagogy, jurisprudence, industrial engineering, and psychology—to name just a few—have all made significant contributions to this field. Therefore, today there is a wide variety of perspectives on work and the corresponding perceptions of human beings.

As an introduction, perhaps a look into another philosophical tradition is necessary—a tradition that is not likely to be accused of offering merely a one-sided and reduced view (Luczak and Rohmert, 1985). The authors of the *Encyclica Laborem Exercens* (Pope John Paul II, 1981) regard work as a positive human good because through work people not only reshape nature to fit their needs but also fulfill themselves in a spiritual sense. They become more human and thus realize creation's divine mandate. According to Luczak and Rohmert, the uniqueness of this definition lies in its ability to stand for itself in every work-related scientific discipline.

The term *work* has always been associated with aspects of burden as well as with those of pride. In history, priority was once given to the first aspect, at other times to the latter (Schmale, 1983). In ancient times work was avoided by people who could afford

it; in contrast, Christianity looked upon work as a task intended by God and elevated successful working within the scope of the *Protestant work ethic* to the standard of salvation, a perception that has often been made responsible for the development of the great advances made during the Industrial Revolution (Weber, 1904/1905).

In connection with the population's attitude toward work, a shift from material to postmaterial values (Inglehart, 1977, 1989) appeared. This silent revolution consists of a slow change from industrial society's appreciation of safety and security to postmaterial society's emphasis on personal liberty. According to Inglehart, this trend is explained largely by development of the welfare state and improvements in education. Inglehart's theory of the silent revolution played an important role in shaping the perception of changing values. Not only the value that has been assigned to work during the various historical eras but also the concept of work itself is different, depending on the ideas of society and humankind as a whole (Hoyos, 1974; Schmale, 1983; Frei and Udris, 1990).

Waged work fulfills, in addition to the assurance of income, a series of psychosocial functions. Research, especially the work on effects of unemployment, indicates the high mental and social benefits of work.

The most important functions are as follows (Jahoda, 1983; Warr, 1984):

1.  *Activity and Competence*. The activity resulting from work is an important precondition for the development of skills. While accomplishing work tasks one acquires skills and knowledge and at the same time the cognition of skills and knowledge (i.e., a sense of competency).

2.  *Structure*. Work structures the daily, weekly, and yearly cycle as well as whole life planning. This is reflected by the fact that many terms referring to time, such as *leisure, vacation*, and *pension*, are definable only in relation to work.

3.  *Cooperation and Contact*. Most professional tasks can be executed only in collaboration with others. This forms an important basis for the development of cooperative skills and creates an essential social field of contact.

4.  *Social Appreciation*. One's own efforts, as well as cooperation with others, lead to social appreciation that in turn produces the feeling of making a useful contribution to society.

5.  *Identity*. The professional role and the task, as well as the experience of possessing the necessary skills and knowledge to master a certain job, serve as a fundamental basis for the development of identity and self-esteem.

How important these functions are is often observed when people lose their jobs or have not yet had the chance to gain working experience. But also in the definition of work by employees themselves, these functions become evident. Despite some contrary claims, work still takes a central position in the lives of many people (Ruiz Quintanilla, 1984). At the same time, distinctions can be observed. Ranking work first does not remain unquestioned any more: Values have become more pluralistic, life concepts more flexible. The number of persons that are not work oriented is increasing; this shows a flexible attitude and is especially the case for younger people (Udris, 1979a,b). This should not be misjudged as a devaluation of work as a sphere of life and least of all as a disappearance of the Protestant work ethic. Young people today differ in several respects from the generation of their parents, which is clearly motivated by the values of an industrial society. Young people have been shaped by the global communication society in which they now live. There is a shift away from values such as obedience or pure fulfillment of duty to values that favor the assertion of needs for self-development and fulfillment (Klages, 1983).

The aspects of work that people consider as most important can be summarized with the help of five keywords. Such listings (e.g., Kaufmann et al., 1982; Hacker, 1986) vary in detail and their degree of differentiation. Essentially, however, they are very much alike:

1.  *Content of work*: completeness of tasks, diversity, interesting tasks, possibility to employ one's knowledge and skills, possibility to learn something new, possibility to take decisions

2.  *Working conditions*: time (duration and position), stress factors (noise, heat, etc.); adequacy of furniture, tools, and spatial circumstances, demanding working speed

3.  *Organizational environment*: job security, promotion prospects, possibilities of further education, information management of the organization

4.  *Social conditions*: opportunities for contact, relations to co-workers and superiors, working atmosphere

5.  *Financial conditions*: wage, social benefits

However, opinions about the weighting of these aspects vary more fundamentally. Schools of work and organizational psychology often differ in the significance they attribute to the various characteristics (Neuberger, 1989). Taylor (1911) gives priority to the economic motive, whereas the human relations movement emphasizes the social aspect (Greif, 1983). Hackman and Oldham (1980) particularly stress the content of work; representative of the sociotechnical system approach (Emery, 1972; Ulich, 1991), the integration of social and technical aspects plays an essential role.

The notion that only the financial aspect is important to workers is widespread. Asking workers themselves, a more differentiated pattern results (Ruiz Quintanilla, 1984; MOW, 1987). When asked about the meaning of various aspects of work in general (keeps me busy, facilitates contacts, is interesting, gives me an income, gives me prestige and status, allows me to serve society), income ranks first, followed by possibilities for contact. Few think that the work itself is interesting and satisfactory. An evaluation of the roles that the tasks, the company, the product, the people, the career, and the money play in one's working life shows that money ranks first again, this time coequally followed by tasks and social contacts. Finally, when asked which of 11 aspects (possibilities to learn, working time, variety, interesting task, job security, remuneration, physical circumstances, and others) is most important to employees, the interesting task has top priority. How should these—at first sight, contrary results—be interpreted?

1.  The great importance that is attached to payment in the first two questions shows that ensuring one's living is seen as a fundamental function of work. It is not surprising that answering the first question, very few people judge work itself as being interesting if one considers that it is about work in general. Work has to fulfill certain conditions to be interesting; it does not obtain this quality automatically.

2.  The second question refers directly to the meaning that the various aspects have in one's own working life. Here, the tasks that one fulfills (i.e., the aspect of content) gain considerable significance.

3.  The third question is directed toward expectations. Here, payment plays an important role, too, but it can now be found on the same level

with other aspects; all are exceeded by one attribute: the desire for an interesting task.

Therefore, remuneration as a fundamental function in terms of income maintenance is of overriding importance for waged work. When talking about one's own working life, it remains the most important aspect, followed by the content of work and the social conditions. Beyond the fundamental function of income maintenance, an exceptionally high remuneration does not have priority; in this context, the aspect of an interesting task ranks first.

## 2 MOTIVATION TO WORK

Science and practice are equally interested in finding out which forces are motivating people to invest energy in a task or a job, in taking up a job at all, in being at the workplace every day, or in working with initiative and interest on the completion of a task. The understanding of work motivation makes it possible to explain why people direct their forces and energy in a certain direction, pursue a set goal, and show certain patterns of behavior and reactions in the job environment of an organization (Heckhausen, 1980; Phillips and Lord, 1980; Wiswede, 1980; McClelland, 1985; Weiner, 1985, 1992; Staw et al., 1986; Katzell and Thompson, 1990; Nicholson et al., 2004).

These considerations about motivation processes in the job environment are predominated by the assumption that behavior and work performance within an organization are influenced and determined by these motivational processes. Although one cannot doubt this observation, it must be noted that motivation cannot be the only determinant of working performance and behavior. Other variables affect the working process of the individual organizational member, too, and a motivation theory also has to take into account such variables as efforts, abilities, expectations, values, and former experiences, to name only a few, if it wants to explain working behavior.

While considering how working behavior is initiated, sustained, stopped, and directed, on which energies it is based and which subjective reactions it can trigger in the organism, most motivation researchers fall back on the two psychological concepts of needs and goals. In doing so, needs are seen as suppliers of energy and trigger the mechanism of the behavioral pattern of a working person. Therefore, a lack of necessities felt by a person at a particular time activates a search process with the intention of eliminating this deficit. Moreover, many theorists assume that the motivation process is goal oriented. Thus, the goal or final result that an employee is striving for in the working process possesses a certain attraction for the person. As soon as the goal is achieved, the lack of necessities is reduced. Thereby, the size of the deficiency is influenced by the individual characteristics of the organizational member (perceptions, personality, attitudes, experiences, values) as well as by a variety of organizational variables (structure, level, span of control, workplace, technology,

leadership, team). The individual characteristics and organizational variables influence the searching behavior shown by an employee trying to attain a goal and determine the amount of energy invested by a person to achieve a goal. The employee will deliver a certain performance that in turn will lead to the expected financial and nonfinancial rewards, and finally, to job satisfaction. It is true that motivation and satisfaction are closely linked, but they are not synonymous. Whereas motivation is to be understood as a predisposition to a specific, goal-oriented way of acting in the working process, job satisfaction is to be seen as a consequence of performance-based rewards. Therefore, the employee can be unsatisfied with the present relation between working behavior, performance, and rewards but might still be highly motivated to fulfill a task, showing initiative, good performance, and working extra hours. In other words, a motivated employee does not have to be satisfied with the various aspects of his or her work.

To a large extent, productivity of organizations depends on the willingness of its employees to use their qualifications in a goal- and task-oriented way. This motivation characterizes an important factor of human productivity in organizations and becomes apparent as quantitative or qualitative work performance as well as in low rates of absenteeism and personnel turnover. As soon as the goals of the organization correspond to those of the employees, the individual motivational states fit into the organizational frame and thus facilitate employee satisfaction. Job and organizational design strive for the creation of conditions under which people can work productively and be satisfied at the same time.

## 3 THEORIES OF WORK MOTIVATION

To explain motivated behavior in the work situation as well as the relationship between behavior and outcome or performance, a series of alternative motivation theories have been developed; some of them are described below. These theories are subdivided into two groups: content theories and process models (Campbell and Pritchard, 1976). The motivation theories of the first category concentrate on the description of the factors motivating people to work. They analyze, among other things, the needs and rewards that drive behavior. In contrast, the process models of work motivation deal primarily with the processes determining execution or omission as well as with the type of execution of an action.

### 3.1 Content Theories

#### 3.1.1 Maslow's Hierarchy of Needs

As a result of psychological experiments and of his own observations, Maslow (1954) formulated a theory that was meant to explain the structure and dynamics of motivation of healthy human beings. In doing so, he distinguished five different levels of needs (Figure 1):

1. *Physiological Needs*. These serve to maintain bodily functions (e.g., thirst, hunger, sexuality,
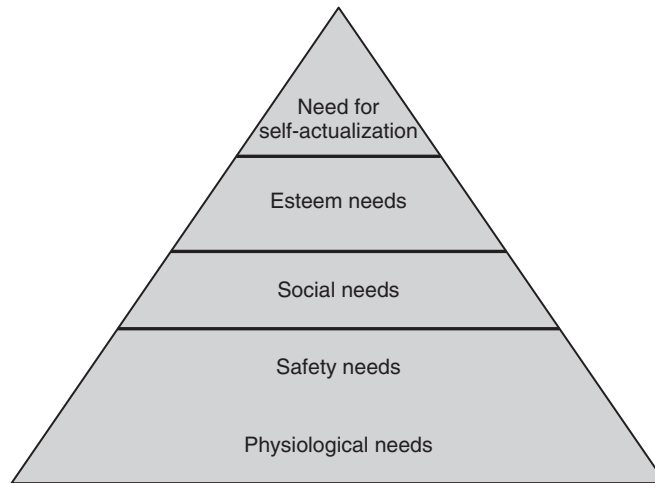
**Figure 1**  Maslow's levels of needs.

tranquility); they manifest themselves as physical deficiencies and are therefore easy to detect.

2. *Safety Needs*. These appear as the desire for safety and constancy, stability, shelter, law, and order; in industrial nations they emerge in their original form only in disaster situations; however, in a culture-specific form they are omnipresent here as well: as a need for a secure job, a savings account, or several kinds of insurances; as resistance to change; and as a tendency to take on a philosophy of life that allows orientation.

3. *Social Needs*. These needs, such as the desire for affection and belonging, aim at the give and take of sympathy and admission to society.

4. *Esteem Needs*. The satisfaction of these needs results in self-confidence and recognition; their frustration leads to feelings of inferiority and helplessness.

5. *Need for Self-Actualization*. This is the desire of human beings to realize their potential abilities and skills.

According to Maslow, these needs are integrated into a hierarchical structure, with physiological needs as the lowest level and the need for self-actualization as the highest. Maslow combines this hierarchy with the thesis that the elementary needs will take effect first; the contents of the higher levels will become important only when the needs of lower levels are satisfied to a certain extent. Only when need levels 1–4 are satisfied does the need for self-actualization get into focus. The single stages show repletion points (i.e., with adequate satisfaction of needs, motivation by means of this need is no longer possible). Striving for self-actualization is the only motive without satisfaction limits and thus remains effective indefinitely; in contrast to deficiency needs 1–4, it is a need for growth serving the perfection of human personality.

The importance of Maslow's concept has to be seen primarily in his verbalization of self-actualization as a human objective; he thus provoked an ongoing discussion (also in industrial companies). The weak spots of this theory lie especially in the difficulties of operationalization and verification and in the fact that the central concept of self-actualization is kept surprisingly vague. Nevertheless, this model is still very popular among practitioners; in the field of research, a critical reception dominates. The practical effectiveness of Maslow's approach is linked more to its plausibility than to its stringency. Still, it has initiated a number of other concepts or has at least influenced them (e.g., Barnes, 1960; McGregor, 1960; Alderfer, 1969).

The extensive criticism of this model (Neuberger, 1974) shows that especially the fascinating claim for universality cannot be confirmed. Apart from the fact that the categories of needs often cannot be distinguished sufficiently, the criticism alludes primarily to the hierarchical order of the needs and the dynamics of their satisfaction. Summarizing some research projects evaluating the validity of Maslow's model (e.g., Salancik and Pfeffer, 1977; Staw et al., 1986), there is little support for the existence of a need hierarchy:

- Most people differ very clearly regarding the degree to which they want a lower need to be satisfied before they concentrate on the satisfaction of a higher need.
- Several categories of needs overlap, and thus an individual need can fall into various categories at the same time.
- Within certain limits, working people are able to find substitutes for the satisfaction of some needs.
- The opportunities and chances given to an employee in the world of work are of great importance for the striving and intention to satisfy certain needs.

Finally, a number of researchers substantiated the fact that the types of needs that working persons are trying to satisfy within their organizations depend on the occupational group they belong to and their values, goals, and standards as well as on the options for need satisfaction offered within their occupational group. These studies have shown that unskilled workers who are offered few possibilities for autonomous work and promotion within an organization stress job security and physical working conditions a lot more than do members of other professional categories. In comparison, skilled workers emphasize the type of work that satisfies or dissatisfies them. Employees in service companies usually focus on the satisfaction of social needs and, as a consequence, on the job satisfaction derived from their social interactions with colleagues and customers.

Engineers concentrate more on the performance needs at the workplace, whereas accountants in the same companies were more concerned about their promotion, even if the promotion did not result in a financial or other material benefit (e.g., Herzberg et al., 1959). Differences of this type in the pursuit of need satisfaction can be explained partially by the fact that the possibilities for the accountant to be creative at work and to develop self-initiative are a lot more restricted than those of an engineer. A study by Porter (1964) shows that the degree of job satisfaction among executives compared to other professions in the same organization was above average, but executives of lower ranks within the organizational hierarchy were a lot less satisfied with their opportunities to work independently, autonomously, and creatively than were employees of higher ranks in the same organizational hierarchy.

### 3.1.2 Alderfer's ERG Theory

Since numerous analyses demonstrate that an excessive differentiation of needs is difficult to operationalize and that their hierarchical structure can be falsified easily, Alderfer (1972) is of the opinion that Maslow's theory is not fully applicable to employees in organizations. In his ERG theory, he reduced the number of need categories—according to Alderfer, Maslow's model shows overlaps among the safety, social, and appreciation needs—and developed a well-elaborated system of relations between these needs. This contains only three levels of basic needs (Figure 2):

1. *Existence Needs* (*E Needs*). These needs comprise the desire for physiological and material well-being and include Maslow's physiological needs, financial and nonfinancial rewards and remuneration, and working conditions.
2. *Relatedness Needs* (*R Needs*). These needs can be subsumed as desire for satisfying interpersonal relationships and contain Maslow's social needs as well as the esteem needs.
3. *Growth Needs* (*G Needs*). These needs can be described as a desire for continued personal growth and development and as the pursuit of self-realization and productivity; therefore, this category forms an overlap of Maslow's esteem needs and self-actualization.

Maslow defines the motivation process of humans, who are always aspiring for the next-higher need level, as a type of progression by satisfaction and fulfillment of the particular needs: The person has to satisfy a specific need first, before the next higher one is activated. Alderfer includes a component of frustration and regression in his model. Thereby, Alderfer's theoretical assumptions are opposed to Maslow's model in several fundamental aspects: ERG theory does not claim that lower level needs have to be satisfied as a precondition for the effectiveness of higher level needs. Moreover, the need hierarchy now works in the opposite direction as well (i.e., if the satisfaction of higher level needs is blocked, the underlying need is reactivated). ERG theory acknowledges that if a higher level need remains unfulfilled, the person might regress to lower level needs that appear easier to satisfy. According to Alderfer's hypothesis of frustration, the power of a need is increased by its frustration, but there is no mandatory association between the needs of the various categories. Therefore, needs already satisfied also serve as motivators as long as they are a substitute for still unsatisfied needs. Another difference is that, unlike Maslow's theory, ERG theory contends that more than one need may be activated at the same time. Finally, ERG theory allows the order of the needs to be different for different people.

The two concepts of fulfillment progression and frustration regression are both constituents of the dynamics
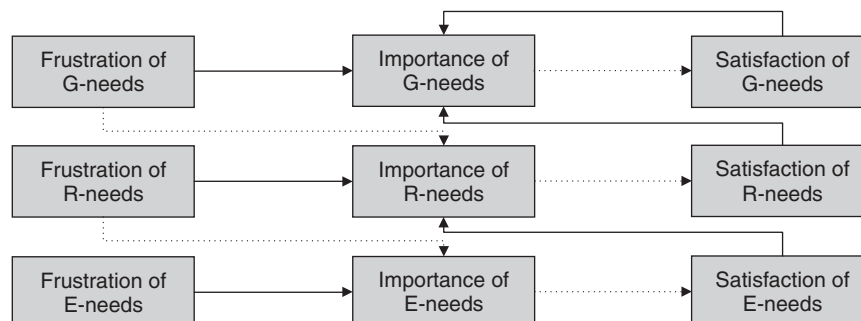


**Figure 2** Alderfer's ERG theory.

of ERG theory. As a consequence of the diverging assumptions of Maslow and Alderfer, different explanations and prognoses about the behavior of employees at their workplace are possible (e.g., Schneider and Alderfer, 1973; Guest, 1984). So far, a completely satisfying proof, especially of the psychological progression and regression processes, has not been successful. On the other hand, the relatively primitive classification of the three needs shows it to be surprisingly acceptable and able to discriminate in several international studies (Elizur et al., 1991; Borg et al., 1993). According to the authors' results, international comparison demonstrates that in case of a lack of financial means and poverty of employees in a certain region the phenomena that are attributed to the complex "existence" come to the fore, whereas in saturated societies especially, growth needs are dominant.

### 3.1.3  McGregor's X and Y Theory

McGregor (1960) supported a direct transfer of Maslow's theory to job motivation. He objected to a theory X that was derived from managerial practice. It starts from the following assumptions: The tasks of management concerning personnel consist in the steering of its performance and motivation as well as in the control and enforcement of company goals. Since without these activities employees face a company's goals in a passive way or resist them, it is necessary to reward, punish, and control. Therefore, a principally negative view of employees prevails among managers. In detail, this view is determined by the wrong ideas that the average person is lazy and inactive, lacks ambition, dislikes responsibility, is egocentric by nature and indifferent to the organization, is greedy and money

oriented, objects to changes from the outset, and is credulous and not very clever.

According to McGregor, this concept of management is destructive for employees' motivation. Hence, McGregor drafts an antithesis building on Maslow's need hierarchy, naming it theory Y. Essentially, it contains the following assumptions: (1) observable idleness, unreliability, dislike of responsibility, and material orientation are consequences of the traditional treatment of the working person by management and (2) motivation in terms of potential for development, the willingness to adapt to organizational goals, and the option to assume responsibility exists in every person, and it is the fundamental task of management to create organizational conditions and to point out ways that allow employees to reach their own goals best when bringing them into agreement with the company's goals.

McGregor proposes the following measures that can ease the restrictions of the possibilities of satisfaction for the employees and facilitate responsible employment for the purpose of Maslow's ideal conception: (1) decentralization of responsibility at the workplace, (2) participation and a consulting management, and (3) involvement of employees in control and evaluation of their own work.

### 3.1.4  Herzberg's Two-Factor Theory

The motivation theory developed by Herzberg et al. (1959) is probably the most popular theory of work motivation (Figure 3). Its central topic is job satisfaction. Results of empirical studies led Herzberg and his colleagues to the opinion that satisfaction and dissatisfaction at the workplace are influenced by various groups of factors. Dissatisfaction does not occur simply
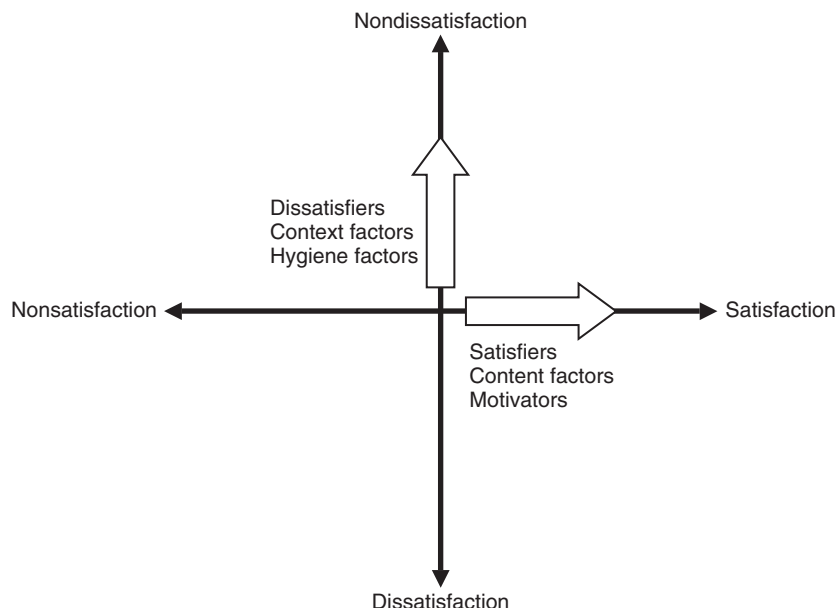


**Figure 3**  Herzberg's two-factor theory.

because of the lack or insufficient value of parameters that otherwise cause satisfaction. Herzberg called the factors that lead to satisfaction *satisfiers*. These are primarily (1) the task itself, (2) the possibility to achieve something, (3) the opportunity to develop oneself, (4) responsibility at work, (5) promotion possibilities, and (6) recognition.

Since these factors are linked directly to the content of work, Herzberg also referred to them as *content factors*. Due to the fact that their positive value leads to satisfaction and consequently motivates for performance, they were seen as the actual motivators. According to Herzberg, for the majority of employees, motivators serve the purpose of developing their professional occupation as a source of personal growth.

In contrast, *dissatisfiers* have to be assigned to the working environment and are therefore also called *context factors*. According to Herzberg, they include especially (1) the design of the surrounding working conditions, (2) the relationship with colleagues, (3) the relationship with superiors, (4) company policy and administration, (5) remuneration (including social benefits), and (6) job security. Since the positive values of these parameters accommodate the employees' need to avoid unpleasant situations in a preventive way, they were also referred to as *hygiene factors*. In many cases content factors allude to intrinsic motivation and context factors to extrinsic motivation.

Since Herzberg's theory also suggests a number of practical solutions to organizational problems and allows predictions about behavior at the workplace to a certain extent and because of the impressive simplicity of the model and its orientation toward the terminology of organizational processes, a large variety of empirical studies have been undertaken to examine the underlying postulates and assumptions (e.g., Lawler, 1973; Kerr et al., 1974; Caston and Braito, 1985). However, many of these studies raised additional questions. The essential objections that can be put forward against Herzberg's model are:

- The restricted validity of data, being based on a small number of occupational groups (only engineers and accountants)
- The oversimplification of Herzberg's construct of motivation or job satisfaction (e.g., satisfaction and dissatisfaction could be based on the working context as well as on the task itself or on both equally)
- The division of satisfaction and dissatisfaction into two separate dimensions
- Lack of consideration of unconscious factors that can have an effect on motivation and dissatisfaction
- Lack of an explanation of why different extrinsic and intrinsic work factors are to influence the performance in a negative or positive way and why various work factors are important
- Lack of consideration of situational variables
- No measurement of job satisfaction as a whole (it is very possible that somebody dislikes parts of his or her job but still thinks that the work is acceptable as a whole)

Existing studies about Herzberg's theory have left many problems unsolved, and it is doubtful whether a relatively simple theory such as Herzberg's can ever shed light on all the questions that are raised here. Despite this criticism, the theory, although being an explanation for job satisfaction rather than a motivation theory, is still very popular. Despite restrictions regarding the methodology and content of this theory, it can be stated that motivators offer a greater motivational potential to intrinsically motivated employees than extrinsic incentives do (e.g., Zink, 1979). Especially when described in a shortened way, Herzberg's concept obviously has such a high plausibility that in industry it still has an astonishing repercussion. The importance of Herzberg's approach is to be seen primarily in the fact that he set the content of work as the focus of attention. This gave numerous companies food for thought and induced manifold change processes. Last but not least, the emphasis on work content had an effect on the dissemination of the so-called *new forms of work design* (Miner, 1980; Ulich, 1991).

### 3.1.5 Hackman and Oldham's Job Characteristics Model

Behavior and attitudes of employees and managers can be influenced to a great extent by a multitude of context variables. Moreover, during recent years the connection and adequate fit of task characteristics or work environment, on the one hand, and the psychological characteristics of the person, on the other (person–job fit), have been studied very intensively and from many different perspectives. Task and work design, especially, formed the focus of interest. These studies were initiated particularly by a motivation model of job and task characteristics (the job characteristics model of work motivation) (Figure 4) developed by Hackman and Oldham (1976, 1980). The model postulates that certain core dimensions of work lead to certain psychological states of the working person which result in specific organizational or personal outcomes. Hackman and Oldham list five job characteristics that cause enhancement of motivation and a higher degree of performance and job satisfaction. Furthermore, they propose that persons with a strong psychological growth need react in a more positive way to tasks containing many core dimensions than do people with a weak growth need. Here, direct reference to Maslow's need hierarchy becomes obvious. A person who currently prioritizes the need for self-actualization is categorized as strong with regard to his or her growth need, whereas somebody operating on the level of safety needs would be seen as somebody with a weak growth need.

The core dimensions of work are the following:

- *Skill variety*: degree to which tasks require different skills or abilities
- *Task identity*: degree to which a person completes a connected piece of work or a task instead of parts or facets of it

*Core job dimensions*          *Critical psychological states*          *Outcomes*



$$\text{Motivation potential} \; = \; \frac{\text{skill variety} + \text{task identity} + \text{task significance}}{3} \; \times \text{autonomy} \times \text{feedback}$$
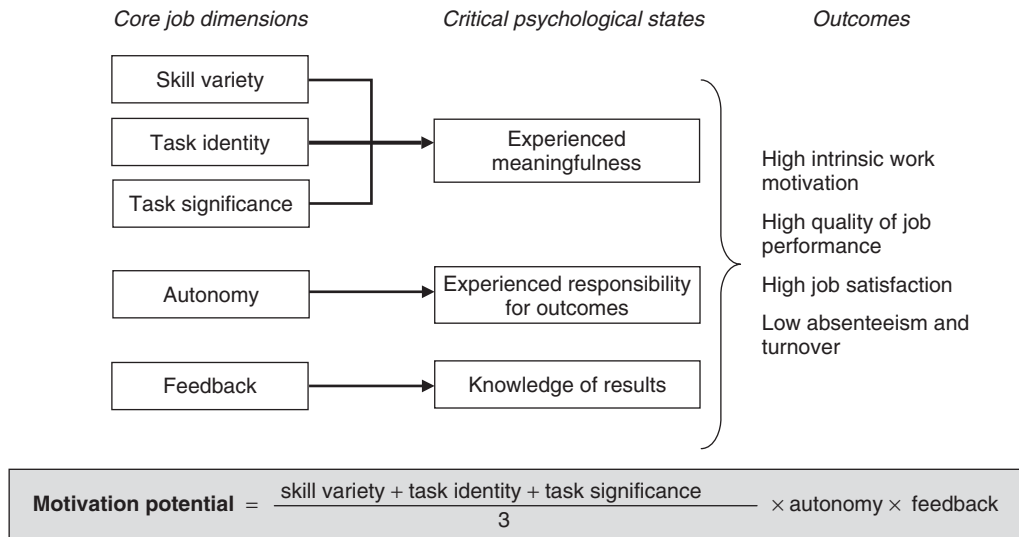
**Figure 4**  Job characteristics model.

- *Task significance*: degree to which work has an effect on the lives and jobs of others
- *Autonomy*: freedom and independence in the accomplishment of work
- *Feedback*: degree to which work provides clear and direct information about success and effectiveness of the performing person

Summarizing these core dimensions, they result in an index of motivational potential of a job by which jobs can be assessed regarding their possibilities to motivate. But this also makes it possible to compare and classify entire work processes depending on whether they are able to help a person to achieve professional and intellectual growth as well as further development. The score can be seen as a measure of adequacy, quality, and success of the present work design. The formula postulates an additive relation for skill variety, task identity, and task significance whereby the single dimensions can compensate for each other. The multiplicative connection between autonomy and feedback does not allow this and can in an extreme case reduce the motivational potential score to zero. Although other authors affirm the strong emphasis on feedback only with reservations, the formula sharply covers important effects on motivation. In many cases an additive combination of the variables shows it to be on a par with the postulated multiplicative formula.

In the center of the model are the *critical psychological states*. These are defined as follows:

- *Meaningfulness of work*: degree to which a person experiences work in general as meaningful, valuable, and worthwhile
- *Responsibility for outcomes*: degree to which a person feels personally responsible for the work that he or she is doing

- *Knowledge of results*: degree to which a person is continually informed about how successful and effective the job done is

When these states are on an acceptable level, it can be presumed that the person feels good and reacts in a positive way toward the job. It is expected that the dimensions skill variety, task identity, and task significance influence the meaningfulness of work experienced by a person. The dimension autonomy presumably affects the experienced responsibility for outcomes. Feedback contributes to knowledge of the actual results. The critical psychological states for their part determine a variety of personal and work outcomes. Several studies show correlations with intrinsic motivation, job satisfaction, absenteeism, and turnover but only low correlations with quality of job performance. Finally, it is assumed that the growth need strength of an employee mediates relations among the other elements of the theory.

To test the theory empirically, Hackman and Oldham (1975) developed the Job Diagnostic Survey (JDS). This instrument allows objective measurement of job dimensions and measures the various psychological states caused by these characteristics, the affective responses to these job characteristics, and the strength of the personal need to grow and to develop oneself. The JDS can be used to identify workplaces or work structures with a high or low motivational potential. Since the motivational potential score provides a comprising index for the overall motivational potential of a job, a low score points to jobs that deserve redesign. For implementation of the theory in practice, a set of action principles has been developed that give instructions about how the core dimensions of work can be improved (Hackman and Suttle, 1977). In practice, the model has reached a comparably high degree of popularity and is used for the improvement of tasks and job design as often as

Herzberg's model. Therefore, scientific literature concerning this model is very extensive (e.g., Fried and Ferris, 1987; Fried, 1991).

Although research has so far supported the model on the whole, the model contains a number of unsolved problems and methodological weaknesses on which further research will have to focus (Roberts and Glick, 1981). There are questions about the tools that measure the various components of the model, about the significance and modality of computation of the motivational potential score, and about the theoretical foundation on which the model is based. It has to be asked whether the five core dimensions measured with the JDS are really independent (Idaszak and Drasgow, 1987; Williams and Bunker, 1993). In addition, the assumed impact of employees' growth need strength has to be examined, especially its significance and direction. Empirical evidence has to be provided as to which way the critical psychological states are caused or influenced by the five core dimensions of work. Although the formula for the computation of motivational potential suggests considerable interactions among the different job characteristics, it is just as possible that employees tend to ascribe more weight to job characteristics that are most beneficial to them (e.g., autonomy, variety). Furthermore, the mediating function of the psychological states is still quite indistinct (Hackman and Oldham, 1980; Miner, 1980; Udris, 1981).

### 3.1.6 McClelland's Theory of Acquired Needs

McClelland's theory of acquired needs (McClelland, 1984, 1985; McClelland et al., 1989) is closely linked to the psychological concepts of learning and is based to a great extent on the works of Murray (1938). McClelland holds the view that many needs are learned by dealing with and mastering the cultural environment in which a person is living (McClelland, 1961). Since these needs are learned from early childhood on, working behavior that is rewarded will occur more often. Applied to an organization, this means that employees can be motivated, by financial and nonfinancial rewards, to be at their workplace on time and regularly as long as these rewards are linked directly to the favored working behavior. As a result of this learning process, people develop certain need configurations that influence their working behavior as well as their job performance.

Together with other researchers (e.g., Atkinson and Feather, 1966), McClelland filtered those needs out of Murray's list of human needs that in his opinion represent the three key needs in human life: (1) the need for achievement (n-ach), (2) the need for affiliation (n-affil), and (3) the need for power (n-pow). These three unconscious motives have a considerable effect on both the short- and long-term behavior of a person (McClelland et al., 1989). The need for achievement is relevant for change behavior and contains continuous improvement of performance. The need for affiliation is important for group cohesion, cooperation, support, and attractiveness in groups. The need for power is of importance for persuasiveness, orientation toward contest and competition, and readiness to combat. There is an important link to the role of a manager that

consists essentially of activating these motivations and thus energize and guide the behavior of subordinates.

The main interest focuses on the need for achievement, and as a consequence, a theory of achievement motivation was formulated (Atkinson and Feather, 1966) that was applied widely in organizational psychology (Stahl, 1986). Achievement motivation corresponds to a relatively stable disposition of behavior or a potential tendency of behavior of an employee in an organization to strive for achievement and success. However, this motivation becomes effective only when being stimulated by certain situational constellations or incentives that lead a person to assume that a certain working behavior will produce the feeling of achievement. The final result is an inner feeling of satisfaction and pride of achievement. The model developed for this purpose is in effect an expectancy–valence model of motivation processes. Working behavior is understood as the resultant of (1) motivation strength, (2) the valence or attractiveness of the incentive that activates motivation, and (3) a person's expectancy that a certain behavior will lead to gaining the incentive. Hence, the corresponding motivation model can be designed as follows:

$$T_S = M_s \times P_s \times I_s$$

A person's tendency ($T_S$) to approach a task is a multiplicative function of the person's strength of the achievement motivation ($M_s$), the subjective probability of success ($P_s$), and the valence or degree of attractiveness of this success or reward ($I_s$). From this assumption, a set of conclusions can be deduced for the processes of job design not only for the selection and promotion of organizational members but also for the preference of certain leadership styles and for the motivation of risk behavior among managers in decision situations (McClelland et al., 1989).

Empirical research on McClelland's model is extensive and shows a set of consistent results. Here are some examples: People who are highly achievement motivated (have high scores on n-ach) prefer job situations in which they bear responsibility, get feedback on a regular basis, and that ask for a moderate attitude toward risks. Such a constellation has a very motivating effect on high achievers. These kinds of people are active mainly in self-dependent fields of activities. Restrictively, it has to be added, though, that high achievers often show less interest in influencing the achievement of others than in personally seeking high achievements. This is why they often are not good as managers. Reciprocally, managers of big companies rarely are n-ach people. The constellation looks a lot different when considering the need for affiliation or for power. Research shows that successful managers frequently have a high need for power and a rather low need for affiliation, a constellation that probably is necessary for efficiency of leadership and that can possibly be deduced from the function or role within the organizational context (Parker and Chusmir, 1992). Finally, Miron and McClelland (1979) point out that for the filling of positions that require a high need for achievement, a combination of selection and training is advised. People with high n-ach scores are selected and

developed by means of achievement trainings with the goal of imparting to these persons a pattern of thought in terms of achievement, success, and profit and to act upon this pattern. The three secondary or learned motivations investigated proved to be very informative and stable with regard to the explanation of working behavior and leadership. However, it is possible that the explication of organizational behavior can be improved by other secondary motives. This applies especially in the area of managerial functions. Therefore, Yukl (1990) adds two more secondary motives to his description of skills, characteristics, and goals of successful managers: the need for security and the need for status. For a better illustration, the five key motivations and their descriptors are listed below:

1. Need for achievement
   - To excel others
   - To attain challenging goals
   - To solve complex problems
   - To develop a better method to do a job
2. Need for power
   - To influence others to change their attitudes and behavior
   - To control people and things
   - To have a position of authority
   - To control information and resources
3. Need for affiliation
   - To need to be liked by others
   - To be accepted as part of a group
   - To relate to others in a harmonious way and to avoid conflicts
   - To take part in enjoyable social activities
4. Need for security
   - To have a secure job
   - To be protected against loss of income
   - To avoid tasks and decisions that include risks or failures
   - To be protected against illness and incapacity for work
5. Need for status
   - To have the right car and to wear the right clothes
   - To work for the right company in the right position
   - To have the privileges of leaders
   - To live in the right neighborhood and to belong to the right club

### 3.1.7   Argyris's Concept

An approach that joins different concepts together is that of Argyris (1964). According to Argyris, work motivation, the competency to solve problems, and emotional well-being are facilitated primarily by a feeling of self-esteem based on *psychological success*. The possibility of defining one's own goals according to one's own needs and values and to control goals in a self-dependent way operates as an important precondition for psychological success. From this emanates a contradiction to the structures of the formal organization that work such that a single employee can control his or her own working conditions to a minimal extent only, that he or she can bring in only a few or very limited skills in his or her work, and that he or she can behave only in a very dependent way.

Only if organizations believe that employees want to apply their skills in the framework of the company's goals and that they want to get involved in relevant decisions will employees be able to behave like grown-ups. On the other hand, if companies are structured differently, employees will behave accordingly: dependent, with little interest, with a short-term perspective; independence of thinking and acting might find their expression only in the development of defense.

Argyris's contribution contains a variety of unclarified points (Greif, 1983). As for Maslow and Herzberg, it is also true for Argyris that interindividual differences are largely disregarded in their concrete meaning for the development of job and organizational structures.

### 3.1.8   Deci and Ryan's Self-Determination Theory

Another metatheory of motivation and personality, the self-determination theory (SDT) of Deci and Ryan (1985), is based on the assumption that humans naturally strive for psychological growth and development. While mastering continuous challenges, the social context plays a vital role; its interaction with the active organism allows predictions about behavior, experience, and development. Deci and Ryan postulate three motivating factors that influence human development: (1) the need for autonomy, (2) the need for competence, and (3) the need for relatedness. They are referred to as basic psychological needs that are innate and universal (i.e., they apply to all people, regardless of gender, group, or culture).

The need for autonomy refers to people's striving for self-determination of goals and actions. Only when perceiving oneself as the origin of one's own actions and not being at the mercy of one's environment can one feel motivated. The successful handling of a task will be perceived as the confirmation of one's own competence only if it was solved mainly autonomously. The need for competence expresses the ambition to perceive oneself as capable of acting effectively in interaction with the environment. The need for relatedness has a direct evolutionary basis and comprises the close emotional bond with another person.

When these three needs are supported by social contexts and are able to be fulfilled by individuals, well-being is reinforced. Conversely, when cultural, contextual, or intrapsychic forces inhibit the fulfillment of the three basic needs, well-being is reduced. In this theory, motivation is seen as a continuum from amotivation to extrinsic motivation to intrinsic motivation. According to SDT, autonomy, competence, and relatedness are

three psychological nutriments that facilitate the progression from amotivation to intrinsic motivation (Ryan and Deci, 2000; Deci, 2002).

SDT contains four subtheories that have been developed to explain a set of motivational phenomena that has emerged from laboratory and field research:

- *Cognitive evaluation theory*: deals with the effects of social contexts on intrinsic motivation
- *Organismic integration theory*: helps to specify the various forms of extrinsic motivation and the contextual factors that either promote or prevent internalization
- *Causality orientations theory*: pictures individual varieties in people's tendencies toward self-determined behavior and toward directing to the environment in a mode that supports their self-determination
- *Basic needs theory*: develops the idea of basic needs and their connection to psychological health and well-being

Cognitive evaluation theory (CET), for example, aims at specifying factors that explain variability in intrinsic motivation and focuses on the need for competence and autonomy. *Intrinsic motivation* means to do an activity for the inherent satisfaction of the activity itself and thus differs from *extrinsic motivation*, referring to the performance of an activity to attain some separable outcome (Ryan and Deci, 2000). According to CET, social-contextual events (e.g., feedback, communication, rewards) conducive to feelings of competence and autonomy during action can enhance intrinsic motivation. Studies showed that intrinsic motivation is facilitated by optimal challenges, effectance-promoting feedback, and freedom from demeaning evaluations (e.g., Deci, 1975; Vallerand and Reid, 1984). However, the principles of CET do not apply to those activities that do not hold intrinsic interest and that do not have the appeal of novelty, challenge, or aesthetic value (Ryan and Deci, 2000).

Contrary to some studies, Deci does not assume an additive connection between intrinsic and extrinsic motivation. Rather, he postulates an interaction between both types of motivation, which means that extrinsic incentives can replace intrinsic motivation (Vansteenkiste and Deci, 2003). In his studies he tested the following hypotheses:

1. If intrinsic motivation makes a person perform an action and this action is recompensed with an extrinsic reward (e.g., money), his or her intrinsic motivation for the particular action decreases.

2. If intrinsic motivation makes a person perform an action and this action is recompensed with verbal encouragement and positive feedback, his or her intrinsic motivation for this particular action increases.

The design of Deci's studies has always been the same: An extrinsic incentive is added to an interesting activity. Then the variance of the intrinsic motivation has been measured on the basis of the dependent variable. He concludes that his experimental results support the hypotheses mentioned above. Money seems to have a negative impact on intrinsic motivation. Verbal encouragement and positive feedback, on the other hand, have a positive impact.

Self-determination theory has been applied within very diverse domains, such as health care, education, sports, religion, and psychotherapy, as well as in industrial work situations. For example, Deci et al. (1989) found that managers' interpersonal orientations toward supporting subordinates' self-determination versus controlling their behavior, correlated with their subordinates' perceptions, affects, and satisfactions. Moreover, the evaluation of an organizational development program focusing on the concept of supporting subordinates' self-determination showed a clearly positive impact on managers' orientations but a less conclusive impact on subordinates. Later studies on the topic of supervisory style support these findings: Participants experienced higher levels of intrinsic motivation under conditions of an autonomy-supportive style than of non-punitive controlling and punitive controlling supervisory styles (Richer and Vallerand, 1995). Researchers were also able to show that the constructs of SDT were equivalent across countries as well. Deci et al. (2001) found that a model derived from SDT in which autonomy-supportive work climates predict satisfaction of the intrinsic needs for competence, autonomy, and relatedness, which in turn predict task motivation and psychological adjustment on the job, was valid in work organizations in the United States as well as in state-owned companies in Bulgaria.

Among the three needs, autonomy is the most controversial. Iyengar and Lepper (1999) presume that cultural values for autonomy are opposed to those of relatedness and group cohesion. They provided experimental evidence showing that the imposition of choices by an experimenter relative to personal choice undermined intrinsic motivation in both Asian Americans and Anglo Americans. However, they also showed that adopting choices made by trusted others uniquely enhanced intrinsic motivation for the Asian group. Their interpretation focused on the latter findings, which they portrayed as challenging the notion that autonomy is important across cultures. Oishi (2000) measured autonomy by assessing people's individualistic values, apparently assuming them to represent autonomy as defined within SDT. On the basis of this measure, Oishi reported that outside of a very few highly individualistic Western nations, autonomous persons were not more satisfied with their lives. Finally, Miller (1997) suggested that in some cultures adherence to controlling pressures yields more satisfaction than does autonomy. Her characterizations of autonomy, like those of Iyengar and Lepper and Oishi, do not concur with SDT's definition.

### 3.1.9  Summary of Content Theories

So far it has not been possible to provide evidence that certain motives are universal. More promising seems the identification of dominant motives or constellations for

certain groups of persons (Six and Kleinbeck, 1989). French (1958) shows, for example, that praise is more effective when its content addresses the dominant motive: Praise for efficiency leads to a better performance in achievement-motivated people, praise for good cooperation, in affiliation-motivated people. Other studies support the importance of the congruence of the person's motivation structure and the organization's structure of incentives.

It is neither possible nor reasonable to fix the number of motivating factors at work once and for all; the differentiation has to differ depending on the purpose of analysis and design. It is striking, though, that from many analyses two factors of higher order result that correspond to a large extent to Herzberg's factors (content and context) (Campbell and Pritchard, 1976; Ruiz Quintanilla, 1984). All concepts mentioned have pointed out the importance of intrinsic motivation by means of holistic and stimulating work contents (Hacker, 1986; Volpert, 1987; Ulich, 1991). The complexity of the problem makes it impossible to choose a theory that is able to serve as an exclusive basis for a satisfying explanation of working behavior in organizations. All theories have in common that they try to explain the "what" of energized behavior, that they recognize that all people possess either congenital or learned and acquired needs, and finally that they reveal nothing about "how" behavior is energized and directed.

## 3.2   Process Models

These motivation theories try to answer the question of how human behavior is energized, directed, and stopped and why humans choose certain ways of behavior to reach their goals. They differ from the content theories especially by stressing cognitive aspects of human behavior and postulating that people have cognitive expectations concerning the goal or final result that is to be reached. According to these instrumentality theories, humans only decide to act if they can achieve something that is valuable for them, and thus an action becomes instrumental for the achievement of a result to which a certain value is attached.

### 3.2.1   Vroom's VIE Theory

Vroom (1964) refers to his instrumentality theory as VIE theory. The central part of this theory is constituted by the three concepts of valence (V), instrumentality (I), and expectancy (E). *Valence* describes the component comprising the attracting and repellent properties of the psychological object in the working environment—payment has a positive, danger a negative valence. Thus, before a job action is initiated, the person is interested in the value of the final result. This valence reflects the strength of the individual desire or the attractiveness of the goal or final result for the person that can be reached by different means. To be able to explain the processes in this situation of selection of alternative actions, Vroom establishes the idea of a result on a first level and a result on a second level. Thereby, the functions of the two other components, instrumentality and expectancy, are defined. An employee may assume that if he or she does a good job, he or she

will be promoted. The degree to which the employee believes in this is an estimation of subjective probability, referred to as *expectancy*. Finally, the expectancy of a person relates to his or her assumed probability that a certain effort will lead to a certain outcome. Thus, in attaining the goal of the first level, the person sees a way to reach the goal of the second level. According to this, the considerations of Vroom's motivation model can be phrased as follows: The valence of the result of the first level (e.g., effort) is determined by the person's estimation of the probability that this first-level result will lead to a row of second-level results (e.g., pay raise or promotion) and the valences linked to it.

Thus, Vroom's model (Figure 5) states that motivation or effort put in by a person to reach his or her goals are a function of his or her expectancy that as a result of his or her behavior a certain outcome will be achieved and of the valence that the result has. If one of the two factors is zero, there is no motivational force and an action for the achievement of a certain result will not take place. Hence, this motivation model provides very concrete explanations for the working behavior of employees in organizational practice. In this context it is important to emphasize that the direct explanations of the VIE model do not refer to working results or job performance but to motivated behavior—to decision making and effort in particular. As to this, the model can be seen as confirmed to a rather wide extent; expected connections to performance are not that strong because there are several mediators between effort and result that have to be taken into account. To make some good explanations with the help of this model, there has to be an appropriate organizational environment; outcomes of actions and their consequences for the persons concerned have to be transparent and calculable consistently. This is where to find important practical implications of the model for leadership and organizational design.

However, despite the advantages offered by Vroom's motivation model, there are some unsolved problems inherent in the model as well that restrict its explanatory power. For example, Vroom provides no information about the effect of those factors that influence the expectancy of an organizational member (e.g., self-esteem, former experiences in similar situations, abilities, leadership). It is also possible that employees misjudge a work situation, possibly because of their needs, emotions, values, or assumptions. This situation may result in employees choosing a nonadequate behavior and in not considering all factors that are relevant. In addition, it could be put forth against Vroom's ideas that to date there has been no research effort to determine how expectations and instrumentalities develop and by which factors they are influenced. Furthermore, the specific operation mode of the model is too complex and too rational to represent human calculations in a realistic way. Besides, an additive model works just as well as a multiplicative one. Altogether, the particular value of the model is that it points out the importance of multiple results or consequences with their probabilities and appraisal being different for each person. In this
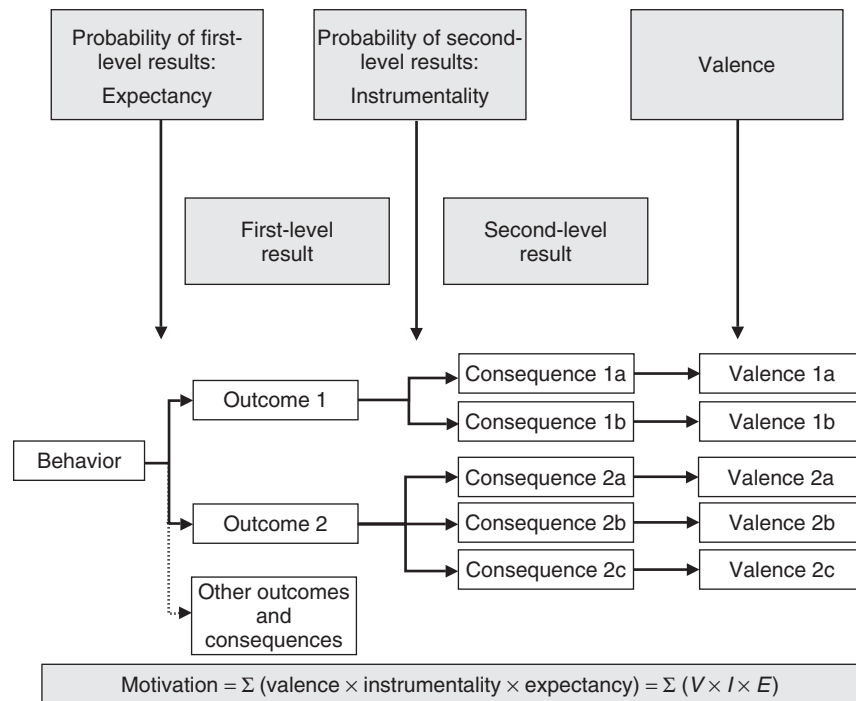
**Figure 5**  Vroom's VIE theory.

respect, it turns against perceptions that are too simple and according to which few motives virtually lead directly to actions and suggests a more complex strategy of analysis.

### 3.2.2  Porter and Lawler's Motivation Model

Among the process models, Porter and Lawler's (1968) model and in succession those of Zink (1979) and Wiswede (1980) have to be pointed out. They consider satisfaction, among other things, as a consequence of external rewards or, with intrinsic motivation, of self-reward for results of actions. The value of these models, which try to integrate various social–psychological principles (e.g., social matching processes, aspiration level, self-esteem, attribution, role perceptions, achievement motivation), is to offer a heuristically effective framework in which relevant psychological theories are related in a systematic way to explain performance and satisfaction.

Porter and Lawler's motivation model (Figure 6) is closely related to Vroom's ideas but focuses more on the special circumstances in industrial organizations. It is a *circulation model* of the relationship between job performance and job satisfaction. With this model, Porter and Lawler describe working behavior in organizations by emphasizing the rational and cognitive elements of human behavior that have been ignored, especially by the content theories. This is particularly true with regard to planning and decision making regarding anticipated future events at the workplace. The two crucial points in this model are:

- *The subjective probability* $E \rightarrow P$: the expectation to achieve a goal with greater effort
- *The probability* $P \rightarrow O$: a good performance will lead to the desired output, considering the valences of these goals

Thus, Porter and Lawler postulate that the motivation of an organizational member to do a good job is determined essentially by two probability factors: by the subjective estimated values $E \rightarrow P$ and $P \rightarrow O$. In other words the individual motivation at the workplace is determined by the probabilities that increased effort leads to better performance and that better performance leads to goals and results that have a positive valence for the person. Moreover, Porter and Lawler state that the two probabilities $E \rightarrow P$ and $P \rightarrow O$ are linked to each other in a multiplicative way. But this multiplicative relationship says as well that as soon as one of the two factors is zero the probability relation between effort and final result also decreases to zero. Explications for observable behavior at the workplace that can be derived from this are evident.

In this model the first component, subjective value of rewards, describes the valence or attractiveness that different outcomes and results of the work done have for the person—different employees have different values for different goals or results. The second component, probability between effort and rewards, refers to the subjective probability with which a person assumes that an increase in effort leads to the receipt of certain results of rewards and remuneration considered as
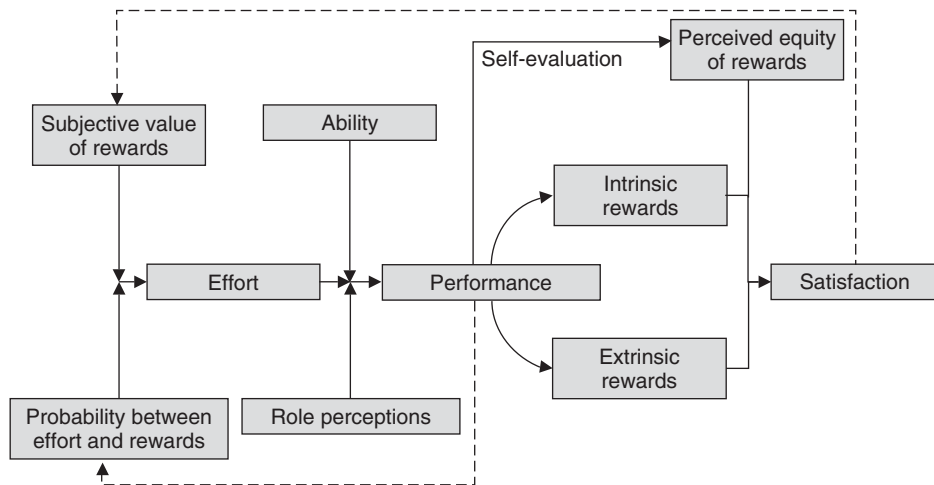
**Figure 6**  Porter and Lawler's motivation model.

valuable and useful by the person. This estimated probability contains in the broader sense the two subjective probabilities specified above: $E \rightarrow P$ and $P \rightarrow O$. The third component of the model consists of the effort of an organizational member to perform on a certain level. In this point, Porter and Lawler's motivation model differs from former theories because it distinguishes between effort (applied energy) and work actually performed (efficiency of work performance). The fourth component, the area of individual abilities and characteristics (e.g., intelligence or psychomotoric skills), has to be mentioned. It sets limits on an employee's accomplishment on a task. These individual characteristics, which are relatively stable, constitute a separate source of interindividual variation in job performance in this model. Role perceptions are based primarily on how an employee interprets success or successful accomplishment of a task at the workplace. They depend on what and in which direction a person will focus his or her efforts. In other words, role perceptions directly influence the relationship between effort and quality of job performance. This is why inadequate role perceptions lead to a situation where an employee obtains wrong or useless work results while showing great effort. Finally, the accomplishment of a job constitutes a sixth component that refers to the level of work performance an employee achieves. Although task execution and work performance play such an important role in organizations, the components themselves, as well as their interactions, are often misunderstood, oversimplified, and wrongly interpreted. Successful accomplishment of a task is, as Porter and Lawler's model shows, influenced by multiple variables and their interactions. It is the resultant of a variety of components and a combination of various parameters and their effects. The component reward consists of two parts: the intrinsic reward, given by oneself, and the extrinsic reward, essentially (but not exclusively) given by a superior. An intrinsic reward is perceived only if the person believes that he or she has mastered a difficult task. On the other

hand, an extrinsic reward can be perceived only if the successful execution of a task is noticed and valued accordingly by a superior, which is often not the case. The last two components of the model are the reward seen as appropriate by the employee and the "satisfaction" of the employee. This component, perceived equity of rewards, refers to the amount of the reward that the employee, based on performance, expects as appropriate and fair from the organization. The degree of satisfaction can be understood as the result of the employee's comparison of the reward actually obtained to the reward considered as appropriate and fair as compensation for the job done. The greater the difference between these two values, the higher will be the degree of satisfaction or dissatisfaction.

The model developed by Porter and Lawler shows in a striking way that a happy employee is not necessarily a productive one. Numerous empirical studies prove the correctness of the assumptions of the model in its essential points (Podsakoff and Williams, 1986; Locke and Latham, 1990; Thompson et al., 1993; Blau, 1993).

Finally, it has to be stated that in the underlying formula of performance, $P = f(M \times A)$ (i.e., performance is a function of the interaction of motivation and ability), another important parameter has been ignored. If we want to explain and predict work performance, it seems more realistic to consider the possibility of achieving a certain goal and showing a certain performance as well. Even if a person is willing and able to do a good job, there might be obstacles reducing or even thwarting success. This is why the formula has to be broadened to $P = f(M \times A \times O)$. A lack of possibilities or options to achieving maximum performance can be found in any work environment. They range from defective material, tools, devices, and machines to a lack of support by superiors and colleagues and inhibiting rules and processes or incomplete information while making decisions.

The model of Porter and Lawler points out precise fields of application in practice: Each component is practicable and the processes are just as easy to understand and to see through. Organizational management can influence the relationship between effort and reward by linking reward directly to work performance. Moreover, it is possible to influence virtually every component in a systematic way because, according to Porter and Lawler, the effects are predictable within certain limits.

### 3.2.3 Adams's Equity Theory

In contrast to the instrumentality theories, which deal essentially with the expectancies of a person at the workplace and with how these expectations influence behavior, the balance theories of motivation focus on interindividual comparisons and on states of tension and their reduction. These are already the general assumptions of this type of motivation model: Behavior is initiated, directed, and sustained by people's attempts to find a kind of internal balance (i.e., to keep their psychological budget balanced). Festinger's (1957) theory of cognitive dissonance serves as a basis for the various versions of balance theories especially designed for work organizations. Simplified, Festinger postulates that discrepant cognitions cause psychological tensions that are perceived as unpleasant and that humans act to reduce these tensions. Hence, if a person has two inconsistent cognitions, the person falls into an aversive state of motivation called *cognitive dissonance*.

The central idea of Adams's equity theory (Adams, 1963, 1965), which has been applied primarily in work organizations and has often been examined (Mikula, 1980), is that employees of an organization make comparisons, on the one hand, between their contributions and the rewards received for it and, on the other, between the contributions and the rewards of relevant other persons in a similar work situation. The choice of the person or group of comparison increases the complexity of the theory. It is an important variable because it can be a person or group within or outside the present organization. As a result, the employee may also compare himself or herself with friends or colleagues in other organizations or from former employments. The choice of the referent is influenced predominantly by information that the employee has about the respective person as well as by the attractiveness of this person. The pertinent research has therefore been interested particularly in the following moderating variables:

- *Gender*. Usually, a person of the same gender is preferred.
- *Duration of Membership in a Company*. The longer the duration of membership, the more frequently a colleague in the same company is preferred.
- *Organizational Level and Training*. The higher the level and the more training a person has, the more likely he or she is to make comparisons with persons outside the organization.

For employees the principle of equity at the workplace is kept when they perceive the ratio of their own contributions (input = $I$) and the rewards obtained (outcome = $O$) as being equivalent to the respective ratio of other persons in the same work situation. Inequity and therefore tension exist for a person when these two ratios are not equivalent. Hence, if this ratio of contribution and reward is smaller or greater for a person than for the referent, the person is motivated to reduce the internal tension produced; one will count as one's contributions everything that one adds personally to a given work situation: psychomotoric or intellectual skills, expertise, traits, or experience. Accordingly, everything that a person obtains and considers as valuable is counted as a reward: remuneration, commendation, appreciation, or promotion. The inner tension perceived by a person pushes for the reestablishment of equity and therefore justice. According to Adams, the strength of the motivated behavior is directly proportional to the amount or strength of tension produced by inequity. Depending on the causes and the strength of the perceived inequity, the person can now choose different alternatives of action that can be predicted by means of the referent. For example, a person can try to get a raise in reward if this is lower than that of the referent chosen. On the other hand, one could increase or decrease one's input by intensifying or reducing one's contributions. If in the case of a perceived inequity neither of these reactions is possible, the person's reaction might be frequent absenteeism from the workplace or even quitting the job. Besides, there are other options as to how to respond: (1) distortion of self-perception, (2) distortion of the perception of others, or (3) variation of the chosen referent.

It is important to state that the contributions as well as the estimation of the rewards and the ratio of the two variables are subject to the perception and judgment of the employee (i.e., they do not necessarily correspond to reality). The majority of research studies on the evaluation of Adams's motivation theory have dealt with the choice of the referent and with payment as a category of reward in work organizations (Husemann et al., 1987; Greenberg, 1988; Summers and DeNisi, 1990; Kulik and Ambrose, 1992). Usually, in these studies for the assessment of the effect of a state of unequal payment in laboratory experiments as well as in field studies, four different conditions have been created: (1) overpayment of the hourly wage, (2) underpayment of the hourly wage, (3) overpayment of the piece wage, and (4) underpayment of the piece wage.

Subjects were assigned in a randomized way to these conditions of inequity, dissonance, and tension. Thus, an employee working under the condition of overpayment/piece wage (i.e., perception of inequity) will improve the quality and reduce the quantity in order to reduce the state of tension because another increase in quantity would augment the state of inequity. In many surveys it turned out to be problematic that the results generally verify the model only under conditions of overpayment and of hourly wage. However, the model could not be supported convincingly concerning overpayment and piece wage (Steers et al., 1996). Moreover, the question of how a person chooses the referent is largely unsolved: whether a person is chosen

within or outside the company and whether the referent is exchanged over the years. Furthermore, there is little knowledge about the strategy of the reduction of tension chosen. Because of a lack of research in this area, it is not yet possible to generalize the applicability of the equity theory with regard to the effect of non-financial rewards. Finally, it seems as if the model's predictable and expectable possibilities to react to solve the problem of inequity that a person can choose from are too limited. Despite these limitations, Adams's motivation model offers the option of explaining and predicting attitudes and reactions of employees (job satisfaction, quitting the job, or absenteeism) on the basis of their rewards and contributions.

### 3.2.4 Locke's Goal-Setting Theory

Locke (1968) holds the view that the conscious goals and aims of persons are the essential cognitive determinants of their behavior. Thereby, values and value judgments play an important role. Humans strive for achievement of their goals to satisfy their emotions and desires. Goals give direction to human behavior, and they guide thoughts and actions. The effect of such goal-setting processes is seen primarily in the fact that they (1) guide attention and action, (2) mobilize effort, (3) increase perseverance, and (4) facilitate the search for adequate strategies of action.

Meanwhile, the goal-setting theory (Figure 7) has attracted wide interest among theorists and practitioners and has received convincing and sustained support by recent research (Tubbs, 1986; Mento et al., 1987). Locke and Latham (1990) point to almost 400 studies dealing solely with the aspect of the difficulty of goals. Through increasing insights into the effect of goal setting on work performance, the original model could be widened.

In organizational psychology, goals serve two different purposes under a motivational perspective: (1) They are set jointly by employees and superiors to serve as a motivational general agreement and mark of orientation that can be aimed at and (2) they can serve as an instrument of control and mechanism of leadership to reach the overall goal of the organization with the help of employees' individual goals.

According to Locke, from a motivational perspective, a goal is something desirable that has to be obtained. As a result, the original theory of 1968 postulates that work performance is determined by two specific factors: the difficulty and specificity of the goal. *Difficulty of the goal* relates to the degree to which a goal represents a challenge and requires an effort. Thereby, to work as an incentive, the goal has to be realistic and achievable. The correctness of this assumption has already been proven by a large number of early studies (Latham and Baldes, 1975; Latham and Yukl, 1975). Locke postulates that the performance of a person can be increased in proportion to the increase in the goal's difficulty until performance reaches a maximum level. *Specificity of goal determination* refers to the degree of distinctness and accuracy with which a goal is set. Correspondingly, the goals to give one's best or to increase productivity are not very specific, whereas the goal to increase turnover by 4% during the next six months is very specific. Goals referring to a certain output or profit or to a reduction in costs are easy to specify. In contrast, goals concerning ethical or social problems and their improvement, such as job satisfaction, organizational culture, image, or working atmosphere, are difficult to grasp in exact terms (Latham and Yukl, 1975).

Set goals do not always result directly in actions but can exist within the person over a longer period of time without a perceivable effect. To become effective, a commitment of the person toward such goals is necessary. The greater the commitment (i.e., the greater the wish to achieve the goal), the more intensive and persistent the person's performance will be influenced by it. It facilitates concentration on action processes and at the same time insulates the person from distractions by potential disturbing variables (e.g., alternative goals). Today, there is abundant research pointing to an extension of the model that would make it possible to meet the complexity of the motivational process concerning setting goals in organizations (Locke and Latham, 1990). A newer version of the theory states that goal-oriented effort is a function not only of the difficulty and specificity of the goal but also of two other goal properties: acceptance and commitment. *Acceptance* refers to the degree to which one views a goal as one's
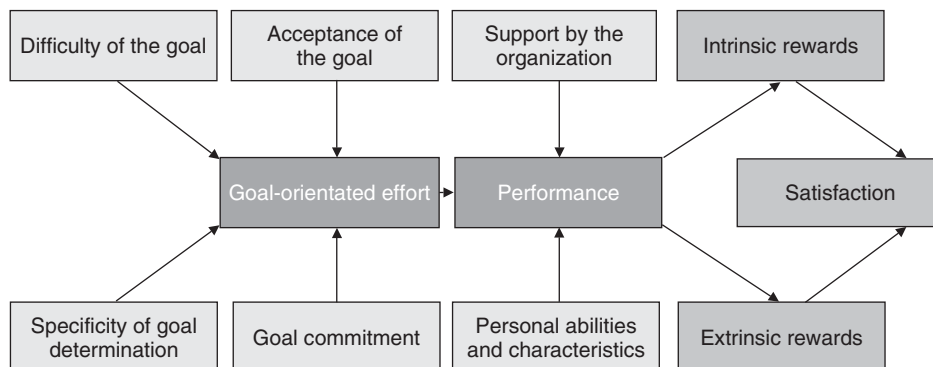


**Figure 7**  Locke's goal-setting theory.

own; *commitment* says something about the degree to which an employee or a superior is personally interested in achieving a goal. Tubbs (1993) pointed out that factors such as active participation while setting goals, challenging at the same time realistic goals, or the certainty that the achievement of the goal will lead to the personally appreciated rewards particularly facilitate goal acceptance and commitment (Kristof, 1996).

The model is being enhanced and improved continuously (Mento et al., 1992; Wofford et al., 1992; Tubbs et al., 1993; Austin and Klein, 1996). The latest developments analyze particularly the role of expectancies, including the differentiation pointed out by Bandura between the *outcome expectation* (the expectation that an action will lead to a certain outcome) and the *efficacy expectation* (the expectation that a person is able to conduct the necessary action successfully) (Bandura, 1982, 1989).

Four central insights are relevant for the practice of organizational psychology in particular:

1. Specific goals (e.g., quotas, marks, or exact numbers) are more effective than vague and general goals (e.g., do your best).

2. Difficult, challenging goals are more effective than relatively easy and common goals. Such goals have to be reachable, though; otherwise, they have a frustrating effect.

3. Accepted goals set in participation are to be preferred over assigned goals.

4. Objective feedback about the advances attained in respect to the goal is absolutely necessary but is not a sufficient condition for the successful implementation of goal setting.

Goal setting offers a useful and important method to motivate employees and junior managers to achieve their goals. These persons will work toward their set goals in a motivated way as long as these are defined exactly and are of a medium difficulty, if they are accepted by employees, and if they show themselves committed to the set goals. The correctness of the assumptions of the theory has been tested in a variety of situations. It turns out that the variables "difficulty" and "specificity" of the goal stand in close relation to performance. Other elements of the theory, such as goal acceptance or commitment, have not been examined that often. Besides, there is little knowledge about how humans accept their goals and how they develop a commitment to certain goals. The question of whether this is a real theory or simply represents an effective technique of motivation has been discussed many times. It has been argued that the process of setting goals constitutes too narrow and rigid a perspective on the employee's behavior. Moreover, it is essential to state that important aspects cannot be quantified that easily. Additionally, goal setting may focus attention on short-term goals, leading to a detriment of long-term considerations. Furthermore, there are other critical appraisals of the theory of goal setting as a motivational instrument of organizational psychology. Setting difficult goals may lead to a higher probability that managers and employees will develop a higher tendency toward risk, which could possibly be counterproductive. In addition, difficult goals can cause stress. Other working areas for which no goals were set might be neglected, and in extreme cases, goal setting may lead to dishonesty and deception.

A practical application of the goal-setting theory is benchmarking. *Benchmarking* refers to the process of comparing the working and/or service processes of one's own company to the best types of processing and production and the results that are detectable in this branch of business. The objective is to identify necessary changes to improve the quality of one's own products and services. The technique of goal setting is used to initiate the necessary activities, to identify the goals that have to be pursued, and to use these as a basis of future actions. To improve one's own work, production, and fabrication, not only the processes within but also those outside the organization are inspected. Through this procedure, several advantages emerge for the company: (1) Benchmarking enables the company to learn from others, (2) the technique places the company in a position to compare itself with a successful competitor, with the objective of identifying strategies of improvement, and (3) benchmarking helps to make a need for change visible in the company by showing how one's own procedures and the assignment of tasks have to be changed and how resources have to be reallocated.

A broader perspective of goal setting in terms of a motivational function is the process of *management by objectives* (MBO). The term refers to a joint process of setting goals with the participation of subordinates and superiors in a company; by doing this, the company's objectives circulate and are communicated top down (Rodgers et al., 1993). Today, MBO is a widespread technique of leadership and motivation featuring substantial advantages: MBO (1) possesses a good potential for motivation, helping to implement the theory of goal setting systematically into the organizational process of a company, (2) stimulates communication, (3) clarifies the system of rewards, (4) simplifies performance review, and (5) can serve managers as a controlling instrument.

Although the technique has to be adjusted to the specific needs and circumstances of the company, there is a general way of proceeding. Top management has to draw up the global objectives of the company and has to stand personally for implementation of the MBO program. After top management has set these goals and has communicated them to the members of the company, superiors and the respective assigned or subordinate employees have to decide jointly on appropriate objectives. Thereby, each superior meets with each employee and communicates the corresponding goals of the division or department. Both have to determine how the employee can contribute to the achievement of these goals in the most effective way. Here, the superior works as a consultant to ensure that the employee sets realistic and challenging goals that are at the same time exactly measurable and verifiable. Finally, it has to be ensured that all resources the employee needs for the achievement of the goals are available. Usually, there are four

basic components of a MBO model: (1) exact description of the objective, (2) participation in the decision making, (3) an explicit period of time, and (4) feedback on the work performed.

Generally, the time frame set for achievement of the objectives is one year. During this period the superior meets on a regular basis with each employee to check progress. It can turn out that because of new information, goals have to be modified or additional resources are necessary. At the end of the set time frame, each superior meets with each employee for a final appraisal conversation to assess to what degree the goals have been reached and the reasons for this conclusion. Such a meeting also serves to revise the proposed figures and performance levels, to determine changes in payment, and as a starting session for a new MBO cycle in the following year.

Overall, goal-setting theory can be considered as being confirmed rather well for individual behavior (Kleinbeck et al., 1990), while its effect on groups or even entire organizations seems to depend on additional conditions that are not yet clarified sufficiently (Miner, 1980). Results are showing that goal setting works better if linked to information about reasonable strategies of action and that both goal setting and strategic information facilitate effort as well as planning behavior (Earley et al., 1987). Moreover, it has to be stressed that the achievement of goals can itself be motivating (Bandura, 1989). Here the thesis of Hacker (1986) is confirmed, stating that tasks not only are directed by motives but also can modify motives and needs.

### 3.2.5 Kelley's Attribution Theory

Research about behavior in organizations has shown that attributions made by managers and employees provide very useful explanations for work motivation. Thus, the theory of attribution (Myers, 1990; Stroebe et al., 1997) offers a better understanding of human behavior in organizations. It is important to point out that, in contrast to other motivation theories, the theory of attribution is, rather, a theory of the relation between personal perception and interpersonal behavior.

As one of the main representatives of this direction of research, Kelley (1967) emphasizes that attribution theory deals with the cognitive processes with which a person interprets behavior as caused by the environment or by characteristics of the actor. Attribution theory mainly asks questions about the "why" of motivation and behavior. Apparently, most causes for human behavior are not observable directly. For this reason, one has to rely on cognitions, especially on perception. Kelley postulates that humans are rational and motivated to identify and understand structures of reasons in their relevant environment. Hence, the main characteristic of attribution theory consists in the search for attributions. At present, one of the most frequently used attributions in organizational psychology is the locus of control. By means of the dimension internal/external, it can be explained whether an employee views his or her work outcome as dependent on either internal or external control (i.e., whether the employee considers himself or herself as able to influence the result personally, e.g.,

through his or her skills and effort, or thinks that the result lies beyond his or her own possibilities of influence and control). It is therefore important that this perceived locus of control has an effect on the performance and satisfaction of the employee. Research studies by Spector (1982) and Kren (1992) point to a correlation of the locus of control and work performance and job satisfaction that is not only statistical in nature. Concerning the relation between motivation and work incentives, it also seems to take a moderating position. Despite the fact that until now the locus of the control dimensions internality and externality for the explanation of work motivation has been the only link on the part of organizational psychology to the attributional approach, it has been suggested repeatedly that other dimensions be examined as well. Weiner (1985) proposes a dimension of stability (fixed vs. variable, e.g., in terms of the stability of internal attribution concerning one's own abilities).

Kelley suggests the dimensions consensus (refers to persons: do others act similarly in this situation?), distinctiveness (refers to tasks: does the person act on this task as on other tasks?), and consistency (refers to time: does the person always act in the same way over time?). These dimensions will influence the type of attribution that is made, for example, by a superior (Kelley, 1973). If consensus, consistency, and distinctiveness are high, the superior will attribute the working behavior of an employee to external (i.e., situational or environmental) causes (e.g., the task cannot be fulfilled better because of external circumstances). If consensus is low, consistency high, and distinctiveness low, causes are attributed to internal or personal factors (e.g., the employee lacks skills, effort, or motivation) (Mitchell and Wood, 1980).

The attributional approach offers promising possibilities for organizational psychology to explain work behavior. However, it has to be mentioned that various sources of error have to be taken into account. For example, there is an obvious tendency of managers to attribute situational difficulties to personal factors (skills, motivation, attitudes). But the reverse also occurs quite frequently: In too many cases, failure is attributed to external factors although personal factors are actually responsible.

### 3.2.6 Summary of Process Models

Process theories refer more closely to actual behavior; they take into account connections between appraisals and results and thereby fill a gap left open by the content theories. Moreover, they do not assume that all people are led by the same motives but emphasize, instead, that each person can have his or her own configuration of desired and undesired facts. They thus open a perspective toward conflicting motives; they can explain why a highly valued behavior might not be executed (e.g., because another one is valued even more or because the expectation to be successful is too small); to the tempting tendency to presume human motives, they oppose the demand not to assume but to study these; and they point out processes that support the connection of highly valued facts and concrete actions (e.g., goal setting, feedback, clear presentation of consequences).

Nevertheless, the question remains unanswered as to what typically is valued as high by which persons—hence, the issue of the contents cannot be evaded. And these are, despite big differences between persons, not so arbitrary that there is nothing that could be said about them. Therefore, the great strength of not assuming any contents becomes a weakness as well. Nevertheless, process models have received a lot of support and they have practical implications. But concerning the contents, not everything should be set aside unseen.

Ultimately, the two approaches deal with completely different aspects. The link between basic motives and specific actions is manifold and indirect, influenced by lots of aspects (e.g., expectations, abilities, situational restrictions); that is why a close direct relationship cannot be expected. To consider the contents (which do not differ much in the various approaches) as usually being effective without presuming them stiffly for each person and to take into account at the same time, the characteristics specified by the process models can provide guidance for practice and theoretical integration possibilities that seem already to have emerged (Locke and Henne, 1986; Six and Kleinbeck, 1989).

### 3.2.7   Recent Developments in Theories of Work Motivation

In the past five years a "revolutionary" new theory of work motivation with practical impact cannot be discovered in the literature. However, under the influence of the management literature, the discussion shifts in focus from the more person-centered view to the more situation-centered view to find, for example, empirical evidence for situations that can be considered as hindering goal-oriented performance items in different elements of the performance picture (Brandstätter and Schnelle, 2007; Nerdinger et al., 2008). The content models are stable in discussion of the "universal" motives but are "unlimited" in the development of motivation factors specified for certain groups and purposes. The process models seem to develop more and more into a section of cognitive theories of goal choice and a section of volitional theories of goal realization. However, both rely on action regulation principles (Nerdinger, 2006; Heckhausen and Heckhausen, 2006).

### 4   POSSIBLE APPLICATIONS OF MOTIVATION MODELS

A question arises as to ways in which the theories of motivation can be applied to the world of work. All of these models can contribute in different ways to a better understanding and predetermination of human behavior and reactions in organizations. This has already been proven for each model. In the following sections we examine the possible applications more closely: involvement and empowerment, remuneration, work and task design, working time, and motivation of various target groups.

### 4.1   Involvement and Empowerment

The role of involvement and empowerment as motivational processes can be seen with regard to need-oriented actions as well as with regard to the postulates of the expectancy theory of motivation. Employee involvement is creating an environment in which people have an impact on decisions and actions that affect their jobs (Argyris, 2001). Employee involvement is neither the goal nor a tool as practiced in many organizations. Rather, it is a management and leadership philosophy about how people are most enabled to contribute to continuous improvement and the ongoing success of their work organization. One form of involvement is *empowerment*, the process of enabling or authorizing a person to think, behave, take action, and control work and decision making in autonomous ways. It is the state of feeling empowered to take control of one's own destiny. Thus, empowerment is a broad concept that aims at reaching involvement in a whole array of different areas of occupation.

Empowerment does not primarily use involvement to reach higher satisfaction and to increase personal performance. The vital point within this process is, rather, the overall contribution of human resources to the efficiency of a company. Employees who can participate in the decision-making process show higher commitment when carrying out these decisions. This addresses simultaneously the two factors in the need for achievement: The person feels appreciated and accepted, and responsibility and self-esteem are increasing.

Furthermore, involvement in decision making helps to clear up expectations and to make the connection between performance and compensation more transparent. Involvement and empowerment can be implemented in various fields, which may concern work itself (e.g., execution, tools, material) or administrative processes (e.g., work planning). Moreover, involvement means participation in relevant decisions concerning the entire company. Naturally, the underlying idea is that employees who are involved in decisions have a considerable impact on their work life, and workers who have more control and autonomy over their work life are motivated to a higher extent, are more productive, and are more satisfied with their work and will thus show higher commitment.

Organizations have experimented for years with lots of different techniques to stimulate involvement and empowerment among their employees and executives.

Quality circles as an integrative approach have attracted particular attention. Their task is to solve work problems with the help of regular discussions among a number of employees. In connection with quality management and qualification programs, quality circles play a crucial role in learning organizations. Not only do they make possible a continuous improvement process and an ideal integration of the experiences of every employee, but they also allow participants to acquire various technical skills and social competences.

To ensure these positive effects, it is crucial that quality circles not pursue solely economic goals. Specific needs and requests of employees regarding improvements in job quality or humanization of work must be

considered equally. However, the evidence is somehow restricted. Although a lot of organizations say that quality circles lead to positive results, there is little research about long-term efficiency (Griffin, 1988).

Another commonly used method of involvement and empowerment of employees is that of work teams. These can be, for example, committees, cross-functional work teams, or interfunctional management teams. They comprise representatives of different departments (e.g., finance, marketing, production) that work together on various projects and activities (e.g., launch of a new product) (Saavedra et al., 1993).

## 4.2   Remuneration

The issue of financial compensation for both executives and employees plays a crucial role in practice (Ash et al., 1985; Judge and Welbourne, 1994; Schettgen, 1996). According to Maslow, payment satisfies the lower needs (physiological and safety needs), and Herzberg considers it as one of the hygienes. Vroom regards payment as a result of the second level that is of particular valence for the employee. According to his VIE theory, the resulting commitment will be high when work leads to a result and a reward highly appreciated by the employee. Adams's equity theory puts emphasis on the relation between commitment and return of work performance. This input/output relation is correlated with a comparison person called a *referent*, a comparison position, or a comparison profession. Perceived inequality leads to attempts at its reduction by changing work commitment or return. Perceived underpayment has, therefore, a negative impact on performance in terms of quality as well as quantity.

The same applies to fringe benefits such as promotions or job security. Three different factors are to be considered which determine the efficiency of payment schemes with regard to the theories of motivation mentioned above: (1) the type of relation between the payment scheme and a person's work performance, (2) the subjective perception of these connections, and (3) different assessments of payment schemes by employees in the same work situation. In addition to that, there are internal performance-oriented benefits, staff shares, and employee suggestion schemes either within or outside the department. Eventually, various additional premiums, such as an increased Christmas bonus or a company pension, have to be mentioned; the latter, especially, is a vital factor in times of diminishing governmental provision.

## 4.3   Work Design

The way in which tasks are combined, the degree of flexibility that executives and employees have, and the existence or lack of important support systems in a company have a direct impact on motivation and on work performance and satisfaction. Motives can be generated within a job by extending work contents and demands.

Organizational psychological research since the 1960s has shown repeatedly that a certain task complexity positively affects work behavior and lots of employees prefer jobs that contain complexity and challenge. But taking a closer look at the actual development of work tasks, one has to conclude that the results noted above have, due largely to economic considerations, generally been ignored, and most work flows have been organized on the basis of economic and technical efficiency. They were primarily attempts to create very specialized work roles and control employees. However, the situation today has changed significantly.

Executives have acknowledged that the most valuable resource available is employee commitment, motivation, and creativity. Only this lets a company stay healthy and competitive in global markets. Technology and downsizing alone can achieve neither flexibility nor new, customer-oriented products and services. It is motivation that must be improved. This can be realized partially by restructuring work tasks and working processes.

### 4.3.1   Strategies of Work Design

**Corrective Work Design**   It is a widespread experience that working systems and work flows have to be changed after their introduction into a company in order to adapt them to specific human needs. Often, these corrections are necessary because of insufficient consideration of anthropometric or ergonomic demands. Such procedures, called *corrective work design*, are always necessary when ergonomic, physiological, psychological, safety–technical, or judicial requirements are not (or not sufficiently) met by planners, design engineers, machine producers, software engineers, organizers, and other responsible authorities. Corrective work design that is at least somewhat effective often causes considerable economic costs. However, its omission can potentially cause physical, psychophysical, or psychosocial harm. Expenditures on corrective work design have to be borne by the companies, whereas the latter costs are carried by the employees affected and thus indirectly by the economy. Both types of costs can be avoided or at least reduced considerably if corrective work design is replaced as far as possible by preventive work design.

**Preventive Work Design**   Preventive work design means that concepts and rules of work science are considered when working systems and work flows are being developed. Hence, possible damage to health and well-being are taken into account when job division between humans and machines is being determined.

**Prospective Work Design**   The strategy of prospective work design arises due to the demand for personality-developing jobs. The criterion of personality development puts an emphasis on the fact that the adult personality develops mainly by dealing with its job. Jobs and working conditions that are personality developing ensure that a person's characteristic strengths can be kept and further developed. Prospective work design means that possibilities of personal development are created intentionally at the stage of planning or reengineering of work systems. This is done by creating a scope of action that can be used and, if possible, extended by employees in different ways. It is crucial for the strategy of prospective work design not to regard it as equivalent to future-oriented work design. Instead, the creation

of work that provides possibilities of development for employees should be seen as a vital feature.

***Differential Work Design***   The principle of differential work design (Ulich, 1990, 1991) takes into account differences between employees: that is, it considers interindividual differences (e.g., different working styles) in dealing with jobs. Employees can choose between different work structures according to their individual preferences and abilities. Since human beings develop through dealing with their jobs, changes among work structures and altering of the structures should be made possible.

The possibility of choosing among alternatives and of correcting a choice, if necessary, means that there will be no need to look for the one best way of organizing jobs and work flows. However, this implies a considerable increase in autonomy and control over one's working conditions. Furthermore, such possibilities of job change lead to a reduction in unbalanced strains.

***Dynamic Work Design***   Dynamic work design does not mean a choice between different existing structures but deals with the possibility of continuously changing and extending existing work structures and creating new ones. Dynamic work design takes into account intraindividual differences (e.g., different learning experiences) in dealing with jobs.

***Empirical Evidence Concerning Differential and Dynamic Work Design***   Without considering interindividual differences, neither optimal personal development nor optimal efficiency can be guaranteed. Differences in cognitive complexity and memory organization may play a role that is just as important as differences in the degree of anticipation, the motivational orientation, the style of learning, or the style of information processing. Empirical data support the assumption that the concept of the one best way, which only needs to be found, constitutes a fundamental and far-reaching error of traditional work design. Moreover, it becomes clear that a standard job structure that is optimal for every employee cannot exist (Zink, 1978). This is consistent with Triebe's (1980, 1981) investigations: In the absence of detailed work schedules, there are interindividually different possibilities as to how to assemble car engines. He observed that workers developed a whole array of different strategies and that these by no means necessarily lead to differing efficiency or effectiveness. Conversely, such results mean that strict work schedules for an operating sequence, which are supposed to be optimal, may sometimes even lead to inefficient work. Differential work design stands out deliberately against the classic search for the one best way in designing work flows. Considering interindividual differences, it is especially appropriate to offer alternative work structures to guarantee optimal personal development in the job.

To take into account processes of personal development (i.e., intraindividual differences in time), the principle of differential work design must be complemented by the principle of dynamic work design. Steinmann and Schreyögg (1980) have pointed out that, when facing choices, some employees might choose the conventional working conditions they are used to. These employees have developed a resigned general attitude and a state of more or less apathetic helplessness because of unchallenging tasks and missing prospects. Hence, it is necessary to develop procedures that make it possible to emphasize a worker's subject position, to reduce barriers to qualification, and to promote readiness for qualification (Alioth, 1980; Ulich, 1981; Baitsch, 1985; Duell and Frei, 1986).

More generally, differential work design can form a link between work design measures, different conditions, and needs of individuals.

It is an important principle of modern work design to offer different work structures. Zülch and Starringer (1984) describe its realization in business by examining the production of electronic flat modules. A macro–work system was created in which differently skilled and motivated employees were simultaneously offered different forms of work organization with different work items. The authors conclude that these new work structures were seen as interesting and motivating.

According to Grob (1985), who provides data and hints for possible extensions, this structure can be applied not only to the production of flat modules but also to all jobs in the company that (1) require several (normally 4–10) employees, (2) have to be carried out frequently in different types and variants, (3) have to be managed with few workshop supplies, and (4) may have a crucial impact on reducing the duration of the cycle time. In this context it is especially important that Zülch and Starringer (1984) were able to prove theoretically that the concept of differential work design can even be realized when facing progressing automation. The production of electronic flat modules for communication devices can serve as an example. It turned out that a useful division into automatic and human operations can be facilitated by not automating all possible operations.

In the beginning, interindividual differences concerning the interaction between individuals and computers were examined almost entirely with regard to the user's role as beginner, advanced user, or expert when dealing with technical systems. It became more and more obvious, however, that the impact of differential concepts goes far beyond that. Hence, Paetau and Pieper (1985) report, with reference to the concept of differential work design, laboratory experiments that examined whether test subjects with approximately the same skills and experiences and given the same work items develop the same preferences for certain systems. Given various office applications, individuals at first preferred a high degree of menu prompting. But with increasing experience, accordance in preferences declined significantly. Due to their results and the experiences and concepts of other authors, Paetau and Pieper (1985) do not see any point in looking for an optimal dialogue design. Demands for programmable software systems, flexible information systems, adaptability of groupware, adaptability of user interfaces, or choices between alternative forms of dialogue put emphasis on the necessary consideration of inter- and intraindividual differences by means of differential and dynamic work design. This has been reflected in the European Community (EC) guideline

90/270/EWG and in International Organization for Standardization (ISO) 9241 (Haaks, 1992; Oberquelle, 1993; Rauterberg et al., 1994). Triebe et al. (1987) conclude that creating possibilities for individualization with the help of individually adaptable user interfaces will probably be one of the most important means of optimizing strain, preventing stress, and developing personality.

Possible achievements and results of the creation of these scopes of development have mainly been examined experimentally. Results of these studies (e.g., Ackermann and Ulich, 1987; Greif and Gediga, 1987; Morrison and Noble, 1987) support the postulate of abolishing generalizing one-best-way concepts in favor of differential work design. Both the participative development of scopes for action and the possible choice between different job structures will objectively increase control in the sense of being able to influence relevant working conditions. This shows at the same time that possibilities of individualization and differential work design are determined at the stage of software development. This is similar for production, where the scope for action is determined mainly by design engineers and planners. Interindividually differing proceedings also matter in design engineers' work. This is revealed in a study by von der Weth (1988) examining the application of concepts and methods of psychological problem solving in design engineering. Design engineers, who acted as test subjects, had to solve a construction task while thinking aloud. Their behavior was registered by video cameras. One of the results that applies to this context says that test persons with adequate problem-solving skills do not show homogeneous procedures. Both the strategy of putting a draft gradually into concrete terms as well as joining solutions of single detailed problems together into a total solution led to success. Thus, the author concludes that an optimal way of reaching a solution that is equally efficient for everybody does not exist. He assumes that this is caused by different styles of behavior which are linked to motivational components such as control needs. As a conclusion, design engineers must be offered a whole array of visual and linguistic possibilities for presenting and linking information. If the system forces a special procedure on users, a lot of creative potential is lost.

***Participative Work Design*** Another strategy, participative work design, has to be considered (Ulich, 1981). In participative work design, all persons concerned with work design measures are included. This participation must focus on all stages of a measure (e.g., including preliminary activities such as evaluation of the actual situation). Participative work design must not result in participation of persons concerned only when one is stuck (e.g., due to technical problems). It must not be a pure measure to get decisions accepted. The various principles overlap and thus cannot often be identified unequivocally. Different strategies of work design pursue different goals that differ not only qualitatively but also in terms of range and time horizon (Ulich, 1980).

The design of work structures implies change in technical, organizational, and social working conditions to adapt structures to workers' qualifications. In this way, they aim at promoting personality development and well-being of workers within the scope of efficient and productive work flows. Criteria are needed to assess jobs and the related design measures concerning this aim. Work psychology and work science can provide a large number of findings and offer support for this (e.g., Oppolzer, 1989; Greif et al., 1991; Leitner, 1993).

### 4.3.2 Characteristics of Task Design

From the point of view of industrial psychology, task design can be seen as an interface between technical or organizational demands and human capabilities (Volpert, 1987). As a consequence, work content and work routine will be determined fundamentally by the design of tasks. Therefore, task design plays a key role in effectiveness, work load, and personality development. Hence, task design takes precedence over the design of work materials and technology, since their use is determined fundamentally by work content and work routine.

Task analysis methods can be used to improve task design. Luczak (1997, p. 341) sees the fundamental idea of task analysis "in a science-based and purpose-oriented method or procedure to determine, what kind of elements the respective task is composed of, how these elements are arranged and structured in a logical or/and timely order, how the existence of a task can be explained or justified ... and how the task or its elements can be aggregated to another entity, composition or compound." Their aim is to transform the task into complete activities or actions.

***Design Criteria*** Tasks should be workable, harmless, and free of impairment and sustain the development of the working person's personality (Luczak et al., 2003). The fundamental aim of task design is to abolish the Tayloristic separation of preparing, planning, performing, and controlling activities (Locke, 2003) so as to make complete activities or actions possible (Hacker, 1986; Volpert, 1987). The results are tasks that offer goal-setting possibilities, the choice between different work modes, and the control of work results. Essentially, it is about granting people a certain scope in decision making.

The attempt in sociotechnical system design (e.g., Alioth, 1980; Trist, 1990) to name other design criteria can be seen as important for the development of abilities and motivation (Ulich, 1993). In addition to the aforementioned autonomy, these are (1) completeness of a task, (2) skill variety, (3) possibilities of social interaction, (4) room for decision making, and (5) possibilities of learning and development. Tasks designed according to these guidelines promote employee motivation, qualifications, and flexibility and are therefore an excellent way to provide and promote the personnel resources of a company in a sensible and economical manner.

Similar dimensions can also be found in the *job characteristics model* of Hackman and Oldham (1976). In the concept of *human strong points* (Dunckel et al., 1993), these criteria are taken up and widened. Human criteria for the assessment and design of tasks and work systems are formulated: (1) Work tasks should have a wide scope concerning actions, decisions, and

time; (2) the working conditions and especially the technology should be easily comprehensible and changeable in accordance with one's own aims; (3) task fulfillment should not be hindered by organizational or technical conditions; (4) the work tasks should require sufficient physical activity; (5) the work tasks should enable dealing with real objects and/or direct access to social situations; (6) human work tasks should offer possibilities for variation; and (7) human work tasks should enable and promote social cooperation as well as direct interpersonal contacts.

According to the definition of work science (Luczak et al., 1989), human criteria are only one theoretically justified way for the assessment and design of work under human aspects (Dunckel, 1996). Neuberger (1985), for example, names other aims of humanization, such as dignity, meaning, security, and beauty. It has to be considered that personality-supporting task design also has a positive effect on the use of technology and on customer orientation; in addition to that, it promises clear economic benefits (Landau et al., 2003). Furthermore, it is interesting to know how work tasks should be designed to create task orientation. Task orientation promotes the development of personality in the process of work and motivates employees to perform tasks without requiring permanent compensation and stimulation from the outside.

***Task Orientation***   Task orientation describes a state of interest and commitment that is created by certain characteristics of the task. Emery (1959) names two conditions for the creation of task orientation: (1) The working person must have the control over the work process and the equipment needed for it and (2) the structural characteristics of a task need to be of a type that sets off in the working person the strength for completing or continuing the work. The extent of control over the work process depends not only on the characteristics of the task or the delegated authority but also, above all, on the knowledge and competence that are brought into dealing with a task.

For those motivational powers that have a stimulating effect on completing or continuing the work, the task itself has to appear to be a challenge with realistic demands (Alioth, 1980). Apart from that, it should be neither too simple nor too complex. Summing up the statements of Emery and Emery (1974), Cherns (1976), and Emery and Thorsrud (1976), the following characteristics of work tasks encourage the process of a task orientation: completeness, skill variety, possibilities for social interaction, autonomy, and possibilities for learning and development. Furthermore, Emery and Thorsrud mention the aspect of meaning. Therefore, work should make a visible contribution to the usefulness of a product for the consumer.

These characteristics correspond so well with the characteristics of tasks derived theoretically by Hackman and Lawler (1971) and Hackman and Oldham (1976) that Trist (1981) points out that this degree of agreement is exceptional in such a new field and has placed work redesign on a firmer foundation than is commonly realized.

***Task Completeness***   An early description of what is now called a *complete task* can be found in Hellpach's article about group fabrication (Hellpach, 1922). He came to the conclusion that it should be a main objective to overcome fragmentation in favor of complete tasks, in the sense of the at least partial restoration of the unity of planning, performing, and controlling. Incomplete tasks show a lack of possibilities for individual goal setting and decision making, for the development of individual working methods, or for sufficiently exact feedback (Hacker, 1987). Research could show, furthermore, that the fragmentation that goes together with classic rationalization strategy can have negative effects on a person in many areas. Restrictions on the scope of action can lead to indisposition and to continuous mental and physical problems. It can also possibly result in the reduction of individual efficiency, especially of mental activity, and passive leisure behavior, as well as in a lower commitment in the areas of politics and trade unions.

Specific consequences for production design resulting from the principle of the complete task may be outlined at this point with the help of some examples:

1.  The independent setting of aims requires a turning away from central control to decentralized workshop control; this creates the possibility of individual decision making within defined periods of time.
2.  Individual preparations for actions require the integration of planning tasks into the workshop.
3.  Choice of equipment can mean, for example, leaving to the constructor the decision of using a drawing board (or forming models by hand) instead of using computer-aided design for the execution of certain construction tasks.
4.  Isolated working processes require feedback as to progress, to minimize the distance and to make corrections possible.
5.  Control with feedback as to the results means transferring the functions of quality control to the workshop itself.

First, a complete task is complete in a sequential sense. Besides mere execution functions, it contains preparation functions (goal setting, development of the way of processing, choosing useful variations in the work mode), coordination functions (divide the tasks among a variety of people), and control functions (get feedback about the achievement of the goals set). Second, complete tasks are complete in a hierarchical regard. They make demands on different alternating levels of work regulation. It should be noted that complete tasks, because of their complexity, can often be designed only as group tasks.

***Job Enrichment***   *Job enrichment* is commonly described as changes regarding the content of an employee's work process. Herzberg (1968) points out that, with the help of this method, motivation is being integrated into an employee's work process to improve

his or her satisfaction and performance. Job enrichment refers to the *vertical* widening of a work role. The aim is to give employees more control concerning planning, performing, and appraisal of their work. Tasks will be organized such that they appear to be a complete module to heighten identity with the task and to create greater variety. Work is to be experienced as meaningful, interesting, and important. Employees will receive more independence, freedom, and heightened responsibility. On top of that, employees will get regular feedback, enabling them to assess their own performance and, if necessary, to adjust it. In this context, customer orientation is of special importance. It will result nearly automatically in direct and regular feedback. The customer can be either internal or external, but the relationship must be direct.

Reports from large companies such as Imperial Chemical Industries (Paul et al., 1969) and Texas Instruments (Myers, 1970) tell of the success of job enrichment programs. Ford (1969) mentioned about 19 job enrichment projects from the American Telephone and Telegraph Company; nine were called extraordinarily successful, nine successful, and one a failure. The success was often rated by means of productivity and quality reference numbers, the rate of times absent and of turnover, and examinations of the attitude of employees.

**Job Enlargement**   *Job enlargement* refers to the *horizontal* expansion of a work role. In this process the number and variety of tasks may be increased to diversify and achieve a motivational effect. Employees will perform several operations on the same product or service. Job enlargement therefore intends to string together several equally structured or simple task elements and, by doing that, to enlarge the work cycle. It becomes obvious that job enlargement touches primarily on the work process, whereas an attempt at job enrichment also concerns the organizational structure. However, it is only the realization of concepts of vertical work expansion that can contribute to overcoming the Tayloristic principle of separating planning and performing and therefore to a work arrangement that develops the personality of an employee. On the other hand, research results have shown that outcomes regarding a heightened challenge or motivation have been rather disappointing (Campion and McClelland, 1993).

**Job Rotation**   *Job rotation* deals with lateral exchange of a work role. If a strong routine in the work becomes a problem for employees, if the tasks are no longer challenging, and the employee is no longer motivated, many companies make use of the principle of rotation to avoid boredom. An employee will usually be transferred from one task to another periodically. This principle is also favorable for the company: Employees with a wider span of experience and abilities allow for more flexibility regarding adaptation to change and the filling of vacancies. On the other hand, this process has disadvantages: Job rotation increases the cost of training, employees will always be transferred to a new position when they are on the highest productive level and have thus reached the highest efficiency,

the process can have negative consequences on a well-operating team, and the process can have negative effects on ambitious employees, who would like to take on particular responsibilities in a chosen position.

**Group Tasks**   Currently, companies employ project teams, quality circles, and working groups as typical forms of teamwork. The main task of *project teams* is to solve interdisciplinary problems. Unlike quality circles and working groups, however, they work together for only a limited period of time and will be dissolved after having found the solution to a certain problem (Rosenstiel et al., 1994). The group will therefore be put together on the basis of professional criteria and will consist mostly of employees in middle and upper management. Well-founded evaluations about the concept of project teams are still missing. That is why Bungard et al. (1993) discover a clear deficiency of research, although project teams gain more and more importance in companies.

It is typical of *quality circles* that groups do not work together continuously; instead, they meet only at regular intervals. Employees get the chance to think about improvements systematically. Attendance is explicitly optional (i.e., employees need to wish to deal with these questions). Other requirements for the success of quality circles are a usable infrastructure in the company (e.g., a conference room and moderation equipment); company support, especially from middle management; and a business culture that is characterized by participation and comprehensive quality thoughts. Behind this concept is the idea that the people affected are better able than anyone else to recognize and solve their own problems. As a side effect, communication among employees will also improve (Wiendieck, 1986a,b).

*Working groups* are organizational units that can regulate themselves within defined boundaries. It therefore is a group that is supposed to solve essential problems with sole responsibility. This work form is, among other things, meant to create motivating work contents and working conditions. The concepts of job enlargement, job enrichment, and job rotation are transferred to the group situation (Rosenstiel et al., 1994; Hackman, 2002).

Psychologically, work in a group has two principal intertwined reasons: (1) The experience of a complete task is possible in modern work processes only where interdependent parts are combined to complete group tasks and (2) the combination of interdependent parts to a common group task makes a higher degree of self-regulation and social support possible. Concerning the first point, Wilson and Trist (1951) as well as Rice (1958) found out that in cases in which the individual task does not allow this, satisfaction can result from cooperation in completing a group task. Concerning the second point, Wilson and Trist are of the opinion that the possible degree of group autonomy can be characterized by how far the group task shows an independent and complete unit. Incidentally, Emery (1959) found out that a common work orientation in a group develops only if the group has a common task for which it can take over responsibility as a group and if it is able to control the work process inside the group.

The common and complete task is what practically all supporters of the sociotechnical approach call a central characteristic of group work (Weber and Ulich, 1993). The existence of a common task and task orientation also has an essential influence on the intensity and length of group cohesion. Work groups whose cohesion is based mainly on socioemotional relations therefore show less stability than do work groups that have a common task orientation (Alioth et al., 1976).

Hackman's (1987) considerations about group work make clear that the organization of work in groups contributes not only to the support of work motivation but also to an increase in work efficiency and therefore in productivity. However, work motivation and efficiency will not develop without organizational efforts and will not remain without any kind of endeavor. A study conducted by a German university together with six well-known companies and more than 200 employees revealed that insufficient adaptation of the organization to the requirements of teamwork is the biggest problem area (Windel and Zimolong, 1998). Another problem is that companies at first sight believe that teamwork is a concept of better value compared to the acquisition of expensive technical systems. Windel and Zimolong stress that even with teamwork investments (e.g., into the qualification of employees) are necessary before the concept pays off in the medium and long terms. For the management of a company, this means that in addition to endurance there needs to be trust in the concept.

Teamwork is associated with a variety of dangers for which a company needs to be prepared (e.g., group targets do not orient themselves toward the overall goals of the company). Therefore, systems are needed that are able to develop complete tasks and that can also be used to orient the motivation potentials toward organizational goals. Arbitrarily used scopes of action and motivation potentials can endanger an organization fundamentally. Such instruments need to include the various areas of responsibility and to turn the work of the group into something measurable in order to compare it to the goals set. In addition, it should offer the possibility of assessing the working results of the group regarding their importance for the entire organization and to give feedback to employees. In this way, the productivity of the organization can be increased because at the same time a high work motivation arises. As a consequence, individual preconditions for performance are used optimally, absenteeism decreases, and work satisfaction increases.

### 4.3.3 Working Time

Motivation, satisfaction, absenteeism, and work performance can be improved within certain limits by means of the implementation of alternative models of working time. It is also possible to explain this improvement in terms of the reduction in need deficiencies, the achievement of a second-level result and its instrumentality, the principle of affiliation, or the motivators developed. The model that is being used most frequently is *flextime*. Here the employee has to stick to certain mandatory working hours, beyond which it is up to him or her how the rest of the workday is arranged. Such a model can be expanded such that extra hours can be saved to create a free day each month. The advantages of this popular model are considerable for both sides and range from the reduction of absenteeism to cost saving and higher productivity and satisfaction to increased autonomy and responsibility at the workplace (Ralston and Flanagan, 1985; Ralston et al., 1985). However, there are a lot of professions for which the model is not applicable. *Job sharing* refers to the division of a job or work week between two or more persons. This approach offers a maximum of flexibility for the employees and for the company.

In the search for new work structures, *teleworking* has been a main point of issue for many years. The term refers to employees who do their job at home at a computer that is connected to their office or their company. The tasks range from programming to processing and analysis of data to the acceptance of orders, reservations, and bookings by telephone. The advantages for the employees and the organization are obvious: For the former it means no traveling to and from work and flexible working hours, and for the latter there is an immense reduction in costs. The disadvantages include a lack of important social contacts and sources of information, and because the employee is no longer integrated into important processes, they might suffer from disadvantages concerning promotions and salary increases.

Other models of working time are also being discussed, such as the *compressed workweek*, with the same number of working hours completed in four days of nine hours each. Supporting this approach, it has been argued that there would be extended leisure time and that employees would not have to travel to and from work during rush hour. It has also been stated that with the help of this model commitment, job satisfaction, and productivity would be increased and costs would be reduced. Extra hours would no longer be necessary, and rates of absenteeism would be lower. Undoubtedly, the acceptance of this model among employees is rather high, but there are also opponents of the approach. They consider the workday as being too long and believe that problems will arise in trying to structure the demands of private life with those of the job.

A working time model that seems to be especially beneficial for older employees or to fight high unemployment is a reduction in the weekly working time without pay compensation. For older employees this eases the transition to retirement. In the scope of measures against unemployment, this model would stand for a fairer allocation of existing work to a greater number of people without increasing total costs. However, for employees it is most important how they are affected personally rather than the positive effects that the model has on a country's unemployment problem.

Another solution that is being discussed, especially for large projects where considerable overtime can be necessary, is to give the employee a time account. As soon as the project is finished, he or she can take up to two months off while receiving his or her regular pay.

### 4.3.4 Motivation of Various Target Groups

In an increasingly diversifying working world (Jackson, 1992; Jackson and Ruderman, 1995) individual differences with respect to needs and expectations are wide. On the part of companies, they are hardly taken into account or taken seriously. For this reason, some motivating measures have little effect. This is why organizations of the future will have to strive for more flexibility regarding the structuring of work and work processes. If they want to maintain the commitment, motivation, and creativity of their managers and employees, they will have to consider family-oriented employees as well as dual-career couples. Organizations of the future will have to show just as much interest in fast-trackers pushing early for leadership responsibility as in employees entering a field that is different from their educational background or midcareer persons wanting to take on a completely new profession.

Different needs and values in various professional groups are crucial for whether or not a motivational measure is effective. In academic professions, especially, employees and managers obtain a considerable part of their intrinsic satisfaction out of their job. They have a very strong and long-lasting commitment toward their field of work, and their loyalty to their special subject is often stronger than that toward their employer. For them, remaining up to date with their knowledge is more important than financial aspects, and they will not insist on a working day with only seven or eight hours and free weekends. What they do possesses a central value in their lives. That is why it is important for these people to focus on their work as their central interest in life. What is motivating them are challenging projects rather than money or leisure time. They wish for autonomy to pursue their own interests and to go their own way. Such persons can be motivated very well by means of further education, training, participation in workshops and conferences—and far less by money.

In future organizations, many employees will be working only temporarily, by project, or part time. People may experience the same working conditions in very different ways. Part-time, project, or temporary work is seen by one group as lacking security or stability, and such employees will not identify themselves with an organization and will show little commitment. But there are also a lot of people for whom this status is convenient. They need a lot of personal flexibility and freedom and are often mothers, older employees, or persons who dislike restrictions by organizational structures. For this group of persons, the long-term prospect of permanent status is more important, and therefore more motivating, than momentary financial incentives. Just as motivating, probably, is the offer of continuous education and training that helps to augment one's market value.

### 4.4 Impact in "Management"

Under the perspective of a practical impact "motivation" plays a considerable role in most newer management discussions of:

- Values of entrepreneurship and sustainability
- Modern principles of leadership and corporate governance in agile enterprises
- Models and strategies for an adaptive and efficient organizational and personnel development
- Identification and realization of innovation chances and/or improvement of competitive processes, especially for hybrid products and the industrialization of services (Luczak and Gudergan, 2007, 2010)

No doubt, the "will to act" is an essential feature of modern management (Hausladen, 2007) that is grounded in human factors research in goal setting and goal realization, for example. Thus human factorsknowledge penetrates into the domain of company organization.

## 5 PRACTICAL EXPERIENCES FROM EUROPEAN STUDIES

Extreme working conditions, which could have been met in the early days of industrialization, with excessive daily hours, children's work, high risks of accidents, and the nonexistence of social security programs, are features of the past, at least in most industrialized countries. In these countries, however, a successive change in attitude with respect to working conditions took place in the early 1960s. Conditions of the working environment such as noise, toxic substances, heat, cold, high physical loads, high levels of concentration, monotonous short-cycle repetitive work, or impaired communication met decreasingly with the pretensions of working people (Kreikebaum and Herbert, 1988; Staehle, 1992). As a result, working people increasingly depreciated shortcomings in work design and answered with "work to rule" (Schmidtchen, 1984) or hidden withdrawal. In Germany, an increased number of complaints about working conditions, increasing dissatisfaction, and decreasing working morale were observed, for instance, by Noelle-Neumann and Strümpel (1984).

Since the early 1970s, the term *quality of working life* (QWL) has developed into a popular issue both in research and in practice. In its early phase it has often been defined as the degree to which employees are able to satisfy important personal needs through their work and experience with the organization. According to the ideas of organizational psychology, projects of organizational development should particularly create a working environment in which the needs of the employees are satisfied. Management saw this movement as a good possibility of increasing productivity and therefore supported it. What were needed were satisfied, dedicated, motivated, and competent employees who were connected emotionally to the organization. These early perceptions of quality of working life have been concretized in the meantime, but also diversified. Like any other movement with broadly and diffusely defined goals, this one produced programs, claims, and procedures differing in respect to content and a variety of at least overlapping terms often being used synonymously. Besides *quality of working life*, there are terms

such as *humanization of work* (Schmidt, 1982; Hettlage, 1983), *sociotechnical systems* (Cummings, 1978), *industrial democracy* (Andriessen and Coetsier, 1984; Wilpert and Sorge, 1984), *structuring of work*, and others.

The typical characteristics of a high-quality work environment can be summarized as follows: (1) adequate and fair payment; (2) a secure and healthy work environment; (3) guaranteed basic rights, including the principle of equality; (4) possibilities for advancement and promotion; (5) social integration; (6) integration of the entire lifetime or life span; (7) an environment that fosters human relations; (8) an organization with social relevance; and (9) an environment that allows employees to have a say or control of decisions concerning them.

Pursuit of these goals for the creation of a work environment with a high-quality work structure reaches back to Herzberg et al. (1959), who assigned them an extremely important role. In research on extrinsic or hygiene factors, Herzberg came to the conclusion that motivation of employees is increased not through higher payment, different leadership styles, or social relations at the workplace but only through considerable change in the type and nature of the work itself (i.e., task design). Certainly, today, there are still a lot of workplaces (e.g., in mass production and in the service sector) that do not meet these requirements, although they have been altered or even replaced to an increasing degree by means of new technologies.

In the 1970s and 1980s, in the light of the quality of working life discussion, many European countries initiated programs of considerable breadth to support research for the humanization of working life. From a motivational point of view, the focus of interest in these action research projects was on (1) avoiding demotivation caused by inadequate workplace and task designs and (2) supporting intrinsic motivation by identifying and designing factors that provide such potential. Researchers basically built on earlier empirical studies that investigated informal groups of workers. Studies were carried out in the United States, where groups in the Hawthorne plant of Western Electric were surveyed in the late 1920s (Roethlisberger and Dickson, 1939), and in Great Britain at the Durham coal mines, where the influences of technology-driven changes on autonomous groups were surveyed in the 1940s (Trist and Bamforth, 1951). In fact, these studies had a considerable influence on such concepts and movements as sociotechnical systems, industrial democracy, quality of working life, and humanization of work, which all are more or less centered around task design and motivation. In the following sections several programs and projects are outlined. This selection is meant neither as a rating of projects nor as a devaluation of programs or projects not discussed.

### 5.1 German Humanization of Working Life Approach

As early as 1922, Hellpach conducted a survey of group work experiments in a German car manufacturing company (Böhrs, 1978). Shifting task design from timed assembly lines toward small groups which were able to organize a certain scope of work autonomously, the company enhanced the content and extent of tasks and therefore fostered intrinsic motivation of the workers. In 1966, the German company Klöckner-Moeller abolished automated assembly lines and changed over to assembling their electrical products at single workplaces and within work groups. Other companies partially followed Klöckner-Moeller, including Bosch, Siemens, BMW, Daimler-Benz, Audi, and Volkswagen (Kreikebaum and Herbert, 1988). However, changes in these companies were by no means as far-reaching as those of some Norwegian or Swedish companies (see Section 5.2.2).

Organizational players in German companies can be divided into (1) representatives of the owners (management), (2) representatives of the employees (works councils), and (3) employees. A works council has to be informed by management of several issues of "information rights." The works council can enter objections to other issues of "participation possibilities," and regarding issues of "participation rights" the works council has to be asked for permission by management: for instance, when it comes to employing new persons. In the German Occupational Constitution Act (Betriebsverfassungsgesetz) of 1972, management and works councils were obligated to regard results of work science in order to design work humanely. Compared with other countries, this was novel.

In 1974, the German Federal Secretary of Research and Technology [Bundesministerium für Forschung und Technologie (BMFT)] initiated the program "Research for the Humanization of Working Life" [Humanisierung des Arbeitslebens (HdA)] and funded about 1600 studies from 1974 to 1988. In 1989, this program was followed by the strategic approach "Work and Technology" [Arbeit und Technik (AuT)], which besides humanization focused particularly on technology and rationalization aspects. In 1999, this program was succeeded by the program "Innovative Work Design—Work of the Future" [Zukunft der Arbeit (ZdA)], where (among others) humanization aspects with respect to the service sector have been of interest. So far, about 3400 single projects have been funded with the help of these three German programs. These projects were accomplished with the assistance of one or more research institutes, which carried out accompanying engineering, psychological, medical, or sociological research. They usually received between 10 and 20% of the funds cited in Figure 8. Objects of research were the survey of actual conditions as well as the implementation and evaluation of solutions with respect to workplace design, task design, or environmental influences in one or more companies or organizations out of all branches and fields that one could imagine.

Therefore, these case studies were carried out primarily with industrial partners from mechanical engineering, the mining and steel-producing industry, the electrical industry, the automotive industry, the clothing industry, the chemical industry, the food-processing industry, and the building industry. However, workplaces were also surveyed in service branches such as the hotel and catering industry, the transport industry (including trucking, harbor, and airports), the retail industry, railway transportation, postal services, merchant shipping, and
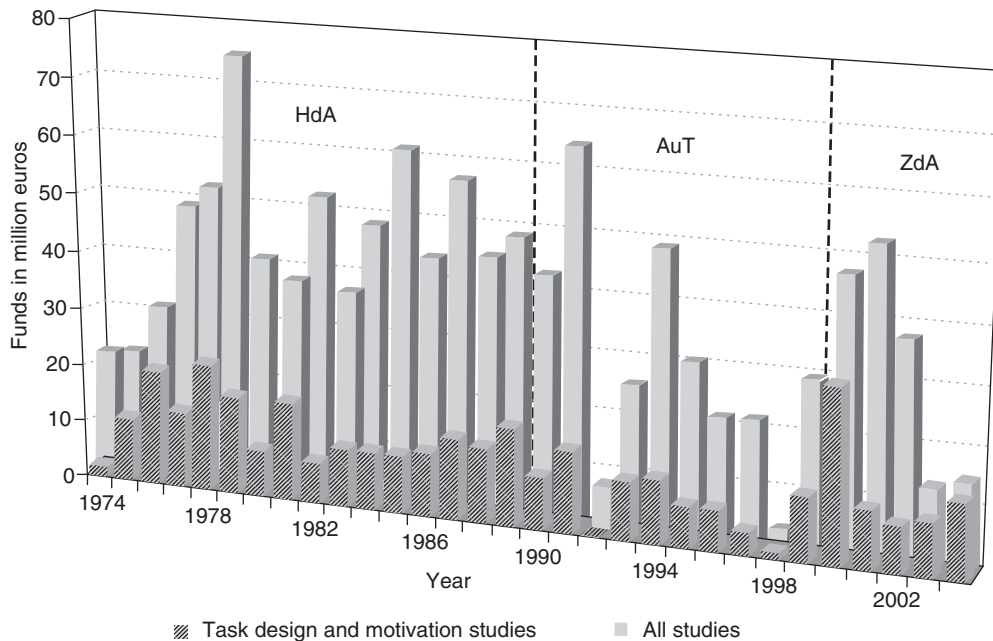
**Figure 8** Funds provided in German ''Humanization of Working Life'' (HdA) studies and the succeeding programs ''Work and Technology'' (AuT) and ''Innovative Work Design—Future of the Work'' (ZdA), in million euros.

health care. In recent years, aspects of humanization have become less important in these projects, which is very unfortunate in light of the fact that employees in the automotive industry, for example, and in the "new economy" faced increasingly inhumane working conditions.

The HdA program, including its successors, AuT and ZdA, is not the only German program that funded research on humanization and quality of working life projects. Regarding the qualification of employees, the program of the Federal Secretary of Economy [Bundesministerium für Wirtschaft (BMFW)] funded projects with 350 million euros from 1994 until 1999, with about 150 million euros being provided by the European Social Fund (ESF) (BMWA, 2003). Additionally, several other programs could be mentioned, such as those initiated by the federal states in Germany.

### 5.1.1 Goals of the HdA Program

In terms of task design and motivation, the HdA program focused initially on avoiding demotivation caused by shortcomings in fulfillment of basic needs of workers. Later, issues such as increasing motivation of workers by introducing new forms of work organization became more and more important. Basic goals as propagated by the Federal Secretary were as follows (Keil and Oster, 1976): (1) development of standards of hazard prevention, reference values, and minimum requirements regarding machines, installations, and workplaces; (2) development of humane working techniques; (3) development of exemplary recommendations and models for work organization and task design; (4) distribution and

application of scientific results and insights; and (5) supporting economy in implementing these insights practically. With respect to particular areas of work design, the focus of interest was on assessing and reducing risks of accidents; environmental influences such as noise, vibration and concussion, and hazardous substances; as well as physical and psychical stress and strain at work.

In the scope of this chapter, the following HdA goal (Keil and Oster, 1976) is of particular interest: Influences of task design should be surveyed with respect to the organization of work processes, structures of decision making and participation, planning of labor utilization, remuneration, and occupational careers as well as satisfaction and motivation. Around 1975, the focus of interest in industrial research projects shifted. Before the shift, projects with a strong focus on reducing risks of accidents or physical stress and strain of workers were funded. After an increase in funding volume in 1975, studies with a focus on introducing new organizational structures in the fields of production or administration began to play a more important role (Kreikebaum and Herbert, 1988).

### 5.1.2 Quantitative Analysis of HdA Studies

Aspects of motivation were of special interest in many of the HdA projects. Even though these aspects played an only partial starring role, they can be considered as a common ground for all studies. Unfortunately, in the beginning of the HdA program, the central funding organization, the German Aerospace Center [Deutsches Zentrum für Luft- und Raumfahrt (DLR)], neglected a systematic documentation of all funded

projects by not forcing the receiving organizations to standardize the documenting of results. Therefore, today a complete investigation of these funded projects is almost impossible.

The quantitative analysis of funded action research studies as depicted in Figures 8-10 draws on the work of Brüggmann et al. (2001), who gathered a database of about 35,000 sets of journal articles from various fields of occupational safety and health (OSH), aiming to identify tendencies in OSH particularly between research work done in different countries. Additionally, project descriptions of more than 4000 projects of HdA and AuT and of other German project-funding institutions, such as the Federal Agency of Occupational Safety and Health [Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BauA)] or the legally obligated Mutual Indemnity Association [Berufsgenossenschaft (BG)], have been gathered. Since only the HdA, AuT, and ZdA studies are of interest in this chapter, the remaining studies and the OSH literature have been ignored. For simplicity, the term "HdA studies" may be used for both AuT and ZdA studies.

In the database there were fields available such as title, accomplishing institutes, funding period, and keywords for studies. Titles of studies combined with keywords provide a high information density which can be compared to articles with available title and abstract. To be able to classify the data sets correctly, an elaborate, hierarchical system of criteria with manifold combinations of logical AND, OR, and NOT matches of thousands of buzzwords was developed by Brüggmann. The classification hierarchy was then validated empirically by several experts.

In this chapter the set of criteria has been expanded and adapted to the scope of task design and motivation. Since the number of available data sets of studies (carried out between 1974 and 2004) as well as the amount of available relevant information varied over time, a representation in relative percentages has been chosen for depicting time-based characteristics; 100% in the diagrams indicates 100% of all data sets that have at least one hit for each criterion.

The results of the quantitative analysis can be used to balance HdA studies with respect to several aspects of task design and motivation. To do so, three boundaries for the respective scope have been chosen. These boundaries, in turn, contain several criteria, which might be of special interest and therefore build up a hierarchical system. In terms of *system theory*, a criterion in one scope can again build up its own scope, such as "task design" in Figure 9. The first scope of interest covered all funded studies versus studies that are relevant for task design and motivation, as depicted in Figure 8. The second scope covered all studies that are relevant for task design and motivation, as depicted in Figure 9.

The first thing that stands out is the continually decreasing proportion of studies that cover aspects of task design, even though the hits for "leadership/ autonomy" could also be counted as "task design" (interpreting them as autonomy according to Hackman and Oldham's job characteristics model). The second remarkable characteristic is the increased proportion of

both "leadership/autonomy" and "incentives." In recent years, especially incentives have become a big issue in funded studies. But aspects of leadership increasingly gained influence. Two representative projects that address questions of leadership are "Modern Services by Innovative Processes of Organization" (MoveOn) and "Flexible Cooperation with Information and Communication Technologies" (SPICE). Breaking "task design" down into "feedback," "task significance," "task identity," or "skill variety," the proportions depicted in Figure 10 can be observed.

Again, light trends can be deduced from the characteristics pictured. The courses of "skill variety" and of "task identity" seem to follow a steady downward tendency. This could be explained by the trend of tasks becoming more and more demanding, making research in this field redundant. An aspect that has become more and more important in recent years is feedback from the task carried out. Aspects of "task significance" gained influence similarly. After all, quantitative analysis gives hints about which fields have been in researchers' and funding organizations' focuses of interest and how these have been objects of shifts over the years.

### 5.1.3  Selected Case Studies

The following is a small selection of typical HdA-funded projects. The aim of this selection is to give an idea of how diversified—in terms of surveyed types of workplaces—the scope of this program was. Furthermore, only case studies from the first years of this program have been selected, since these studies reached a certain degree of recognition within the German scientific community and all of these studies had the character of a role model with respect to succeeding HdA projects. However, many other studies within the HdA program revealed countless scientifically and practically valuable results and insights as well. For more details regarding these studies, the reader is referred to BMFT (1981).

***Electrical Components Industry: The Case of Bosch***  The original title of the Bosch 1 study was "Personalentwicklungsorientierte Arbeitsstrukturierung" ("Structuring of Work with Focus on the Development of Employees"). It was accomplished by the Institute of Work Science of the Technical University of Darmstadt (IAD), the Institute of Production Technology and Automation of the Fraunhofer-Gesellschaft in Stuttgart (IPA), the Institute of Sociology of the University of Karlsruhe (IfS), and the Working Group of Empirical Research in Education in Heidelberg (AfEB), in cooperation with the Robert Bosch GmbH, a global player in the electric and automotive industry. The project was funded from 1974 to 1980 with about 10 million euros.

The primary goal was the development of new forms of work organization in the field of assembling products of varying complexity, such as car radios, cassette decks, TV sets, speakers, electrical tools, and dishwashers. With respect to task design and motivation in many of the plants considered, the contents of work were expanded in order to realize the concept of job enrichment. Technically, this was done by decreasing the degree of automation. In the plant at Herne, tasks
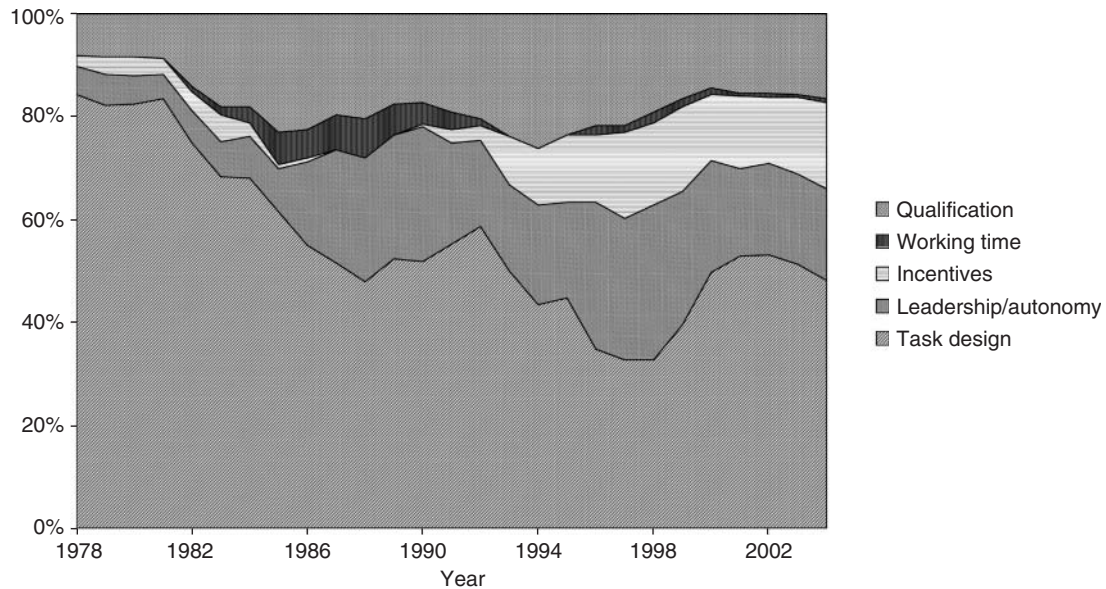
**Figure 9** Proportions (five-year moving averages) of several criteria of HdA studies relevant for task design and motivation.
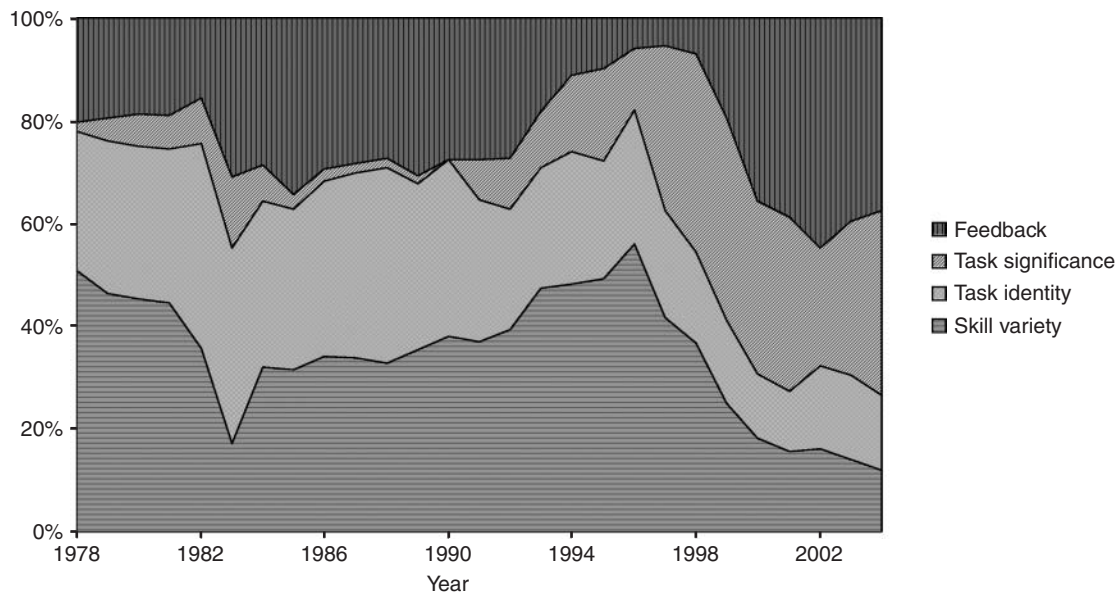


**Figure 10** Proportions (five-year moving averages) of several criteria of task design studies.

for the assembly of car speakers were considered. The time for a task for which an employee was responsible increased from 0.3 to 1.5 min on average. In the plant at Hildesheim, the assembly of car cassette decks was considered. Here, the increase in time was from about 1.0 to about 5.5 min. Considering a flow assembly with an average working time of up to 1 min for each worker, the time spent for one assembly could be

extended to about 1 h by combining functionally different tasks. Additionally, logistical tasks were done by the group. Other projects in the plants of Hildesheim, Leinfelden, and Dillingen were concluded with similar results. The project team established a "learning on the job" concept in these projects, which was close to the job rotation approach. Additionally, a qualification approach for implementing advanced social structures was applied

and evaluated. The scope of actions and decision making was increased by a decoupling of conveyor belts, and the cycle of tasks by means of buffers. Therefore, employees could, to a certain degree, dispose of their own work. The researchers predicted a great potential in the self-disposing of work systems. However, strong participation by employees was considered essential. In some of the groups surveyed, they even found that workers took over disposing activities without any authority.

### Automotive Industry: The Case of Volkswagen

The German title of this case study was "Untersuchung von Arbeitsstrukturen im Bereich der Aggregatefertigung der Volkswagen AG Salzgitter" ("Survey of Structures of Work in the Field of Aggregate-Assembly of Volkswagen"). It was accomplished by the Institute of Work Science of the Technical University of Darmstadt (IAD), the Institute of Work Psychology of the ETH Zurich (IfAP), and the Institute of Production Technology and Automation of the Fraunhofer-Gesellschaft in Stuttgart (IPA) in cooperation with Volkswagen AG, one of the biggest car manufacturers in the world. The project was funded from 1975 to 1978 with about 5.5 million euros.

The primary goal of this project was the analysis of conventional and new forms of work structures in the field of manufacturing aggregates in the automotive industry. Therefore, a qualitative as well as a quantitative evaluation of person-specific as well as monetary criteria was aimed at, both cross-sectioned and longitudinal-sectioned, with a time frame of three years. The objectives of the research were the concepts of (1) conventional assembly line with pallets, (2) intermittent transfer assembly, and (3) assembly groups. In all three alternatives, the variables of feasibility, tolerability, reasonability, and satisfaction were investigated. Remarkable at this point were the open-minded employees of Volkswagen: 268 of 450 potential participants volunteered to participate in the project.

The alternative of assembly groups embodied the concept of job enrichment, where tasks formerly automated were now handled additionally by the group. Even though this implied partially increased stress for the persons involved, the overall distribution of stress was perceived as being more favorable. After all, psychologists found that satisfaction of workers in assembly groups was significantly higher than in the other alternatives. Furthermore, this work was perceived as being more demanding.

Finally, Volkswagen evaluated the results and came to the following conclusions:

1. Many improvements in work and task design will be accounted for in future corporate planning.

2. Stress resulting from work in all three alternatives was on a tolerable level; only partially significant differences could be verified.

3. The proposed new form of work organization competed with established rules, agreements, and legal regulations for distribution of tasks in companies.

4. Mandatory preconditions for implementing a comprehensive process of qualification were individual skills and a methodical proceeding incorporating adequate tools and techniques.

5. Decisions regarding the introduction of new forms of work organization depended on expectations about long-term improvements resulting from that change; the evaluation of economical issues was crucial in that respect.

6. The research project made clear that, in opposition to common positions, a change in working conditions does not necessarily lead to substantial improvement. But it appeared that increased consideration of employees' desires and capabilities regarding assignment to tasks is leading to motivation of these employees.

From an economical point of view, assembly groups were considered to be cost-effective only for small lot sizes of up to 500 motors a day.

### Clothing Industry: A Case of a Total Branch

The German title of this branch project was "Neue Arbeitsstrukturen in der Bekleidungsindustrie—Branchenvorhaben" ("New Structures of Work in the Clothing Industry—Branch Projects"). It was accomplished by the Institute of Operations Research in Berlin (AWF), the Institute of Economical Research in Munich (IFO), the Country's Institute of Social Research in Dortmund, the Institute of Stress Research of the University of Heidelberg, and the Research Institute of Hochenstein, in cooperation with the German clothing union, the German Association of Clothing Industry, as well as the companies Weber, Bierbaum & Proenen, Bogner, Patzer, and Windsor. The project was funded from 1977 to 1993 with about 24.5 million euros. In fact, this was the first German project that investigated an entire branch of industry.

The primary goal of this project was the identification of possible and convertible improvements in clothing production processes. With respect to task design and motivation, changes in organizational structures and participation of employees were considered as a necessary condition to realize these improvements rather than as the actual object of research. However, it was found that shortcomings regarding task design were evident before the project started: unchallenging work; short cycle times; one-sided physical stress; piece-rate-, quality-, and time-based remuneration; low scope of disposing one's own tasks; social isolation; and demotivating leadership by directing and controlling. Only a few months after improvements in workplace and task design had been implemented (e.g., setting up groups with an enhanced scope of action), the first positive results regarding job satisfaction, communicativeness, increased qualification, and degree of performance were observed. Hence, it was possible to prove that the changes did indeed lead to increased motivation of employees. These new structures also proved to be useful in economic terms. The involved companies soon began to implement these structures in other fields as well.

### Services in Public Administration: The Case of the Legal Authority of the City of Hamburg

The German title of this project was "Verbesserung der Arbeitbedingungen bei gleichzeitiger Steigerung der Effizienz des Gerichts durch Einführung von Gruppenarbeit in den Geschäftstellen" ("Improvement of Working Conditions and Increase of Efficiency of the Legal Authority by Introduction of Group Work in Offices"). It was accomplished by the Consortium for Organizational Development in Hamburg and the Research Group for Legal-Sociology in Hannover in cooperation with the Legal Authority of the City of Hamburg. The project was funded from 1977 to 1981 with about 0.8 million euros.

The primary goals of the project were improving working conditions, improving employee job satisfaction, guaranteeing efficient operation, and decreasing the duration of processes. Therefore, group offices were implemented which could react more flexibly in the case of a varying workload. Additionally, a cutback of hierarchical structures combined with improved participation and qualification of employees should be reached. Together with a better design of communication structures and flows of information, improved service should be reached. Therefore, the researchers established four model groups: two in the field of civil law and two in the field of criminal law. The results implied that introduction of group work in large courts would be an adequate way of encountering dysfunctions by means of division of labor. Regarding the introduction of novel office technology, the results implied that an appropriate accompaniment by organizational changes can be seen as necessary. Otherwise, partial overload of employees would be possible.

### Metal Work Industry: The Case of Peiner AG

In terms of employee participation, the *Peiner model* attracted considerable attention within the German scientific community. Therefore, this project is presented here in more depth. The German Research Institute of the Friedrich-Ebert-Stiftung accomplished this action research study from 1975 to 1979 in cooperation with Peiner AG, an incorporated company in the metal work industry with about 2000 employees at that time. In the period of 1973–1974, one year before the project started, Peiner AG closed its balance sheet with 10 million euros of losses. Up to 1977–1978, these losses were reduced by 80%. The following description of the project is based on the final report of the Research Institute of the Friedrich-Ebert-Stiftung (Fricke et al., 1981).

At the beginning of the study, machinery and installation at plant I, which was mainly producing screws, were in bad shape. Production in plant I was characterized by small lot sizes and short delivery times. Due to an unsteady supply of incoming orders, utilization of both workers and machines was changing with some degree of uncertainty. Therefore, employees had an average employment guarantee of only a few days. Furthermore, wages were coupled with particular activities in the production process. Since workers had to be extremely flexible concerning the tasks they had to perform in one day (which was not meant as sanitizing

work in the sense of job rotation but was born from the necessity of the production situation), wages could differ from day to day. Altogether the situation represented considerable uncertainty for the workers. The focus of interest of the team of researchers from the Friedrich-Ebert-Stiftung was division ZII of plant I. In terms of the production flow, chipping, which was carried out in ZII, succeeded the warm-forming and cold-forming divisions. At the start of the project, 47 employees worked at ZII, where high rates of fluctuation to other divisions were observed.

With respect to workplace design, tremendous shortcomings could be found: Due to the nonergonomical shapes of machines, machine workplaces did not leave workers a choice of whether they prefer to work standing or sitting. The working heights of machine workplaces were not personally adjustable. For small persons it was not even possible to place a small platform in front of the machines. Reaching spaces for machines in ZII were designed without any consideration of percentiles of human arms. Therefore, while working in a standing position, joints, muscles, ligaments, and the vertebral column could easily be overstrained. Some control pedals on a special machine forced workers to stay solely on one foot during an entire shift of eight hours. Each raw part had to be placed into the machines manually. Especially with heavy parts, this was painful for workers. Due to boxes, bins, hand gears, actuators, raw parts, parts of machines, or tools around the machine and behind the workers, freedom of movement was cut down dramatically. These types of enforced bearing caused tremendous impairments of the human body. To actuate machines, workers had to expend enormous force. Changing from machine to machine made these workers suffer from adjustment pain, which could last a few days. Furthermore, these workers often suffered from inflammation of the synovial sheath of tendon. To keep track for piece-rate purposes, frequent clearance of chipping boxes was neglected until the boxes were overly full. Emptying 35-kg heavy boxes at a height of 1.50 m often led to overstraining by female workers and sometimes by male workers, too. To refill the cutting oil emulsion in machines, workers had to tow heavy buckets of this fluid. Various other findings indicated tremendous shortcomings in terms of today's standards of occupational safety and health.

With respect to task design, several factors that promoted physical and psychical overstrain could be observed at ZII in 1975. In addition to socially unfavorable conditions, this led to systematic demotivation and therefore to considerable withdrawal of workers at ZII, manifesting in a high rate of fluctuation. The situation in 1975 could be characterized by a high degree of division of work. A foreman was responsible for assigning jobs to machines or workers. He received all necessary papers a few weeks before the start of production from the division of job preparation. However, nobody could tell when the jobs would arrive at ZII. Usually, if the job arrived in the preliminary division, the respective craftsman sent a status message. If jobs were urgent, deadlines were short, or utilization of ZII was low, the craftsman headed in person for the next job. Division of

labor was then realized by providing several services, which were carried out by setters, mechanics, weighing machine workers, pallet jack drivers, inspectors, and shop floor typists.

Machinery workers did not have fixed workplaces and therefore changed on the basis of demand between different machines. Maintenance work on machines was carried out only if it was really necessary. The pallet jack drivers got direct orders from the foremen as to where to provide which raw materials and where to put which final materials. The setters got direct orders from the foremen as to which tools to prepare and in which machines tools had to be changed. Machinery workers got to know on which machines to work next only during the actual shift. Beginning work on a particular machine, the piece-rate ticket had to be stamped. The actual work consisted of depositing up to five parts in parallel into the machines in which they were processed. Each of these processes had to be controlled particularly. If processing of a particular piece was finished, the carriages had to be reset in the starting position by the worker. Work cycles were up to 1.8 s per piece. If the target was 100 pieces within 5 min, the performance rate was about 140%. Short cycles carried out over an entire shift of 8 h caused tremendous stress, resulting from monotony. Additionally, the workers had to spent a certain degree of attention to avoid injury to their hands, to control the quality of processed parts and to track the even operation of the machines, to refill cooling emulsion if necessary, and to request a setter if tools were about to lose sharpness. Even though workers spent this degree of attention, they soon lapsed into an automated working mode where they basically reacted habitually. Another source of demotivation arose out of the enforced cooperation between machinery workers and setters. Since machinery workers' wages depended on piece rates, the workers usually reacted angrily if the setters did not appear instantly or did not work quickly enough (from their point of view). Setters, in turn, felt provoked by machinery workers.

This unchallenging work, short cycle times, permanent attention in combination with piece-rate based remuneration, authoritarian leadership, and the other environmental conditions mentioned constituted an enormous source of stress for these employees. In consequence, this led to a cliff-hanging atmosphere, with separation of single workers, as well as competition and conflicts between co-workers.

In carrying out the project, researchers aimed basically at the following goals: First, the source of imagination would come from committed employees' ideas about how to improve work and task design, which Fricke (1975) called *qualification to innovate*. One of the goals was determining the social conditions and requirements that are necessary to mediate, apply, and unfold qualifications to innovate. At the same time, these processes of mediating and applying should indicate how these qualifications to innovate could look and what impact they could have. This goal was based on empirical findings that employees may, in fact, have qualifications enabling them to formulate innovative changes, yet they were hindered by a variety

of environmental factors, including firmly established organizational structures or resistance from colleagues or superiors. Since from a scientific point of view qualification to innovate is a potential for action, it would have to be observed to enable drawing conclusions about influencing variables. In a normal work environment, however, this observation would not have been possible, for the reasons mentioned above. Therefore, the project was planned as an action research project in which participants were enabled and encouraged to formulate and express such potential.

Furthermore, the researchers aimed at developing and testing approaches for organizing systematic processes of employees' participation in changing work design and task design. This procedure was meant to provide a frame of action for any employee to contribute to the design of working conditions. Therefore, again an action research approach had to be chosen. Finally, together with workers at ZII, actual improvements in workplace and task design were developed, which could be implemented upon approval by management. Hence, the three goals could be pursued in an integrative, simultaneous fashion.

The project could be sectioned into seven phases. The core of the project's phases were workshop weeks, where employees discussed solutions together with researchers. Additionally, project groups were built consisting of employees, ombudsmen, members of the works council, and experts. Actual solutions and suggestions for improvements were proposed to the plant's management and, when approved, were implemented. In each of these phases, countless discussions and meetings took place with superiors and management as well as with the works council. Additionally, several economic, ergonomic, and medical surveys were conducted.

The systematic procedure for employees' participation made clear that even unskilled workers are both willing and capable of participating in the design of workplaces and tasks, and therefore employees do have qualifications to innovate. With respect to the second goal, the following results could be presented. The systematic procedure of employees' participation turned out to be one possible way toward decentralization of decision-making structures, not only in industrial organizations. The research revealed preventable problems occurring when employees are not involved in workplace and task design processes. Participatory workplace and task design can lead to an increase in productivity. Fricke et al. (1981) note that such gains in productivity must not be misused. Therefore, agreements regarding distribution of resulting time in the form of rest periods, reduction in working time, and looking ahead to technical–organizational changes have to be met. From their point of view, these new approaches could be useful completions to existing legal forms of organizational participation. With respect to the third goal, the results of the six workshops should be mentioned. Altogether, 150 pages of suggestions by employees at ZII are evidence of the innovative potential and motivation of these workers.

A major result of this project was an official agreement between Peiner AG and employees: "Participation

of Employees in Designing Work Places, Tasks and Environment" in 1979 (Fricke et al., 1981). The results of the project with respect to task design and motivation of employees at ZII were not overwhelming but could be seen as a starting point for further research projects. Employees confirmed that, in addition to improvements in physical working conditions, they gained considerably in self-confidence, everyone felt "free," everyone was more capable of discussing needs and ideas, and everybody talked increasingly about working conditions and task design. Employees learned how to express their needs and who to contact in certain situations. Employee motivation to innovate in their direct work environment and task design increased significantly.

## 5.2 European Approaches to Humanization of Work

### 5.2.1 Employee Participation in Europe

Especially during the Industrial Revolution, voices that criticized inhumane working conditions in industry gained increasing influence, leading to the development of unions and labor parties. In many European countries this political influence resulted over time in legal regulations that at least assured a minimum of human rights and human dignity for industrial employees.

Scandinavian countries particularly, but also France and the Netherlands, legally codified these rights, including several forms of employee participation, in specific labor acts. With the renewed European Union guideline RL 89/391, a sort of constitution for occupational safety and health was enacted in 1989, which for the first time contained the concept of employee participation (Kohte, 1999). In this guideline as well as in country-specific legislations, participation is always implemented in the form of democratic institutions within companies.

Regarding distribution of power, two types of participation can be distinguished: (1) unilateral participation, where rules of communication and decision making are implemented either by management or by employee representatives (usually, by unionlike institutions), or (2) multilateral participation, where rules of communication and decision making may have been negotiated between management and employee representatives (Kißler, 1992). In neither of these approaches do actual workers have much influence on, or are in charge of, their actual work environment, including task design.

In opposition to legally implemented forms of participation, a management-driven form of participation, called *quality cycles* or *quality circles* (Kahan and Goodstadt, 1999), has become widespread in recent years. In these quality circles, employees gather on a regular basis to discuss actual work-related problems such as work organization, task design, or qualification matters. In many cases these cycles are well accepted by employees since they are considered as opportunities for advancement. However, programs and institutions that are implemented parallel to the actual work processes are likely not to benefit sufficiently from the inherent potential and intrinsic motivation of employees (Kißler, 1992; Sprenger, 2002). However, organizing work within autonomous groups can be seen as one

possible way to overcome these shortcomings, although both really humanize work in terms of needs and dignity as well as meeting economic demands.

### 5.2.2 European QWL Programs

In 1975, the European Foundation for the Improvement of Living and Working Conditions was established by the European Union. This organization comprises members of a respective country's governments, economies, and unions. In the years 1993–1998, this foundation carried out "Employee Direct Participation in Organisational Change" (EPOC), a major program of research dealing with the nature and extent of direct participation and new forms of work organization (Sisson, 2000). Major results were (1) the insight that a significant number of managers consider new forms of work organization as beneficial for reaching conventional business performance goals, such as output, quality, and reduction in throughput time, as well as reducing sickness and absenteeism; (2) that companies adapting new forms of work organization will probably stabilize themselves in long-term perspective, and therefore employment may increase in these companies; and (3) that there are surprisingly only a handful of organizations that actually practice integrated approaches. Uncountable programs and projects aiming at the improvement of QWL have been carried out in most European countries in recent decades.

***Norway*** In the 1960s, Norway's social partners, under the guidance of the psychologist Einar Thorsrud, initiated the "Norwegian Industrial Democracy Project" (NIDP) (Gustavsen, 1983; Kreikebaum and Herbert, 1988; Elden, 2002). Researchers around Thorsrud further developed basic ideas that had originated at the Tavistock Institute in London in the 1950s and invented new conceptual tools (Elden, 2002). The dominating form of research was the action research approach (Gustavsen, 1983). In the focus of interest was the survey of new forms of work organization. In fact, Thorsrud and his colleagues introduced the concept of autonomous groups in several industrial companies, such as Christiana Spigerverk, Hunsfos, and Norsk Hydro (Kreikebaum and Herbert, 1988). In the course of a project with Norsk Hydro, groups were in charge of an entire process beginning with the actual production up to shipping of the product. Since the new tasks were less physically straining but technically more demanding, an enhancement in employee qualifications became necessary.

The common feature in the Norwegian projects was the systematic empowerment of employees to design their own work environment. Some necessary conditions of empowerment are depicted in Table 1.

In 1977, as a "spin-off product" of the effort of all groups and organizations involved, the government enacted the Norwegian Worker Protection and Work Environment Act. This law contained regulations regarding participation of employees. In contrast to conventional legislation, participation in designing their own work environment was mentioned explicitly. In the following years, several agreements between players in the Norwegian economy and endorsements in legislation were aimed at improving QWL further in Norway.

**Table 1  Some Necessary Conditions for Empowering Participation**

| Norwegian Model | Other Models | Significant Common Features |
|---|---|---|
| Institutional and political support at ''higher'' levels | Some parity of power prior to participation | A rejection of conventional organizational design and sociotechnical systems as a source of empowerment |
| High levels of cooperation and conflict | Systematic development of bases of power | Recognition that participation can be either cooperative or empowering |
| A vision of how work should be organized | Overcoming resistance to empowerment by the powerless | Recognition of significant differences between organizational and political democracy |
| ''Do-it-yourself'' participative research | | Empowerment as learning legitimates new realities and possibilities for action from the bottom up |
| Researchers act as ''colearners,'' not as experts in charge of change | | |

*Source*: Based on Elden (2002).

**Sweden**   From the late 1960s to the present, Sweden has made considerable efforts to improve conditions of work, therefore humanizing work environments and contents. However, these reforms were in fact born in a debate between ideological voices that focused on humanization aspects and practical voices that focused on rationalization aspects in order to countervail increasing employment of foreign workers. In the 1970s, several legislative acts regarding humanization of work were enacted: Act on Employee Representation on Boards, Security of Employment Act, Promotion of Employment Act, Act on the Status of Shop Stewards, Worker Protection or Safety Act, Act on Employee Participation in Decision Making, and Work Environment Act (Albrecht and Deutsch, 2002).

The majority of Swedish research programs draw on the Swedish Fund of Work Environment [Arbetsmiljöfonden, (AMFO)]. The AMFO is a state authority that is financed by the Swedish employers. Regarding task design and motivation, two large projects can be mentioned that gained considerable recognition within the international scientific community. The first is the case of Saab-Scania, where in the late 1960s and early 1970s, 130 production groups and 60 development groups were established. As a result of these steps in the plant at Södertälje, the rate of fluctuation decreased from 100% in 1968–1969 to 20% in 1972 (Kreikebaum and Herbert, 1988). Furthermore, in the case of motor assembly, the work cycles were decoupled from automated assembly lines. The company soon began to introduce these new forms of work organization in other plants as well.

The second project we mention is the case of Volvo. No other company in the world was as radical at that time in terms of abolishing automated assembly lines. Research experiments focusing on aspects of autonomous groups were conducted in seven plants, of which the plant of Torslanda was the largest. Uncountable experiments with the 8500 employees at Torslanda provided valuable insights into how best to introduce such groups. The new plants at Skövde and Kalmar were later built with the knowledge gathered in Torslanda.

In recent years, AMFO has carried out several succeeding programs for improving QWL. One of these programs was the program for "Leadership, Organization and Participation" (LOM), which from 1985 to 1991 accomplished 72 change projects in 148 public and private organizations. The program was funded with 5 million euros (Gustavsen, 1990; Naschold, 1992). In the period 1990–1995, the "Working Life Fund" [Arbetslivsfonden (ALF)] funded 24,000 workplace programs with about 1500 million euros (Hofmaier and Riegler, 1995). To recognize the weight of this program, one should note that Sweden's overall population was only about 8 million people at that time. The Swedish government established this fund last (but not least) in fear of an overheating economy. Therefore, several regenerating funds were established by AMFO in which employers had to spend up to 10% of their profits. Out of these funds they were able to finance, for example, development programs for their employees for a period of five years.

Actually, a national program for "Sustainable Work Systems and Health" is being carried out by the Swedish Agency of Innovative Systems (VINNOVA) from 1999 to 2006. The main goals of this program regarding QWL are keeping sustainability of organizational structures as well as integrating job design with organizational design. The program is accomplished as an action research approach as well as an action learning approach. The program will be funded with about 23 million euros (Brödner and Latniak, 2002).

**France**   Traditionally, employees' participation in France was based on interaction between the two organizational players: (1) committees of employees (delegués du personnel) and (2) representatives of the employer (comité d'entreprise). The role of the delegués du personnel was to formulate complaints of employees about working conditions, mainly regarding issues of occupational safety and health (Kißler, 1988). In 1982, the

French Secretary of Labour enacted "Auroux's Act" (Lois Auroux). With that a third player gained influence: the employee itself. Based on co-determination-groups (groupes d'expression) employees got the right to directly influence their own working conditions.

In 1973, the French government founded the National Agency for Improvement of Work Conditions [Agence Nationale pour l'Amélioration des Conditions de Travail (ANACT)], which consists of representatives of the government, the economy, and the unions. ANACT, often in association with other French organizations, such as the Improvement of Work Condition Fund [Le Fonds pour l'Amélioration des Conditions de Travail (FACT)], funded several research activities with a focus on occupational health and safety and issues of QWL. Additionally, ANACT provides offices all over France where companies can be consulted in questions of workplace design, work organization, and so on.

In 1983, the French Ministry of Research launched the "Mobilize Technology, Employment, Work Program" [Mobilisateur Technology, Emploi, Travail (TET)] (Tanguy, 1986). This program aimed at establishing research potential and an academic community for investigating, among others, forms of work organization. In 1989, the "Mobilize Man, Work and Technology" [Homme, Travail et Technologie (HTT)] program succeeded. The aim of this program was to investigate all dimensions of work, such as physical, physiological, psychological, social, and organizational. The second goal of this program was to conduct increasing action research.

In 1984, the National Center for Scientific Research [Centre National de la Recherche Scientifique (CNRS)] launched the "Interdisciplinary Research Program on Technology, Work, and Lifestyles" [Programme Interdisciplinaire de Recherche sur les Technologies, le Travail et les Modes de Vie (PIRTTEM)]. This program focused mainly on projects on the development of technology and respective influences on work organization, especially on employees accepting or not accepting new technologies.

From 1983 to 1985, for example, a consortium of ANACT, TET, and the "Action for Improving Work Conditions in the Alsace" [Action pour l'Amérilation des Conditions de Travail en Alsace (ACTAL)], as well as the CNRS research group of Group Lyonnais de Sociologie Industrielle (GLYSI) and a management consultant, accomplished an action research project in the Moulhouse plant of the car manufacturer Peugeot. The project was called "Social and Organizational Impact of Automation and Robotics" [Impact Social et Organisationnel des Automatismes et de la Robotique (ISOAR)] (Coffineau and Sarraz, 1992). The project basically aimed at preparing the company for future investment in automation technology. New organizational and social equilibriums resulting from that change were to be identified. Employees in the affected parts of the plant were to be developed and qualified accordingly. Therefore, a new concept of participation consisting of three hierarchical levels was established. The first level dealt with shop floor working groups comprising superiors, foremen, workers, and union members. The second level consisted of superiors, a production engineering group, union members, and public representatives. Finally, the third level consisted of representatives from management, the unions, and public organizations.

***Great Britain***   In the advent of European humanization approaches, Great Britain's companies could not provide legally codified forms of employee participation (i.e., there were neither any works councils nor any forms of written agreements on the management level) (Heller et al., 1980). However, some of the roots of humanization of working life can be found in Great Britain: namely, in research at the Tavistock Institute of Human Relations in London, where the concepts of sociotechnical systems and the quality of working life (QWL) emerged. Most notably, the Tavistock coal-mining study of Trist and Bamforth (1951) contributed to the insight that technical innovations in the workplace and task design cannot be applied without regard to the social impact these changes can have on employees. Trist discovered that workers in the Durham coal mines who for decades had worked in autonomous groups barely accepted new, technology-driven forms of work which forced them to abandon the social relations that had grown up among the old group. Furthermore, the impact of concepts such as content of work, extent of work, order of tasks, and degrees of control and feedback on the work's output and on the motivation of workers was investigated. The results can be seen as the foundations of well-known concepts: job enrichment, job enlargement, job rotation, and work organization in autonomous groups.

Even though we could mention other projects that investigated humanization and participation aspects in Great Britain (e.g., several British car manufacturers, the aerospace industry, Indian textile mills), no governmental programs comparable to HdA, AuT, AMFO, or ANACT were accomplished at that time. In 1999, the British Prime Minister carried out The "Partnership at Work Fund," which until 2004 has funded projects with about 14 million pounds sterling. The program focuses on improving relations among organizational players where issues of employee participation are of particular interest (Brödner and Latniak 2002).

## 5.3   Summary of European Studies

Changing actual working conditions as well as scientific insights and implications for work design naturally resulted in conflicts among the involved organizational and societal players. Employer federations argued that humanization must not lead to a shift in responsibilities and power. Beyond that, however, voices from this side admitted that humanization goals do not necessarily compete with economical goals. Employee federations criticized many of the studies as leading to an implementation of measures of rationalization, with the consequence of increased unemployment. Apart from that, employee federations such as unions or works committees widely supported the humanization efforts.

Although these humanization studies have been carried out considering different forms of work and workplaces in different types of companies and branches,

one common aspect can be identified in all the projects: task design and motivation. Whether surveying the Durham coal mines, the assembly lines of Bosch, Saab, Volvo, or Volkswagen, the clothing industry, or authorities of cities and countries in retrospect, if one talks about enlarging work contents and extents, the scope of responsibilities, the possibilities for organizing work in groups, and employee qualifications, one also talks about factors that may influence the intrinsic motivation of employees. Speaking with Maslow: social or esteem needs; speaking with Alderfer: growth needs; speaking with Herzberg: satisfiers and dissatisfiers—these were all addressed substantially in these studies. Herzberg's two-factor theory and Hackman and Oldham's job characteristics model can be seen especially as a basic source of inspiration for most of the practical task design solutions that have been surveyed in the action research studies mentioned.

But also in those studies that surveyed primarily the influences of work environment on employee safety and health (which have not been taken into account in Figure 8), issues of motivation actually provide the common ground. That is, whether these studies focused on development of standards of hazard prevention or on reducing physical stress and strain for workers, these issues can be connected to fulfillment of basic needs of employees according to the content theories of motivation.

Recapitulating, one could argue that from an employer's point of view the humanization of work (as well as the quality of working life, work life balance, etc.) debates led to a better understanding of the needs and motives of employees. Facing the tremendous change in attitude with respect to working conditions that occurred in the 1960s, these insights provided room for reducing demotivation of employees substantially, therefore improving productivity and quality of work results. However, there are voices (Sprenger, 2002) that postulate a new shift of employee attitudes with respect to motivational techniques and incentive systems. They give warning of focusing on manipulating employees' extrinsic motivation. In their opinion, such forms of leadership can easily lead to incentive-dependent employees in the best case—or to demotivated employees who feel that they are being treated as immature and are not taken seriously in the worst case. From an employee's point of view, these debates and the resulting changes can be assessed as a noteworthy contribution to improvement in the quality of working life. For single employees, however, these new forms of work organization often come together with increased stress and strain, which can be partially balanced by increased efforts at qualification.

Considering all the new challenges that the information age has evoked for working conditions, one can recognize that there is still a lot of research to be conducted. The development of new forms of work organization leading to increasingly demanding, highly complex tasks has generated a new source of stress for employees which cannot simply be countervailed by measures of qualification. Along with inhumane pressure of time and a competitive culture, this results

more and more in psychosocial diseases such as the well-known burnout syndrome. Furthermore, there is little knowledge of the long-term effects of manipulative motivational techniques on employee motivation and achievement potential. From a practical point of view, commonsense task design that accounts for both motives and the dignity of human beings appears to be one of the keys in facing these challenges.

## 5.4 Recent Developments

Beginning in the 1990s and reinforced by employment problems in 2000++ the focus of most governmental programs for the improvement of working conditions changed from a physical product and production type view in terms of technology to the until-then underrated and underrepresented "services" (Luczak, 1999; Luczak et al., 2004). Nowadays this development has reached the international scientific community and multiple practitioners' groups under the heading "service engineering" (Salvendy and Karwowski, 2010) as a new branch of a comprehensive service science approach. No doubt that human factors thoughts in service product development processes, service production organizations, and teaching approaches for the new discipline of service engineering play a dominant role in this context: "Task and motivation"–related knowledge forms a kernel of ideas and competencies transferable from physical goods' production to service production (Luczak and Gudergan, 2010).

Besides a shift to the task type "service" the "humanization" approaches in public programs were oriented to "good work" too. The term was invented and propagated by the unions to counteract the employers' tendency to precarious forms of work and to find new ways to compensate the tendency to a shrinking membership. In fact the discussion centers around criteria and their combination (Fuchs, 2009; Prümper and Richenhagen, 2009; Landau, 2010) that have a lot to do with "task design and motivation": Good work is:

- Senseful and satisfactory (work satisfaction)
- Qualifying with open ways to career and development possibilities
- Stable and regular in employment
- Well paid (just and reasonable) and balanced in terms of a work to family life account
- Sane, not only safe
- Limited in resource consumption, especially emotional stressors, by leadership and company culture
- Balanced in work intensity and working hours
- Communicative to colleagues, having free information flows, and influential in terms of self-set design possibilities
- And so on

On the whole, the basic idea is that work/tasks can be perceived by the working person as being a fountain of well-being and of personality development and an improvement of self-esteem.

## REFERENCES

Ackermann, D., and Ulich, E. (1987), "On the Question of Possibilities and Consequences of Individualisation of Human–Computer Interaction," in *Psychological Issues of Human–Computer Interaction in the Work Place*, M.Frese, E.Ulich, and W. Dzida, Eds., North-Holland, Amsterdam, pp. 131–145.

Adams, J. S. (1963), "Toward an Understanding of Inequity," *Journal of Abnormal and Social Psychology*, Vol. 67, pp. 422–436.

Adams, J. S. (1965), "Inequity in Social Change," in *Advances in Experimental Social Psychology*, L. Berkowitz, Ed., Academic, New York.

Albrecht, S., and Deutsch, S. (2002), "The Challenge of Economic Democracy: The Case of Sweden," in *Economic Democracy: Essays and Research on Workers' Empowerment*, W. P. Woodworth, Ed., Sledgehammer, Pittsburgh, PA.

Alderfer, C. P. (1969), "An Empirical Test of a New Theory of Human Needs," *Organizational Behavior and Human Performance,* Vol. 4, pp. 142–175.

Alderfer, C. P. (1972), *Existence, Relatedness, and Growth: Human Needs in Organizational Settings*, Free Press, New York.

Alioth, A. (1980), *Entwicklung und Einführung alternativer Arbeitsformen*, Hans Huber, Bern, Switzerland.

Alioth, A., Martin, E., and Ulich, E. (1976), "Semi-Autonomous Work Groups in Warehousing," in *Proceedings of the 6th Congress of the International Ergonomics Association, Washington, DC*, pp. 187–191, Human Factors Society, Santa Monica, CA

Andriessen, E. J. H., and Coetsier, P. L. (1984), "Industrial Democratization," in *Handbook of Work and Organizational Psychology*, P. J. D.Drenth, H.Thierry, P. J.Willems, and C. J.de Wolff, Eds., Wiley, Chichester, West Sussex, England.

Argyris, C. (1964), *Integrating the Individual and the Organization*, Wiley, New York.

Argyris, C. (2001), "Empowerment: The Emperor's New Clothes," in *Creative Management*, J. Henry, Ed., Sage Publications, London, pp. 195–201.

Ash, R. A., Lee, Y.-L., and Dreher, G. F. (1985), "Exploring Determinants of Pay Satisfaction," in *Proceedings of the APA Meeting*, Los Angeles, pp. 22–27.

Atkinson, J. W., and Feather, N. T., Eds. (1966), *A Theory of Achievement Motivation*, Wiley, New York.

Austin, J. T., and Klein, H. J. (1996), "Work Motivation and Goal Striving," in *Individual Differences and Behavior in Organizations*, K. R. Murphy, Ed., Jossey-Bass, San Francisco.

Baitsch, C. (1985), *Kompetenzentwicklung und partizipative Arbeitsgestaltung*, Lang, Bern, Switzerland.

Bandura, A. (1982), "Self-Efficacy Mechanism in Human Agency," *American Psychologist,* Vol. 37, pp. 122–147.

Bandura, A. (1989), "Self-Regulation of Motivation and Action through Internal Standards and Goal Systems," in *Goal Concepts in Personality and Social Psychology*, L. A. Pervin, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 19–85.

Barnes, L. B. (1960), *Organizational Systems and Engineering Groups: A Comparative Study of Two Technical Groups in Industry*, Division of Research, Harvard Business School, Boston.

Blau, G. (1993), "Operationalizing Direction and Level of Effort and Testing Their Relationships to Individual Job Performance," *Organizational Behavior and Human Decision Processes*, Vol. 55, No. 1, pp. 152–170.

BMFT (1981), "Das Programm 'Forschung zur Humanisierung des Arbeitslebens': Ergebnisse und Erfahrungen arbeitsorientierter Forschung, 1974–1980," in *Schriftenreihe "Humanisierung des Arbeitslebens,"* Vol. 1, *Bundesminister für Forschung und Technologie*, Ed., Campus Verlag, Frankfurt on Main, Germany.

BMWA (2003),"Abschlussbericht zum operationellen Programm in der Bundesrepublik im Rahmen der Gemeinschaftsinitiative ADAPT im Förderzeitraum, 1994–1999," ARINCO-Nr. 94.DE.05.077, ESF-Nr. 946001 D8, Bonn, Germany, available: http://www.equal-de.de/download/Abschlussbericht-GI%20BESCHAEFTIGUNG.doc, accessed July 28, 2004.

Böhrs, H. (1978), "Gruppenfabrikation 1922 in einer deutschen Automobilfabrik," in *Veröffentlichungen des REFA zur Humanisierung der Arbeit 2: Artikel aus REFA-Nachrichten und fortgeschrittene Betriebsführung*, REFA, Darmstadt, Germany, 1976–1978.

Borg, I., Braun, M., and Haeder, M. (1993), "Arbeitswerte in Ost- und Westdeutschland, unterschiedliche Gewichte, aber gleiche Struktur," *ZUMA Nachrichten*, Vol. 33, pp. 64–82.

Brandstättler, V., and Schnelle, J. (2007), in *Handbuch der Arbeits- und Organisationspsychologie*, H. Schuler and K. Sonntag, Eds., Hogrefe, Göttingen, pp. 51–65.

Brödner, P., and Latniak, E. (2002), "Moderne Arbeitsformen für Innovation und Wettbewerbsfähigkeit: Nationale Förderprogramme zur Entwicklung neuer Formen der Arbeitsorganisation," European Union Technical Report EMPL-2002-12380-00-00-DE-TRA-00 (EN); available: http://iat-info.iatge.de/aktuell/veroeff/ps/broedner02c.pdf, accessed July 28, 2004.

Brüggmann, M., Rötting, M., and Luczak, H. (2001), "International Comparison of Occupational Safety and Health Research: A Review Based on Published Articles," *International Journal of Occupational Safety and Ergonomics*, Vol. 7, No. 4, pp. 387–401.

Bungard, W., Antoni, C. H., and Lehnert, E. (1993), *Gruppenarbeitskonzepte mittlerer Industriebetriebe: Forschungsbericht*, Ehrenhof-Verlag, Ludwigshafen, Germany.

Campbell, J. P., and Pritchard, R. D. (1976), "Motivation Theory in Industrial and Organizational Psychology," in *Handbook of Industrial and Organizational Psychology*, M. D. Dunette, Ed., Rand McNally, Chicago, pp. 63–130.

Campion, M. A., and McClelland, C. L. (1993), "Follow-up and Extension of the Interdisciplinary Costs and Benefits of Enlarged Jobs," *Journal of Applied Psychology*, Vol. 78, No. 3, pp. 339–351.

Caston, R. J., and Braito, R. (1985), "A Specification Issue in Job Satisfaction Research," *Sociological Perspectives*, April, pp. 175–197.

Cherns, A. (1976), "The Principles of Organizational Design," *Human Relations*, Vol. 29, pp. 783–792.

Coffineau, A., and Sarraz, J. P. (1992), "Partizipatives Management und Unternehmensberatung: Das Projekt ISOAR bei Peugeot-Mulhouse," in *Management und Partizipation in der Automobilindustrie: Zum Wandel der Arbeitsbeziehungen in Deutschland und Frankreich*, L. Kißler, Ed., Centre d'Information et de Recherche sur l'Allemagne Contemporaine, Paris.

Cummings, T. G. (1978), "Self-Regulating Work Groups: A Socio-Technical Synthesis," *Academy of Management Review*, Vol. 3, pp. 625–634.

Deci, E. L. (1975), *Intrinsic Motivation*, Plenum, New York.

Deci, E. L. (2002), *Handbook of Self-Determination Research*, University of Rochester Press, Rochester, NY.

Deci, E. L., and Ryan, R. M. (1985), *Intrinsic Motivation and Self-Determination in Human Behavior*, Plenum, New York.

Deci, E. L., Connell, J. P., and Ryan, R. M. (1989), "Self-Determination in a Work Organization," *Journal of Applied Psychology*, Vol. 74, pp. 580–590.

Deci, E. L., Ryan, R. M., Gagné, M., Leone, D. R., Usunov, J., and Kornazheva, B. P. (2001), "Need Satisfaction, Motivation, and Well-Being in the Work Organizations of a Former Eastern Bloc Country," *Personality and Social Psychology Bulletin*, Vol. 27, pp. 930–942.

Duell, W., and Frei, F. (1986), *Leitfaden für qualifizierende Arbeitsgestaltung*, TÜV Rheinland, Köln, Germany.

Dunckel, H. (1996), *Psychologisch orientierte Systemanalyse im Büro*, Hans Huber, Bern, Switzerland.

Dunckel, H., Volpert, W., Zölch, M., Kreutner, U., Pleiss, C., and Hennes, K. (1993), *Kontrastive Aufgabenanalyse im Büro: Der KABA-Leitfaden*, Teubner, Stuttgart, Germany.

Earley, P. Ch., Wojnaroski, P., and Prest, W. (1987), "Task Planning and Energy Expended: Exploration of How Goals Influence Performance," *Journal of Personality and Social Psychology*, Vol. 72, pp. 107–114.

Elden, M. (2002), "Socio-Technical Systems Ideas as Public Policy in Norway: Empowering Participation through Worker-Managed Change," in *Economic Democracy: Essays and Research on Workers' Empowerment*, W. P. Woodworth, Ed., Sledgehammer, Pittsburgh, PA.

Elizur, D., Borg, I., Hunt, R., and Magyari-Beck, I. (1991), "The Structure of Work Values: A Cross Cultural Comparison," *Journal of Organizational Behavior*, Vol. 12, pp. 21–38.

Emery, F. E. (1959), *Characteristics of Socio-Technical Systems*, Document 527, Tavistock Institute of Human Relations, London.

Emery, F. E. (1972), "Characteristics of Socio-Technical Systems," in *Job Design*, L. E. Davis and J. C. Taylor, Eds., Penguin Books, Harmondsworth, Middlesex, England, pp. 177–198.

Emery, F. E., and Emery, M. (1974), *Participative Design*, Australian National University, Canberra, Australia.

Emery, F. E., and Thorsrud, E. (1976), *Democracy at Work*, Martinus Nijhoff, Leiden, The Netherlands.

Festinger, L. (1957), *A Theory of Cognitive Dissonance*, Row, Peterson, Evanston, IL.

Ford, N. (1969), *Motivation through the Work Itself*, American Management Association, New York.

Frei, F., and Udris, I., Eds. (1990), *Das Bild der Arbeit*, Hans Huber, Bern, Switzerland.

French, E. G. (1958), "Effects of the Interaction of Motivation and Feedback on Task Performance," in *Motives in Fantasy, Action, and Society*, J. W. Atkinson, Ed., Van Nostrand, Princeton, NJ, pp. 400–408.

Fricke, E., Fricke, W., Schönwälder, M., and Stiegler, B. (1981), "Qualifikation und Beteiligung, 'Das Peiner Modell,'" in *Schriftenreihe Humanisierung des Arbeitslebens*, Vol. 12, *Bundesminister für Forschung und Technologie*, Ed., Campus Verlag, Frankfurt on Main, Germany.

Fricke, W. (1975), *Arbeitsorganisation und Qualifikation: Ein industriesoziologischer Beitrag zur Humanisierung der Arbeit*, Neue Schriftenreihe des Forschungsinstituts der Friedrich-Ebert-Stiftung, Vol. 119, Gesellschaft, Bonn Bad-Godesberg, Germany.

Fried, Y. (1991), "Meta-Analytic Comparison of the Job Diagnostic Survey and Job Characteristics Inventory as Correlates of Work Satisfaction and Performance," *Journal of Applied Psychology*, Vol. 76, pp. 690–697.

Fried, Y., and Ferris, G. R. (1987), "The Validity of the Job Characteristics Model: A Review and Meta-Analysis," *Personnel Psychology*, Vol. 40, pp. 287–322.

Fuchs, T. (2009), "Der DGB-Index Gute Arbeit," in *Arbeitsgestaltung als Zukunftaufgabe*, E. Kistler and F. Mussmann, Eds., VSA Verlag, Hamburg, Germany, pp. 186–222.

Greenberg, J. (1988), "Equity and Workplace Status," *Journal of Applied Psychology*, Vol. 73, No. 4, pp. 606–613.

Greif, S. (1983), *Konzepte der Organisationspsychologie: Eine Einführung in grundlegende theoretische Ansätze*, Hans Huber, Bern, Switzerland.

Greif, S., and Gediga, G. (1987), "A Critique and Empirical Investigation of the 'One-Best-Way-Models' in Human–Computer Interaction," in *Psychological Issues of Human–Computer Interaction in the Work Place*, M.Frese, E.Ulich, and W. Dzida, Eds., North-Holland, Amsterdam, pp. 357–377.

Greif, S., Bamberg, E., and Semmer, N., Eds. (1991), *Psychischer Stress am Arbeitsplatz*, Hogrefe Verlag, Göttingen, Germany.

Griffin, R. W. (1988), "A Longitudinal Assessment of the Consequences of Quality Circles in an Industrial Setting," *Academy and Management Journal*, Vol. 31, pp. 338–358.

Grob, R. (1985), *Flexibilität in der Fertigung*, Springer-Verlag, Berlin.

Guest, D. (1984), "What's New in Motivation?" *Personnel Management*, Vol. 16, pp. 20–23.

Gustavsen, B. (1983), *Some Aspects of the Development of Social Science Work Research in Scandinavia*, Internationales Institut für Vergleichende Gesellschaftsforschung und Arbeitspolitik, Berlin.

Gustavsen, B. (1990), "Demokratische Arbeitspolitik mit LOM in Schweden," in *Jahrbuch Arbeit + Technik*, W. Fricke, Ed., Verlag J. H. W. Dietz, Bonn, Germany, pp. 213–228.

Haaks, D. (1992), *Anpaßbare Informationssysteme*, Verlag für Angewandte Psychologie, Göttingen, Germany.

Hacker, W. (1986), *Arbeitspsychologie: Psychische Regulation von Arbeitstätigkeiten*, Hans Huber, Bern, Switzerland.

Hacker, W. (1987), "Software-Ergonomie: Gestalten rechnergestützter Arbeit?" in *Software-Ergonomie '87: Nützen Informationssysteme dem Benutzer?* W. Schönpflug and M. Wittstock, Eds., Teubner, Stuttgart, Germany, pp. 31–54.

Hackman, J. R. (1987), "The Design of Work Teams," in *Handbook of Organizational Behavior*, J. W. Lorsch, Ed., Prentice Hall, Englewood Cliffs, NJ, pp. 315–342.

Hackman, J. R. (2002), *Leading Teams: Setting the Stage for Great Performances*, Harvard Business School Press, Boston.

Hackman, J. R., and Lawler, E. E. (1971), "Employee Reactions to Job Characteristics," *Journal of Applied Psychology*, Vol. 55, pp. 259–286.

Hackman, J. R., and Oldham, G. R. (1975), "Development of the Job Diagnostic Survey," *Journal of Applied Psychology*, Vol. 60, pp. 159–170.

Hackman, J. R., and Oldham, G. R. (1976), "Motivation through the Design of Work: Test of a Theory," *Organizational Behavior and Human Performance*, Vol. 16, pp. 250–279.

Hackman, J. R., and Oldham, G. R. (1980), *Work Redesign*, Addison-Wesley, Reading, MA.

Hackman, J. R., and Suttle, J. L., Eds. (1977), *Improving Life at Work: Behavioral Science Approaches to Organizational Change*, Goodyear, Santa Monica, CA.

Hausladen, I. (2007), Management am Puls der Zeit—Strategien, Konzepte und Methoden TCW—TransferCentrum, München.

Heckhausen, H. (1980), *Motivation und Handeln: Lehrbuch der Motivationspsychologie*, Springer-Verlag, Berlin.

Heckhausen, J., and Heckhausen, H., Eds. (2006), *Motivation und Handeln (Motivation and Action)*, Springer-Publications, Berlin.

Heller, F. A., Tynan, O., and Hitchon, B. (1980), "Der Beitrag der Arbeitsbestimmung zur Arbeitsgestaltung: am Beispiel von British Leyland," in *Arbeitsstrukturierung durch Verhandlung*, H. W. Hetzler, Ed., Forschungsstelle für Betriebswirtschaft und Sozialpraxis, Munich–Mannheim.

Hellpach, W. (1922), "Sozialpsychologische Analyse des betriebstechnischen Tatbestandes 'Gruppenfabrikation,'" in *Gruppenfabrikation*, R. Lang and W. Hellpach, Eds., Springer-Verlag, Berlin, pp. 5–186.

Herzberg, F. (1968), "One More Time: How Do You Motivate Employees?" *Harvard Business Review*, Vol. 46, pp. 53–62.

Herzberg, F., Mausner, B., and Snyderman, B. (1959), *The Motivation to Work*, Wiley, New York.

Hettlage, R. (1983), "Humanisierung der Arbeit: Über einige Zusammenhänge zwischen Wirklichkeitsbildern und Wirklichkeit," *Die Betriebswirtschaft*, Vol. 43, No. 3, pp. 395–406.

Hofmaier, B., and Riegler, C. H. (1995), "Staatlich geförderte Unternehmensentwicklung in Schweden," *Arbeit*, Vol. 3, No. 4, pp. 249–270.

Hoyos, C. Graf (1974), *Arbeitspsychologie*, Kohlhammer, Stuttgart, Germany.

Husemann, R. C., Hatfield, J. D., and Miles, E. W. (1987), "A New Perspective on Equity Theory," *Academy of Management Review*, Vol. 12, pp. 222–234.

Idaszak, J. R., and Drasgow, F. (1987), "A Review of the Job Diagnostic Survey: Elimination of a Measurement Artifact," *Journal of Applied Psychology*, Vol. 72, pp. 69–74.

Inglehart, R. (1977), *The Silent Revolution: Changing Values and Political Styles among Western Publics*, Princeton University Press, Princeton, NJ.

Inglehart, R. (1989), *Kultureller Umbruch: Wertwandel in der westlichen Welt*, Campus Verlag, Frankfurt on Main, Germany.

Iyengar, S. S., and Lepper, M. R. (1999), "Rethinking the Value of Choice: A Cultural Perspective on Intrinsic Motivation," *Journal of Personality and Social Psychology*, Vol. 76, pp. 349–366.

Jackson, S. E. (1992), *Diversity in the Workplace: Human Resources Initiatives*, Guilford, New York.

Jackson, S. E., and Ruderman, M. N. (1995), *Diversity on Work Teams*, American Psychological Association, Washington, DC.

Jahoda, M. (1983), *Wieviel Arbeit braucht der Mensch? Arbeit und Arbeitslosigkeit im 20. Jahrhundert*, Julius Beltz, Weinheim, Germany.

Judge, T. A., and Welbourne, T. M. (1994), "A Confirmatory Investigation of the Dimensionality of the Pay Satisfaction Questionnaire," *Journal of Applied Psychology*, Vol. 79, pp. 461–466.

Kahan, B., and Goodstadt, M. (1999), "Continuous Quality Improvement and Health Promotion: Can CQI Lead to Better Outcomes?" *Health Promotion International*, Vol. 14, No. 1, pp. 83–91.

Katzell, R. A., and Thompson, D. E. (1990), "Work Motivation: Theory and Practice," *American Psychologist*, Vol. 45, pp. 149–150.

Kaufmann, I., Pornschlegel, H., and Udris, I. (1982), "Arbeitsbelastung und Beanspruchung," in *Belastungen und Streß bei der Arbeit*, L. Zimmermann, Ed., Rowohlt Verlag, Reinbek, Germany, pp. 13–48.

Keil, G., and Oster, A. (1976), *Humanisierung des Arbeitslebens*, Honnefer Buchverlag, Bad Honnef, Germany.

Kelley, H. H. (1967), "Attribution Theory in Social Psychology," in *Nebraska Symposium on Motivation*, D. Levine, Ed., University of Nebraska Press, Lincoln, NE, pp. 192–238.

Kelley, H. H. (1973), "The Process of Causal Attribution," *American Psychologist*, Vol. 28, pp. 107–128.

Kerr, S., Harlan, A., and Stogdill, R. M. (1974), "Preference for Motivator and Hygiene Factors in a Hypothetical Interview Situation," *Personnel Psychology*, Vol. 25, pp. 109–124.

Kißler, L. (1988), "Die 'Software' der Modernisierung: Partizipationsexperimente in deutschen und französischen Betrieben," in *Modernisierung der Arbeitsbeziehungen: Direkte Arbeitnehmerbeteiligung in deutschen und französischen Betrieben*, L. Kißler, Ed., Centre d'Information et de Recherche sur l'Allemagne Contemporaine, Paris.

Kißler, L. (1992), "Direkte Arbeitnehmerbeteiligung und Wandel der betrieblichen Arbeitsbeziehungen als Managementaufgabe: Erste Ergebnisse einer empirischen Untersuchung in der deutschen und französischen Automobilindustrie, Das Beispiel Peugeot-Mulhouse," in *Modernisierung der Arbeitsbeziehungen: Direkte Arbeitnehmerbeteiligung in deutschen und französischen Betrieben*, L. Kißler, Ed., Centre d'Information et de Recherche sur l'Allemagne Contemporaine, Paris.

Klages, H. (1983), "Wertwandel und Gesellschaftskrise in der sozialstaatlichen Demokratie," in *Krise der Arbeitsgesellschaft?*, H. Matthes, Ed., Campus Verlag, Frankfurt on Main, Germany, pp. 341–352.

Kleinbeck, U., Quast, H.-H., Thierry, H., and Häcker, H., Eds. (1990), *Work Motivation*, Lawrence Erlbaum Associates, Mahwah, NJ.

Kohte, W. (1999), *Die Stärkung der Partizipation der Beschäftigten im betrieblichen Arbeitsschutz*, Vol. 9, Hans-Böckler-Stiftung, Düsseldorf, Germany.

Kreikebaum, H., and Herbert, K. J. (1988), *Humanisierung der Arbeit: Arbeitsgestaltung im Spannungsfeld ökonomischer, technologischer und humanitärer Ziele*, Gabler Verlag, Wiesbaden, Germany.

Kren, L. (1992, September), "The Moderating Effects of Locus of Control on Performance Incentives and Participation," *Human Relations*, pp. 991–999.

Kristof, A. L. (1996), "Person–Organization Fit: An Integrative Review of Its Conceptualizations, Measurement, and Implications," *Personnel Psychology*, Vol. 49, pp. 1–49.

Kulik, C. T., and Ambrose, M. L. (1992), "Personal and Situational Determinants of Referent Choice," *Academy of Management Review*, Vol. 17, pp. 212–237.

Landau, K. (2010). Editorial to a special issue of the Zeitschrift fur Arbeitswissenschaft on "Good Work," *Zeitschrift für Arbeitswissenschaft*, Vol. 64, No. 1, pp. 1–2.

Landau, K.; Luczak, H.; Keith, H.; Rösler, D.; Schaub, K.; Winter, G. (2003): AKTIV/Innovative Konzepte/Bilanz erfolgreicher Veränderungen in der Arbeitsgestaltung und Unternehmensorganisation. Ergonomia Verlag, Stuttgart

Latham, G. P., and Baldes, J. J. (1975), "The Practical Significance of Locke's Theory of Goal Setting," *Journal of Applied Psychology*, Vol. 60, pp. 187–191.

Latham, G. P., and Yukl, G. A. (1975), "A Review of Research on the Application of Goal Setting in Organizations," *Academy of Management Journal*, Vol. 18, pp. 824–845.

Lawler, E. E., III (1973), *Motivation in Work Situations*, Brooks/Cole, Monterey, CA.

Leitner, K. (1993), "Auswirkungen von Arbeitsbedingungen auf die psychosoziale Gesundheit," *Zeitschrift für Arbeitswissenschaft*, Vol. 47, No. 2, pp. 98–107.

Locke, E. A. (1968), "Toward a Theory of Task Motivation and Incentives," *Organizational Behavior and Human Performance*, Vol. 3, pp. 157–189.

Locke, E. A. (2003), "The Ideas of Frederick W. Taylor: An Evaluation," in *Operations Management: A Supply Chain Approach*, D. L. Waller, Ed., Thomson, London, pp. 73–91.

Locke, E., and Henne, D. (1986), "Work Motivation Theories," in *International Review of Industrial and Organizational Psychology*, C. L. Cooper and I. T. Robertson, Eds., Wiley, Chichester, West Sussex, England, pp. 1–35.

Locke, E. A., and Latham, G. P. (1990), *A Theory of Goal Setting and Task Performance*, Prentice Hall, Englewood Cliffs, NJ.

Luczak, H. (1997), "Task Analysis," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York, pp. 340–416.

Luczak, H. (1999), *Servicemanagement mit System (Systematic service management) –Erfolgreiche Methoden für die Investitionsgüter-Industrie (successful methods for investment-goods industries)*, Springer, Berlin.

Luczak, H., and Gudergan, G. (2007), "Organisation industrieller Dienstleistungen und Innovationsfähigkeit [Organization of Industrial Services and Innovation Capabilities] " in *Management am Puls der Zeit –Strategien, Konzepte und Methoden*, I. Hausladen, Ed., TCW-TransferCentrum, München.

Luczak, H., and Gudergan, G. (2010), "The Evolution of Service Engineering—Toward the Implementation of Designing Integrative Solutions, " in *Introduction to Service Engineering*, G. Salvendy and W. Karwowski, Eds., Wiley, Hoboken, NJ.

Luczak, H., and Rohmert, W. (1985), "Ansätze zu einer anthropologischen Systematik arbeitswissenschaftlicher Erkenntnisse," *Zeitschrift für Arbeitswissenschaft*, Vol. 39, No. 3, pp. 129–142.

Luczak, H., Volpert, W., Raeithel, A., and Schwier, W. (1989), *Arbeitswissenschaft. Kerndefinition–Gegenstandkatalog–Forschungsgebiete*, TÜV Rheinland, Köln, Germany.

Luczak, H., Schmidt, L., and Springer, J. (2003), "Gestaltung von Arbeitssystemen nach ergonomischen und gesundheitsförderlichen Prinzipien," in *Neue Organisationsformen im Unternehmen: Ein Handbuch für das moderne Management*, H.-J. Bullinger, H.-J. Warnecke, and E. Westkämper, Eds., Springer-Verlag, Berlin, pp. 421–458.

Luczak, H., Reichwald, R., and Spath, D. (2004), *Service Engineering in Wissenschaft und Praxis [Service Engineering in science and practice]*, Gabler, Wiesbaden.

Maslow, A. (1954), *Motivation and Personality*, Harper & Row, New York.

McClelland, D. C. (1961), *The Achieving Society*, Van Nostrand, Princeton, NJ.

McClelland, D. C. (1984), *Motives, Personality, and Society*, Praeger, New York.

McClelland, D. C. (1985), *Human Motivation*, Scott, Foresman, Glenview, IL.

McClelland, D. C., Koestner, R., and Weinberger, J. (1989), "How Do Self-Attributed and Implicit Motives Differ?" *Psychological Review*, Vol. 96, No. 4, pp. 690–702.

McGregor, D. M. (1960), *The Human Side of Enterprise*, McGraw-Hill, New York.

Mento, A. J., Steel, R. P., and Karren, R. J. (1987), "A Meta-Analytic Study of the Effects of Goal Setting on Task Performance," *Organizational Behavior and Human Decision Processes*, Vol. 39, pp. 52–83.

Mento, A. J., Klein, H. J., and Locke, E. A. (1992), "Relationship of Goal Level to Valence and Instrumentality," *Journal of Applied Psychology*, Vol. 77, No. 4, pp. 395–405.

Mikula, G., Ed. (1980), *Gerechtigkeit und soziale Interaktion*, Hans Huber, Bern, Switzerland.

Miller, J. G. (1997), "Cultural Conceptions of Duty: Implications for Motivation and Morality," in *Motivation and Culture*, D.Munro, J. F.Schuhmaker, and A. C.Carr, Eds., Routledge, New York, pp. 178–192.

Miner, J. B. (1980), *Theories of Organizational Behavior*, Dryden, Hinsdale, IL.

Miron, D., and McClelland, D. C. (1979), "The Impact of Achievement Motivation Training on Small Business," *California Management Review*, Vol. 21, pp. 13–28.

Mitchell, T. R., and Wood, R. E. (1980), "Supervisor's Responses to Subordinate Poor Performance: A Test of an Attribution Model," *Organizational Behavior and Human Performance*, Vol. 25, pp. 123–138.

Morrison, P. R., and Noble, G. (1987), "Individual Differences and Ergonomic Factors in Performance on a Videotex-Type Task," *Behavior and Information Technology*, Vol. 6, pp. 69–88.

MOW International Research Team (1987), *The Meaning of Work*, Academic, London.

Murray, H. A. (1938), *Exploration in Personality*, Oxford University Press, New York.

Myers, D. G. (1990), *Social Psychology*, McGraw-Hill, New York.

Myers, M. S. (1970), *Every Employee a Manager: More Meaningful Work through Job Enrichment*, McGraw-Hill, New York.

Naschold, F. (1992), *Den Wandel organisieren: Erfahrungen des schwedischen Entwicklungsprogramms 'Leitung, Organisation, Mitbestimmung' (LOM) im internationalen Wettbewerb*, Rainer Bohn Verlag, Berlin.

Nerdinger, F.W. (2006), "Motivierung [How To Motivate]," in *Handbuch Personalpsychologie*, H. Schuler, Ed., Hogrefe, Göttingen, pp. 385–407.

Nerdinger, F.W., Blickle, G., and Schaper, N. (2008), *Arbeits- und Organisationspsychologie (Psychology of work and organization)*, Springer, Heidelberg.

Neuberger, O. (1974), *Theorien der Arbeitszufriedenheit*, Kohlhammer, Stuttgart, Germany.

Neuberger, O. (1985), *Arbeit*, Enke Verlag, Stuttgart, Germany.

Neuberger, O. (1989), "Organisationstheorien, " in *Organisationspsychologie: Enzyklopädie der Psychologie*, Vol. D/III/3, E. Roth, Ed., Hogrefe Verlag, Göttingen, Germany, pp. 205–250.

Nicholson, N., Herzberg, F., Simon, W., Sprenger, R., and Kühl, S. (2004), *Motivation: Was Manager und Mitarbeiter antreibt*, Wirtschaftsverlag Carl Ueberreuter, Frankfurt, Germany.

Noelle-Neumann, E., and Strümpel, B. (1984), *Macht Arbeit krank? Macht Arbeit glücklich?* Piper Verlag, Munich.

Oberquelle, H. (1993), "Anpaßbarkeit von Groupware als Basis für die dynamische Gestaltung von computergestützter Gruppenarbeit," in *Software-Ergonomie in der Gruppenarbeit*, U. Konradt and L. Drisis, Eds., Westdeutscher Verlag, Opladen, Germany, pp. 37–54.

Oishi, S. (2000), "Goals as Cornerstones of Subjective Well-Being: Linking Individuals and Cultures," in *Culture and Subjective Well-Being*, E. Diener and E. Suh, Eds., MIT Press, Cambridge, MA, pp. 87–112.

Oppolzer, A. (1989), *Handbuch Arbeitsgestaltung*, VSA-Verlag, Hamburg, Germany.

Paetau, M., and Pieper, M. (1985), "Differentiell-dynamische Gestaltung der Mensch–Maschine–Kommunikation," in *Software-Ergonomie '85: Mensch–Computer-Interaktion*, H. J. Bullinger, Ed., *Berichte des German Chapter of the ACM*, Vol. 24, Teubner, Stuttgart, Germany, pp. 316–324.

Parker, B., and Chusmir, L. H. (1992), "Development and Validation of a Life-Success Measures Scale," *Psychological Reports*, Vol. 70, No. 1, pp. 627–637.

Paul, P. J., Robertson, K. B., and Herzberg, F. (1969), "Job Enrichment Pays Off," *Harvard Business Review*, Vol. 47, pp. 61–78.

Phillips, J. S., and Lord, R. G. (1980), "Determinants of Intrinsic Motivation: Locus of Control and Competence Information as Components of Deci's Cognitive Evaluation Theory," *Journal of Applied Psychology*, Vol. 65, pp. 211–218.

Podsakoff, P. M., and Williams, L. J. (1986), "The Relationship between Job Performance and Job Satisfaction," in *Generalizing from Laboratory to Field Settings: Research Findings from Industrial-Organizational Psychology, Organizational Behavior, and Human Resource Management*, E. Locke, Ed., Lexington Books, Lexington, MA, pp. 207–254.

Pope John Paul II (1981), *Enzyklika Laborem Exercens: Über die menschliche Arbeit*, Georg Bitter, Recklinghausen, Germany.

Porter, L. W. (1964), *Organizational Patterns of Managerial Job Attitudes*, American Foundation for Management Research, New York.

Porter, L. W., and Lawler, E. E. (1968), *Managerial Attitudes and Performance*, Dorsey, Homewood, IL.

Prümper, J., and Richenhagen, G. (2009), "Arbeitswissenschaftliche Bewertung des DGB-Index 'Gute Arbeit,'" *Zeitschrift fur Arbeitswissenschaft*, Vol. 63, pp. 175–187.

Ralston, D. A., and Flanagan, M. F. (1985, April), "The Effect of Flextime on Absenteeism and Turnover for Male and Female Employees," *Journal of Vocational Behavior*, pp. 206–217.

Ralston, D. A., Anthony, W. P., and Gustafson, D. J. (1985), "Employees May Love Flextime, But What Does It Do to the Organization's Productivity?" *Journal of Applied Psychology*, Vol. 70, pp. 272–279.

Rauterberg, M., Spinas, P., Strohm, O., Ulich, E., and Waeber, D. (1994), *Benutzerorientierte Software-Entwicklung*, Teubner, Stuttgart, Germany.

Rice, A. K. (1958), *Productivity and Social Organization: The Ahmedabad Experiment*, Tavistock Institute of Human Relations, London.

Richer, S. F., and Vallerand, R. J. (1995), "Supervisors' Interactional Styles and Subordinates' Intrinsic and Extrinsic Motivation," *Journal of Social Psychology*, Vol. 135, pp. 707–722.

Roberts, K. H., and Glick, W. (1981), "The Job Characteristics Approach to Task Design: A Critical Review," *Journal of Applied Psychology*, Vol. 66, pp. 193–217.

Rodgers, R., Hunter, J. E., and Rodgers, D. I. (1993), "Influence of Top Management Commitment on Management Program Success," *Journal of Applied Psychology*, Vol. 78, pp. 151–155.

Roethlisberger, F., and Dickson, W. (1939), *Management and the Worker: An Account of a Research Program Conducted by the Western Electric Company, Chicago*, Harvard University Press, Cambridge, MA.

Rosenstiel, L. V., Hockel, C., and Molt, W., Eds. (1994), *Handbuch der angewandten Psychologie: Grundlagen, Methoden, Praxis*, Ecomed, Landsberg, Germany.

Ruiz Quintanilla, S. A. (1984), "Bedeutung des Arbeitens: Entwicklung und empirische Erprobung eines sozialwissenschaftlichen Modells zur Erfassung arbeitsrelevanter Werthaltungen und Kognitionen," dissertation, Technische Universität Berlin.

Ryan, R. M., and Deci, E. L. (2000), "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being," *American Psychologist*, Vol. 55, No. 1, pp. 68–79.

Saavedra, R., Earley, P., and Van Dyne, L. (1993), "Complex Interdependence in Task-Performing Groups," *Journal of Applied Psychology*, Vol. 78, pp. 61–72.

Salancik, G. R., and Pfeffer, J. (1977), "An Examination of Need-Satisfaction Models of Job Attitudes," *Administrative Science Quarterly*, Vol. 22, pp. 427–456.

Salvendy, G. and Karwowski, W., Eds., (2010): Introduction to Service Engineering. Wiley, Hoboken, N. J.

Schettgen, P. (1996), *Arbeit, Leistung, Lohn: Analyse- und Bewertungsmethoden aus sozioökonomischer Perspektive*, Enke Verlag, Stuttgart, Germany.

Schmale, H. (1983), *Psychologie der Arbeit*, Klett-Cotta, Stuttgart, Germany.

Schmidt, G. (1982), "Humanisierung der Arbeit," in *Einführung in die Arbeits- und Industriesoziologie*, W.Litek, W.Rammert, and G. Wachtler, Eds., Campus Verlag, Frankfurt on Main, Germany.

Schmidtchen, G. (1984), *Neue Technik–Neue Arbeitsmoral: Eine sozialpsychologische Untersuchung über Motivation in der Metallindustrie*, Deutscher Institutsverlag, Köln, Germany.

Schneider, B., and Alderfer, C. P. (1973), "Three Studies of Need Satisfaction in Organizations," *Administrative Science Quarterly*, Vol. 18, pp. 489–505.

Sisson, K. (2000), "Direct Participation and the Modernisation of Work Organisation," European Foundation for the Improvement of Living and Working Conditions, available: http://www.eurofound.eu.int/publications/files/EF0029EN.pdf, accessed May 28, 2004.

Six, B., and Kleinbeck, U. (1989), "Motivation und Zufriedenheit in Organisationen," in *Organisationspsychologie: Enzyklopädie der Psychologie*, Vol. D/III/3, E. Roth, Ed., Hogrefe Verlag, Göttingen, Germany, pp. 348–398.

Spector, P. E. (1982), "Behavior in Organizations as a Function of Employee's Locus of Control," *Psychological Bulletin*, Vol. 91, pp. 482–497.

Sprenger R. K. (2002), *Mythos Motivation: Wege aus einer Sackgasse*, Campus Verlag, Frankfurt on Main, Germany.

Staehle W. H. (1992), "Das Bild vom Arbeitnehmer im Wandel der Arbeitgeber-Arbeitnehmer-Beziehungen," in *Modernisierung der Arbeitsbeziehungen: Direkte Arbeitnehmerbeteiligung in deutschen und französischen Betrieben*, L. Kißler, Ed., Centre d'Information et de Recherche sur l'Allemagne Contemporaine, Paris.

Stahl, M. J. (1986), *Managerial and Technical Motivation: Assessing Needs for Achievement, Power, and Affiliation*, Praeger, New York.

Staw, B. M., Bell, N. E., and Clausen, J. A. (1986), "The Dispositional Approach to Job Attitudes," *Administrative Science Quarterly*, Vol. 31, pp. 56–77.

Steers, R. M., Porter, L. W., and Bigley, G. A. (1996), *Motivation and Leadership at Work*, McGraw-Hill, New York.

Steinmann, H., and Schreyögg, G. (1980), "Arbeitsstrukturierung am Scheideweg," *Zeitschrift für Arbeitswissenschaft*, Vol. 34, pp. 75–78.

Stroebe, W., Hewstone, M., and Stephenson, G. M., Eds. (1997), *Sozialpsychologie*, Springer-Verlag, Heidelberg, Germany.

Summers, T. P., and DeNisi, A. S. (1990), "In Search of Adam's Other: Reexamination of Referents Used in the Evaluation of Pay," *Human Relations*, Vol. 43, No. 6, pp. 120–133.

Tanguy, L. (1986), *L'introuvable relation formation/emploi: un état des recherches en France*, Ministère de la Recherche et de l'enseignement superieur, programme mobilisateur technologie, emploi, travail, La Documentation Française, Paris.

Taylor, F. W. (1911), *The Principles of Scientific Management*, Harper, New York.

Thompson, E. P., Chaiken, S., and Hazlewood, J. D. (1993), "Need for Cognition and Desire for Control as Moderators of Extrinsic Reward Effects," *Journal of Personality and Social Psychology*, Vol. 64, No. 6, pp. 987–999.

Triebe, J. K. (1980), "Untersuchungen zum Lernprozeß während des Erwerbs der Grundqualifikationen (Montage eines kompletten Motors): Arbeits- und sozialpsychologische Untersuchungen von Arbeitsstrukturen im Bereich der Aggregatefertigung der Volkswagenwerk AG, Bonn," BMFT HA 80-019, BMFT, Bonn, Germany.

Triebe, J. K. (1981), "Aspekte beruflichen Handelns und Lernens," dissertation. Universität Bern.

Triebe, J. K., Wittstock, M., and Schiele, F. (1987), *Arbeitswissenschaftliche Grundlagen der Software-Ergonomie*, Wirtschaftsverlag NW, Bremerhaven, Germany.

Trist, E. (1990), "Sozio-technische Systeme: Ursprünge und Konzepte," *Organisationsentwicklung*, Vol. 4, pp. 10–26.

Trist, E., and Bamforth, K. (1951), "Some Social and Psychological Consequences of the Longwall Method of Coal-Getting," *Human Relations*, Vol. 4, pp. 3–38.

Trist, E. L. (1981), *The Evaluation of Sociotechnical Systems: Issues in the Quality of Working Life*, Ontario Quality of Working Life Centre, Toronto, Ontario, Canada.

Tubbs, M. E. (1986), "Goal Setting: A Meta-Analytic Examination of the Empirical Evidence," *Journal of Applied Psychology*, Vol. 71, pp. 474–483.

Tubbs, M. E. (1993), "Commitment as a Moderator of the Goal–Performance Relation: A Case for Clearer Construct Definition," *Journal of Applied Psychology*, Vol. 78, pp. 86–97.

Tubbs, M. E., Boehme, D. M., and Dahl, J. G. (1993), "Expectancy, Valence, and Motivational Force Functions in Goal-Setting Research," *Journal of Applied Psychology*, Vol. 78, No. 3, pp. 361–373.

Udris, I. (1979a), "Ist Arbeit noch länger zentrales Lebensinteresse?" *Psychosozial*, Vol. 2, No. 1, pp. 100–120.

Udris, I. (1979b), "Central Life Interests: Werte, Wertwandel, Arbeitspsychologie," *Schweizerische Zeitschrift für Psychologie*, Vol. 38, pp. 252–259.

Udris, I. (1981), "Streß in arbeitspsychologischer Sicht," in *Stress*, J. R. Nitsch, Ed., Hans Huber, Bern, Switzerland, pp. 391–440.

Ulich, E. (1980), "Psychologische Aspekte der Arbeit mit elektronischen Datenverarbeitungssystemen," *Schweizerische Technische Zeitschrift*, Vol. 75, pp. 66–68.

Ulich, E. (1981), "Subjektive Tätigkeitsanalyse als Voraussetzung autonomieorientierter Arbeitsgestaltung," in *Beiträge zur psychologischen Arbeitsanalyse*, F. Frei and E. Ulich, Eds., Hans Huber, Bern, Switzerland, pp. 327–347.

Ulich, E. (1990), "Individualisierung und differentielle Arbeitsgestaltung, " in *Ingenieurpsychologie: Enzyklopädie der Psychologie*, Vol. D/III/2, C. Graf Hoyos and B. Zimolong, Eds., Hogrefe Verlag, Göttingen, Germany, pp. 511–535.

Ulich, E. (1991), *Arbeitspsychologie*, Poeschel, Stuttgart, Germany.

Ulich, E. (1993), "Gestaltung von Arbeitstätigkeiten," in *Lehrbuch Organisationspsychologie*, H. Schuler, Ed., Hans Huber, Bern, Switzerland, pp. 189–208.

Vallerand, R. J., and Reid, G. (1984), "On the Causal Effects of Perceived Competence on Intrinsic Motivation: A Test of Cognitive Evaluation Theory," *Journal of Sport Psychology*, Vol. 6, pp. 94–102.

Vansteenkiste, M., and Deci, E. L. (2003), "Competitively Contingent Rewards and Intrinsic Motivation: Can Losers Remain Motivated?" *Motivation and Emotion*, Vol. 27, No. 4, pp. 273–299.

Volpert, W. (1987), "Psychische Regulation von Arbeitstätigkeiten," in *Arbeitspsychologie: Enzyklopädie der Psychologie*, Vol. D/III/1, U. Kleinbeck and J. Rutenfranz, Eds., Hogrefe Verlag, Göttingen, Germany, pp. 1–42.

von der Weth, R. (1988), "Konstruktionstätigkeit und Problemlösen," in *Rechnerunterstützte Konstruktion*, E. Frieling, and H. Klein, Eds., Hans Huber, Bern, Switzerland, pp. 32–39.

Vroom, V. H. (1964), *Work and Motivation*, Wiley, New York.

Warr, P. (1984), "Work and Unemployment," in *Handbook of Work and Organizational Psychology*, P. J. D.Drenth, H.Thierry, P. J.Williams, and C. J. de Wolff, Eds., Wiley, Chichester, West Sussex, England, pp. 413–443.

Weber, M. (1904/1905), "Die protestantische Ethik und der Geist des Kapitalismus," *Archiv für Sozialwissenschaft und Sozialpolitik*, Vol. 20, pp. 1–54, and Vol. 21, pp. 1–110.

Weber, W., and Ulich, E. (1993), "Psychological Criteria for the Evaluation of Different Forms of Group-Work in Advanced Manufacturing Systems," in *Human–Computer Interaction: Application and Case Studies*, M. J. Smith and G. Salvendy, Eds., Elsevier, Amsterdam, pp. 26–31.

Weiner, B. (1985), *Human Motivation*, Springer-Verlag, New York.

Weiner, B. (1992), *Human Motivation: Metaphors, Theories, and Research*, Sage Publications, Newbury Park, CA.

Wiendieck, G. (1986a), "Warum Qualitätszirkel? Zum organisationspsychologischen Hintergrund eines neuen Management-Konzeptes," in *Qualitätszirkel als Instrument zeitgemäßer Betriebsführung*, G. Wiendieck and W. Bungard, Eds., Verlag Moderne Industrie, Landsberg on Lech, Germany, pp. 61–75.

Wiendieck, G. (1986b), "Widerstand gegen Qualitätszirkel: Eine Idee und Ihre Feinde," in *Qualitätszirkel als Instrument zeitgemäßer Betriebsführung*, G. Wiendieck and W. Bungard, Eds., Verlag Moderne Industrie, Landsberg on Lech, Germany, pp. 207–223.

Williams, E. S., and Bunker, D. R. (1993), "Sorting Outcomes: A Revision of the JCM," in *The Job Characteristics Model: Recent Work and New Directions*, D. R. Bunker and E. D. Williams, Eds., Academy of Management, Atlanta, GA.

Wilpert, B., and Sorge, A. (1984), *International Perspectives in Organizational Democracy*, Wiley, Chichester, West Sussex, England.

Wilson, A. T. M., and Trist, E. L. (1951), *The Bolsover System of Continuous Mining*, Document 290, Tavistock Institute of Human Relations, London.

Windel, A., and Zimolong, B. (1998), "Nicht zum Nulltarif: Erfolgskonzept Gruppenarbeit," *Rubin*, Vol. 2, pp. 46–51.

Wiswede, G. (1980), *Motivation und Arbeitsverhalten*, E. Reinhardt, Munich.

Wofford, J. C., Goodwin, V. L., and Premack, S. (1992), "Meta-Analysis of the Antecedents of Personal Goal Level and of the Antecedents and Consequences of Goal Commitment," *Journal of Management*, Vol. 18, No. 3, pp. 595–615.

Yukl, G. A. (1990), *Skills for Managers and Leaders*, Prentice Hall, Englewood Cliffs, NJ.

Zink, K. (1978), "Zur Begründung einer zielgruppenspezifischen Organisationsentwicklung," *Zeitschrift für Arbeitswissenschaft*, Vol. 32, pp. 42–48.

Zink, K. J. (1979), *Begründung einer zielgruppenspezifischen Organisationsentwicklung auf der Basis von Untersuchungen von Arbeitszufriedenheit und Arbeitsmotivation*, O. Schmidt, Köln, Germany.

Zülch, G., and Starringer, M. (1984), "Differentielle Arbeitsgestaltung in Fertigungen für elektronische Flachbaugruppen," *Zeitschrift für Arbeitswissenschaft*, Vol. 38, pp. 211–216.

# CHAPTER 15

# JOB AND TEAM DESIGN

**Frederick P. Morgeson**
**The Eli Broad Graduate School of Management**
**Michigan State University**
**East Lansing, Michigan**


**Michael A. Campion**
**Krannert School of Management**
**Purdue University**
**West Lafayette, Indiana**


**Patrick F. Bruning**
**Krannert School of Management**
**Purdue University**
**West Lafayette, Indiana**

## 1 INTRODUCTION

### 1.1 Job Design

Job design is an aspect of managing organizations that is so commonplace it often goes unnoticed. Most people realize the importance of job design when an organization or new plant is starting up, and some recognize the importance of job design when organizations are restructuring or changing processes. But fewer people realize that job design may be affected as organizations change markets or strategies, managers use their discretion in the assignment of tasks on a daily basis, people in the jobs or their managers change, the workforce or labor markets change, or there are performance, safety, or satisfaction problems. Fewer yet realize that job design change can be used as an intervention to enhance organizational goals (Campion and Medsker, 1992).

It is clear that many different aspects of an organization influence job design, especially an organization's structure, technology, processes, and environment. These influences are beyond the scope of this chapter, but they are dealt with in other references (e.g., Davis, 1982; Davis and Wacker, 1982). These influences impose constraints on how jobs are designed and will play a major role in any practical application. However, it is the assumption of this chapter that considerable discretion exists in the design of jobs in most situations, and the job (defined as a set of tasks performed by a worker)

is a convenient unit of analysis in both developing new organizations or changing existing ones (Campion and Medsker, 1992).

The importance of job design lies in its strong influence on a broad range of important efficiency and human resource outcomes. Job design has predictable consequences for outcomes including the following (Campion and Medsker, 1992):

1. Productivity
2. Quality
3. Job satisfaction
4. Training times
5. Intrinsic work motivation
6. Staffing
7. Error rates
8. Accident rates
9. Mental fatigue
10. Physical fatigue
11. Stress
12. Mental ability requirements
13. Physical ability requirements
14. Job involvement
15. Absenteeism
16. Medical incidents
17. Turnover
18. Compensation rates

According to Louis Davis, one of the most prolific writers on job design in the engineering literature over the last 35 years, many of the personnel and productivity problems in industry may be the direct result of the design of jobs (Davis, 1957; Davis et al., 1955; Davis and Taylor, 1979; Davis and Valfer, 1965; Davis and Wacker, 1982, 1987). Unfortunately, people mistakenly view the design of jobs as technologically determined and inalterable. However, job designs are actually social inventions. They reflect the values of the era in which they were constructed. These values include the economic goal of minimizing immediate costs (Davis et al., 1955; Taylor, 1979) and theories of human motivation (Steers and Mowday, 1977; Warr and Wall, 1975). These values, and the designs they influence, are not immutable givens but are subject to modification (Campion and Medsker, 1992; Campion and Thayer, 1985).

The question then becomes: What is the best way to design a job? In fact, there is no single best way. There are several major approaches to job design, each derived from a different discipline and reflecting different theoretical orientations and values. This chapter describes these approaches, their costs and benefits, and tools and procedures for developing and assessing jobs in all types of organizations. It highlights trade-offs which must be made when choosing among different approaches to job design. This chapter also compares the design of jobs for individuals working independently to the design of work for teams, which is an alternative to designing jobs at the level of individual workers. This chapter presents the advantages and disadvantages of designing work around individuals compared to designing work for teams and provides advice on implementing and evaluating the different work design approaches.

## 1.2  Team Design

The major approaches to job design typically focus on designing jobs for individual workers. However, the approach to work design at the level of the group or team, rather than at the level of individual workers, is gaining substantially in popularity, and many U.S. organizations are experimenting with teams (Guzzo and Shea, 1992; Hoerr, 1989). New manufacturing systems (e.g., flexible, cellular) and advancements in our understanding of team processes not only allow designers to consider the use of work teams but often seem to encourage the use of team approaches (Gallagher and Knight, 1986; Majchrzak, 1988).

In designing jobs for teams, one assigns a task or set of tasks to a team of workers, rather than to an individual, and considers the team to be the primary unit of performance. Objectives and rewards focus on team, not individual, behavior. Depending on the nature of its tasks, a team's workers may be performing the same tasks simultaneously or they may break tasks into subtasks to be performed by individuals within the team. Subtasks can be assigned on the basis of expertise or interest, or team members might rotate from one subtask to another to provide variety and increase breadth of skills and flexibility in the workforce (Campion and Medsker, 1992; Campion et al., 1994b).

Some tasks are of a size or complexity or otherwise seem to naturally fit into a team job design, whereas others may seem to be appropriate only at the individual job level. In many cases, though, there may be a considerable degree of choice regarding whether one organizes work around teams or individuals. In such situations, the designer should consider advantages and disadvantages of the use of the job and team design approaches with respect to an organization's goals, policies, technologies, and constraints (Campion et al., 1993).

## 2  JOB DESIGN APPROACHES

This chapter adopts an interdisciplinary perspective on job design. Interdisciplinary research on job design has shown that different approaches to job design exist. Each is oriented toward a particular subset of outcomes, each has disadvantages as well as advantages, and trade-offs among approaches are required in most job design situations (Campion, 1988, 1989; Campion and Berger, 1990; Campion and McClelland, 1991, 1993; Campion and Thayer, 1985; Edwards et al., 1999, 2000; Morgeson and Campion, 2002, 2003).

While not new, contemporary work design researchers and practitioners have begun to reintegrate social and contextual aspects of employees' work with the characteristics traditionally studied by job design. These approaches to work design have since led to new approaches and have become incorporated into new assessment tools (Morgeson and Humphrey, 2006; Morgeson et al., 2010; Humphrey et al., 2007; Grant and

Parker, 2009). Building off and integrating the suggestions made in Campion's (1988; Campion and Thayer, 1985) interdisciplinary model of job design (MJDQ), the work design questionnaire (WDQ) represents a new tool with which to assess work design (Morgeson and Humphrey, 2006). This measure broadens the scope, discussion, and measurement of job design through the use of three broad categories of work characteristics (motivational, social, and work context). The WDQ assesses the job and its link to the worker's social and physical context and allows job designers to assess important yet infrequently studied aspects of work design such as knowledge/ability characteristics and social characteristics. A key difference between the MJDQ and WDQ is the perspective from which the job design is assessed. In the original MJDQ each perspective (mechanistic, motivational, perceptual–motor, and biological) is proposed to assess a different set of design principles intended to create different outcomes and thus appeal to a different set of stakeholders (Campion and Thayer, 1985). On the other hand, the WDQ aims to include the concerns of each approach captured in the MJDQ (with more emphasis on the motivational, perceptual–motor, and biological approaches than the mechanistic approach), along with social and contextual concerns in a single approach to designing better work. Based on a framework developed by Morgeson and Campion (2003), the authors used three categories to integrate aspects of work design (motivational, social, and contextual). The four major approaches to job design are reviewed below with a discussion of the applicability of the WDQ characteristics included. Table 1 summarizes the job design approaches and Tables 2 and 3 provide specific recommendations according to the MJDQ (Table 2) and the WDQ (Table 3). The team design approach is reviewed in Section 3.

## 2.1 Mechanistic Job Design Approach

### 2.1.1 Historical Development

The historical roots of job design can be traced back to the idea of the division of labor, which was very important to early thinking on the economies of manufacturing (Babbage, 1835; Smith, 1776). Division of labor led to job designs characterized by specialization and simplification. Jobs designed in this fashion had many advantages, including reduced learning time, saved time from not having to change tasks or tools, increased proficiency from repeating tasks, and development of specialized tools and equipment.

A very influential person for this perspective was Frederick Taylor (Hammond, 1971; Taylor, 1911). He explicated the principles of scientific management, which encouraged the study of jobs to determine the "one best way" to perform each task. Movements of skilled workmen were studied using a stopwatch and simple analysis. The best and quickest methods and tools were selected, and all workers were trained to perform the job the same way. Standard performance levels were set, and incentive pay was tied to the standards. Gilbreth (1911) also contributed to this design approach. With time-and-motion study, he tried to eliminate wasted

movements by the appropriate design of equipment and placement of tools and materials.

Surveys of industrial job designers indicate that this "mechanistic" approach to job design has been the prevailing practice throughout this century (Davis et al., 1955; Taylor, 1979). These characteristics are also the primary focus of many modern-day writers on job design (e.g., Mundel, 1985; Niebel, 1988) and are present in such newer techniques as lean production (Parker, 2003). The discipline base for this approach is early or "classic" industrial engineering.

### 2.1.2 Design Recommendations

Table 2 provides a brief list of statements that describe the essential recommendations of the mechanistic approach. In essence, jobs should be studied to determine the most efficient work methods and techniques. The total work in an area (e.g., department) should be broken down into highly specialized jobs assigned to different employees. The tasks should be simplified so skill requirements are minimized. There should also be repetition in order to gain improvement from practice. Idle time should be minimized. Finally, activities should be automated or assisted by automation to the extent possible and economically feasible.

### 2.1.3 Advantages and Disadvantages

The goal of this approach is to maximize efficiency, in terms of both productivity and utilization of human resources. Table 1 summarizes some human resource advantages and disadvantages that have been observed in research. Jobs designed according to the mechanistic approach are easier and less expensive to staff. Training times are reduced. Compensation requirements may be less because skill and responsibility are reduced. And because mental demands are less, errors may be less common. Disadvantages include the fact that extreme use of the mechanistic approach may result in jobs so simple and routine that employees experience low job satisfaction and motivation. Overly mechanistic, repetitive work can lead to health problems such as repetitive-motion disorders.

## 2.2 Motivational Job Design Approach

### 2.2.1 Historical Development

Encouraged by the human relations movement of the 1930s (Hoppock, 1935; Mayo, 1933), people began to point out the negative effects of the overuse of mechanistic design on worker attitudes and health (Argyris, 1964; Blauner, 1964). Overly specialized, simplified jobs were found to lead to dissatisfaction (Caplan et al., 1975) and adverse physiological consequences for workers (Johansson et al., 1978; Weber et al., 1980). Jobs on assembly lines and other machine-paced work were especially troublesome in this regard (Salvendy and Smith, 1981; Walker and Guest, 1952). These trends led to an increasing awareness of employees' psychological needs.

The first efforts to enhance the meaningfulness of jobs involved the opposite of specialization. It was recommended that tasks be added to jobs, either at the

**Table 1 Advantages and Disadvantages of Various Job Design Approaches**

| Approach/Discipline Base References) | Recommendations | Benefits | Costs |
|---|---|---|---|
| **Mechanistic/classic industrial engineering** (Gilbreth, 1911; Taylor, 1911; Niebel, 1988) | Increase in:<br>• Specialization<br>• Simplification<br>• Repetition<br>• Automation<br>Decrease in:<br>• Spare time | Decrease in:<br>• Training<br>• Staffing difficulty<br>• Making errors<br>• Mental overload and fatigue<br>• Mental skills and abilities<br>• Compensation | Increase in:<br>• Absenteeism<br>• Boredom<br>Decrease in:<br>• Satisfaction<br>• Motivation |
| **Motivational/organizational psychology** (Hackman and Oldham, 1980; Herzberg, 1966) | Increase in:<br>• Variety<br>• Autonomy<br>• Significance<br>• Skill usage<br>• Participation<br>• Feedback<br>• Recognition<br>• Growth<br>• Achievement | Increase in:<br>• Satisfaction<br>• Motivation<br>• Involvement<br>• Performance<br>• Customer service<br>• Catching errors<br>Decrease in:<br>• Absenteeism<br>• Turnover | Increase in:<br>• Training time/cost<br>• Staffing difficulty<br>• Making errors<br>• Mental overload<br>• Stress<br>• Mental skills and abilities<br>• Compensation |
| **Perceptual-motor/experimental psychology, human factors** (Salvendy, 1987; Sanders and McCormick, 1987) | Increase in:<br>• Lighting quality<br>• Display and control quality<br>• User-friendly equipment<br>Decrease in:<br>• Information-processing requirements | Decrease in:<br>• Making errors<br>• Accidents<br>• Mental overload<br>• Stress<br>• Training time/cost<br>• Staffing difficulty<br>• Compensation<br>• Mental skills and abilities | Increase in:<br>• Boredom<br>Decrease in:<br>• Satisfaction |
| **Biological/physiology, biomechanics, ergonomics** (Astrand and Rodahl, 1977; Tichauer, 1978; Grandjean, 1980) | Increase in:<br>• Seating comfort<br>• Postural comfort<br>Decrease in:<br>• Strength requirements<br>• Endurance requirements<br>• Environmental stressors | Decrease in:<br>• Physical abilities<br>• Physical fatigue<br>• Aches and pains<br>• Medical incidents | Increase in:<br>• Financial cost<br>• Inactivity |

*Source*: Adapted from Campion and Medsker (1992).
*Note:* Advantages and disadvantages based on findings in previous interdisciplinary research (Campion, 1988, 1989; Campion and Berger, 1990; Campion and McClelland, 1991, 1993; Campion and Thayer, 1985). Table adapted from Campion and Medsker (1992).

same level of responsibility (i.e., job enlargement) or at a higher level (i.e., job enrichment) (Ford, 1969; Herzberg, 1966). This trend expanded into a pursuit of identifying and validating characteristics of jobs that make them motivating and satisfying (Griffin, 1982; Hackman and Oldham, 1980; Turner and Lawrence, 1965). This approach considers the psychological theories of work motivation (e.g., Steers and Mowday, 1977; Vroom, 1964); thus this "motivational" approach draws primarily from organizational psychology as a discipline base.

A related trend following later in time but somewhat comparable in content is the sociotechnical approach (Emory and Trist, 1960; Pasmore, 1988; Rousseau,

1977). It focuses not only on the work but also on the technology itself and the relationship of the environment to work and organizational design. Interest is less on the job and more on roles and systems. Keys to this approach are work system and job designs that fit their external environment and the joint optimization of both social and technical systems in the organization's internal environment. Though this approach differs somewhat in that consideration is also given to the technical system and external environment, it is similar in that it draws on the same psychological job characteristics that affect satisfaction and motivation. It suggests that as organizations' environments are becoming

**Table 2  Multimethod Job Design Questionnaire**

(Specific Recommendations from Each Job Design Approach)
*Instructions:* Indicate the extent to which each statement is descriptive of the job using the scale below. Circle answers to the right of each statement.
*Please use the following scale*:

| | |
|---|---|
| **(5)** | Strongly agree |
| **(4)** | Agree |
| **(3)** | Neither agree nor disagree |
| **(2)** | Disagree |
| **(1)** | Strongly disagree |
| **( )** | Leave blank if do not know or not applicable |

*Mechanistic Approach*

1. Job specialization: The job is highly specialized in terms of purpose, tasks, or activities.  1 2 3 4 5
2. Specialization of tools and procedures: The tools, procedures, materials, etc., used on this job are highly specialized in terms of purpose.  1 2 3 4 5
3. Task simplification: The tasks are simple and uncomplicated.  1 2 3 4 5
4. Single activities: The job requires you to do only one task or activity at a time.  1 2 3 4 5
5. Skill simplification: The job requires relatively little skill and training time.  1 2 3 4 5
6. Repetition: The job requires performing the same activity(s) repeatedly.  1 2 3 4 5
7. Spare time: There is very little spare time between activities on this job.  1 2 3 4 5
8. Automation: Many of the activities of this job are automated or assisted by automation.  1 2 3 4 5

*Motivational Approach*

9. Autonomy: The job allows freedom, independence, or discretion in work scheduling, sequence, methods, procedures, quality control, or other decision making.  1 2 3 4 5
10. Intrinsic job feedback: The work activities themselves provide direct and clear information as to the effectiveness (e.g., quality and quantity) of job performance.  1 2 3 4 5
11. Extrinsic job feedback: Other people in the organization, such as managers and co-workers, provide information as to the effectiveness (e.g., quality and quantity) of job performance.  1 2 3 4 5
12. Social interaction: The job provides for positive social interaction such as team work or co-worker assistance.  1 2 3 4 5
13. Task/goal clarity: The job duties, requirements, and goals are clear and specific.  1 2 3 4 5
14. Task variety: The job has a variety of duties, tasks, and activities.  1 2 3 4 5
15. Task identity: The job requires completion of a whole and identifiable piece of work. It gives you a chance to do an entire piece of work from beginning to end.  1 2 3 4 5
16. Ability/skill-level requirements: The job requires a high level of knowledge, skills, and abilities.  1 2 3 4 5
17. Ability/skill variety: The job requires a variety of knowledge, skills, and abilities.  1 2 3 4 5
18. Task significance: The job is significant and important compared with other jobs in the organization.  1 2 3 4 5
19. Growth/learning: The job allows opportunities for learning and growth in competence and proficiency.  1 2 3 4 5
20. Promotion: There are opportunities for advancement to higher level jobs.  1 2 3 4 5
21. Achievement: The job provides for feelings of achievement and task accomplishment.  1 2 3 4 5
22. Participation: The job allows participation in work-related decision making.  1 2 3 4 5
23. Communication: The job has access to relevant communication channels and information flows.  1 2 3 4 5
24. Pay adequacy: The pay on this job is adequate compared with the job requirements and with the pay in similar jobs.  1 2 3 4 5
25. Recognition: The job provides acknowledgment and recognition from others.  1 2 3 4 5
26. Job security: People on this job have high job security.  1 2 3 4 5

*Perceptual/Motor Approach*

27. Lighting: The lighting in the workplace is adequate and free from glare.  1 2 3 4 5
28. Displays: The displays, gauges, meters, and computerized equipment on this job are easy to read and understand.  1 2 3 4 5
29. Programs: The programs in the computerized equipment on this job are easy to learn and use.  1 2 3 4 5
30. Other equipment: The other equipment (all types) used on this job is easy to learn and use.  1 2 3 4 5

**Table 2  (*continued*)**

| | | | | | | |
|---|---|---|---|---|---|---|
| 31. | Printed job materials: The printed materials used on this job are easy to read and interpret. | 1 | 2 | 3 | 4 | 5 |
| 32. | Workplace layout: The workplace is laid out such that you can see and hear well to perform the job. | 1 | 2 | 3 | 4 | 5 |
| 33. | Information input requirements: The amount of information you must attend to in order to perform this job is fairly minimal. | 1 | 2 | 3 | 4 | 5 |
| 34. | Information output requirements: The amount of information you must output on this job, in terms of both action and communication, is fairly minimal. | 1 | 2 | 3 | 4 | 5 |
| 35. | Information-processing requirements: The amount of information you must process, in terms of thinking and problem solving, is fairly minimal. | 1 | 2 | 3 | 4 | 5 |
| 36. | Memory requirements: The amount of information you must remember on this job is fairly minimal. | 1 | 2 | 3 | 4 | 5 |
| 37. | Stress: There is relatively little stress on this job. | 1 | 2 | 3 | 4 | 5 |

*Biological Approach*

| | | | | | | |
|---|---|---|---|---|---|---|
| 38. | Strength: The job requires fairly little muscular strength. | 1 | 2 | 3 | 4 | 5 |
| 39. | Lifting: The job requires fairly little lifting and/or the lifting is of very light weights. | 1 | 2 | 3 | 4 | 5 |
| 40. | Endurance: The job requires fairly little muscular endurance. | 1 | 2 | 3 | 4 | 5 |
| 41. | Seating: The seating arrangements on the job are adequate (e.g., ample opportunities to sit, comfortable chairs, good postural support). | 1 | 2 | 3 | 4 | 5 |
| 42. | Size differences: The workplace allows for all size differences between people in terms of clearance, reach, eye height, leg room, etc. | 1 | 2 | 3 | 4 | 5 |
| 43. | Wrist movement: The job allows the wrists to remain straight without excessive movement. | 1 | 2 | 3 | 4 | 5 |
| 44. | Noise: The workplace is free from excessive noise. | 1 | 2 | 3 | 4 | 5 |
| 45. | Climate: The climate at the workplace is comfortable in terms of temperature and humidity, and it is free of excessive dust and fumes. | 1 | 2 | 3 | 4 | 5 |
| 46. | Work breaks: There is adequate time for work breaks given the demands of the job. | 1 | 2 | 3 | 4 | 5 |
| 47. | Shift work: The job does not require shift work or excessive overtime. | 1 | 2 | 3 | 4 | 5 |

*For jobs with little physical activity due to single work station add:*

| | | | | | | |
|---|---|---|---|---|---|---|
| 48. | Exercise opportunities: During the day, there are enough opportunities to get up from the work station and walk around. | 1 | 2 | 3 | 4 | 5 |
| 49. | Constraint: While at the work station, the worker is not constrained to a single position. | 1 | 2 | 3 | 4 | 5 |
| 50. | Furniture: At the work station, the worker can adjust or arrange the furniture to be comfortable (e.g., adequate legroom, foot rests if needed, proper keyboard or work surface height). | 1 | 2 | 3 | 4 | 5 |

*Source*: Table adapted from Campion (1988). See supporting references and related research (e.g., Campion and McClelland, 1991, 1993; Campion and Thayer, 1985) for reliability and validity information. Scores for each approach are calculated by averaging applicable items.

increasingly turbulent and complex, organizational and job design should involve greater flexibility, employee involvement, employee training, and decentralization of decision making and control, and a reduction in hierarchical structures and the formalization of procedures and relationships (Pasmore, 1988).

Surveys of industrial job designers have consistently indicated that the mechanistic approach represents the dominant theme of job design (Davis et al., 1955; Taylor, 1979). Other approaches to job design, such as the motivational approach, have not been given as much explicit consideration. This is not surprising because the surveys only included job designers trained in engineering-related disciplines, such as industrial engineering and systems analysis. It is not necessarily certain that other specialists or line managers would adopt the same philosophies, especially in recent times. Nevertheless, there is evidence that even fairly naive job designers (i.e., college students in management classes) also adopt the mechanistic approach in job

design simulations. That is, their strategies for grouping tasks were primarily based on such factors as activities, skills, equipment, procedures, or location. Even though the mechanistic approach may be the most natural and intuitive, this research has also revealed that people can be trained to apply all four approaches to job design (Campion and Stevens, 1991). The motivational characteristics of the WDQ are an extension of this motivational approach to job design. This set of characteristics is based on the idea that high levels of these characteristics make work more motivating, satisfying, and enriching. Subcategories of these characteristics include task characteristics (task variety, task significance, task identity, and feedback from the job) and knowledge characteristics (job complexity, information processing, problem solving, skill variety, and specialization).

Building off of the ideas presented in Morgeson and Humphrey's WDQ, scholars are beginning to examine the social aspects of work design and how they interact

**Table 3 Work Design Questionnaire**

(Specific Recommendations from Each Job Design Approach)
*Instructions:* Indicate the extent to which each statement is descriptive of the job using the scale below.
Circle answers to the right of each statement.
*Please use the following scale*:

| | |
|---|---|
| **(5)** | Strongly agree |
| **(4)** | Agree |
| **(3)** | Neither agree nor disagree |
| **(2)** | Disagree |
| **(1)** | Strongly disagree |
| **( )** | Leave blank if do not know or not applicable |

### Task Characteristics

*Autonomy/Work Scheduling Autonomy*

1. The job allows me to make my own decisions about how to schedule my work.    1 2 3 4 5
2. The job allows me to decide on the order in which things are done on the job.    1 2 3 4 5
3. The job allows me to plan how I do my work.    1 2 3 4 5

*Autonomy/Decision-Making Autonomy*

4. The job gives me a chance to use my personal initiative or judgment in carrying out the work.    1 2 3 4 5
5. The job allows me to make a lot of decisions on my own.    1 2 3 4 5
6. The job provides me with significant autonomy in making decisions.    1 2 3 4 5

*Autonomy/Work Methods Autonomy*

7. The job allows me to make decisions about what methods I use to complete my work.    1 2 3 4 5
8. The job gives me considerable opportunity for independence and freedom in how I do the work.    1 2 3 4 5
9. The job allows me to decide on my own how to go about doing my work.    1 2 3 4 5

*Task Variety*

10. The job involves a great deal of task variety.    1 2 3 4 5
11. The job involves doing a number of different things.    1 2 3 4 5
12. The job requires the performance of a wide range of tasks.    1 2 3 4 5
13. The job involves performing a variety of tasks.    1 2 3 4 5

*Task Significance*

14. The results of my work are likely to significantly affect the lives of other people.    1 2 3 4 5
15. The job itself is very significant and important in the broader scheme of things.    1 2 3 4 5
16. The job has a large impact on people outside the organization.    1 2 3 4 5
17. The work performed on the job has a significant impact on people outside the organization.    1 2 3 4 5

*Task Identity*

18. The job involves completing a piece of work that has an obvious beginning and end.    1 2 3 4 5
19. The job is arranged so that I can do an entire piece of work from beginning to end.    1 2 3 4 5
20. The job provides me the chance to completely finish the pieces of work I begin.    1 2 3 4 5
21. The job allows me to complete work I start.    1 2 3 4 5

*Feedback from Job*

22. The work activities themselves provide direct and clear information about the effectiveness (e.g., quality and quantity) of my job performance.    1 2 3 4 5
23. The job itself provides feedback on my performance.    1 2 3 4 5
24. The job itself provides me with information about my performance.    1 2 3 4 5

### Knowledge Characteristics

*Job Complexity*

25. The job requires that I only do one task or activity at a time (reverse scored).    1 2 3 4 5
26. The tasks on the job are simple and uncomplicated (reverse scored).    1 2 3 4 5
27. The job comprises relatively uncomplicated tasks (reverse scored).    1 2 3 4 5
28. The job involves performing relatively simple tasks (reverse scored).    1 2 3 4 5

**Table 3** (*continued*)

*Information Processing*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 29. | The job requires me to monitor a great deal of information. | 1 | 2 | 3 | 4 | 5 |
| 30. | The job requires that I engage in a large amount of thinking. | 1 | 2 | 3 | 4 | 5 |
| 31. | The job requires me to keep track of more than one thing at a time. | 1 | 2 | 3 | 4 | 5 |
| 32. | The job requires me to analyze a lot of information. | 1 | 2 | 3 | 4 | 5 |

*Problem Solving*

| | | | | | | |
|---|---|---|---|---|---|---|
| 33. | The job involves solving problems that have no obvious correct answer. | 1 | 2 | 3 | 4 | 5 |
| 34. | The job requires me to be creative. | 1 | 2 | 3 | 4 | 5 |
| 35. | The job often involves dealing with problems that I have not met before. | 1 | 2 | 3 | 4 | 5 |
| 36. | The job requires unique ideas or solutions to problems. | 1 | 2 | 3 | 4 | 5 |

*Skill Variety*

| | | | | | | |
|---|---|---|---|---|---|---|
| 37. | The job requires a variety of skills. | 1 | 2 | 3 | 4 | 5 |
| 38. | The job requires me to utilize a variety of different skills in order to complete the work. | 1 | 2 | 3 | 4 | 5 |
| 39. | The job requires me to use a number of complex or high-level skills. | 1 | 2 | 3 | 4 | 5 |
| 40. | The job requires the use of a number of skills. | 1 | 2 | 3 | 4 | 5 |

*Specialization*

| | | | | | | |
|---|---|---|---|---|---|---|
| 41. | The job is highly specialized in terms of purpose, tasks, or activities. | 1 | 2 | 3 | 4 | 5 |
| 42. | The tools, procedures, materials, and so forth used on this job are highly specialized in terms of purpose. | 1 | 2 | 3 | 4 | 5 |
| 43. | The job requires very specialized knowledge and skills. | 1 | 2 | 3 | 4 | 5 |
| 44. | The job requires a depth of knowledge and expertise. | 1 | 2 | 3 | 4 | 5 |

### Social Characteristics

*Social Support*

| | | | | | | |
|---|---|---|---|---|---|---|
| 45. | I have the opportunity to develop close friendships in my job. | 1 | 2 | 3 | 4 | 5 |
| 46. | I have the chance in my job to get to know other people. | 1 | 2 | 3 | 4 | 5 |
| 47. | I have the opportunity to meet with others in my work. | 1 | 2 | 3 | 4 | 5 |
| 48. | My supervisor is concerned about the welfare of the people that work for him/her. | 1 | 2 | 3 | 4 | 5 |
| 49. | People I work with take a personal interest in me. | 1 | 2 | 3 | 4 | 5 |
| 50. | People I work with are friendly. | 1 | 2 | 3 | 4 | 5 |

*Interdependence/Initiated Interdependence*

| | | | | | | |
|---|---|---|---|---|---|---|
| 51. | The job requires me to accomplish my job before others complete their jobs. | 1 | 2 | 3 | 4 | 5 |
| 52. | Other jobs depend directly on my job. | 1 | 2 | 3 | 4 | 5 |
| 53. | Unless my job gets done, other jobs cannot be completed. | 1 | 2 | 3 | 4 | 5 |

*Interdependence/Received Interdependence*

| | | | | | | |
|---|---|---|---|---|---|---|
| 54. | The job activities are greatly affected by the work of other people. | 1 | 2 | 3 | 4 | 5 |
| 55. | The job depends on the work of many different people for its completion. | 1 | 2 | 3 | 4 | 5 |
| 56. | My job cannot be done unless others do their work. | 1 | 2 | 3 | 4 | 5 |

*Interaction Outside Organization*

| | | | | | | |
|---|---|---|---|---|---|---|
| 57. | The job requires spending a great deal of time with people outside my organization. | 1 | 2 | 3 | 4 | 5 |
| 58. | The job involves interaction with people who are not members of my organization. | 1 | 2 | 3 | 4 | 5 |
| 59. | On the job, I frequently communicate with people who do not work for the same organization as I do. | 1 | 2 | 3 | 4 | 5 |
| 60. | The job involves a great deal of interaction with people outside my organization. | 1 | 2 | 3 | 4 | 5 |

*Feedback from Others*

| | | | | | | |
|---|---|---|---|---|---|---|
| 61. | I receive a great deal of information from my manager and co-workers about my job performance. | 1 | 2 | 3 | 4 | 5 |
| 62. | Other people in the organization, such as managers and co-workers, provide information about the effectiveness (e.g., quality and quantity) of my job performance. | 1 | 2 | 3 | 4 | 5 |
| 63. | I receive feedback on my performance from other people in my organization (such as my manager or co-workers). | 1 | 2 | 3 | 4 | 5 |

**Table 3** (*continued*)

| | | | | | |
|---|---|---|---|---|---|
| ***Work Context*** | | | | | |

*Ergonomics*

| | | | | | |
|---|---|---|---|---|---|
| 64. | The seating arrangements on the job are adequate (e.g., ample opportunities to sit, comfortable chairs, good postural support). | 1 | 2 | 3 | 4 | 5 |
| 65. | The workplace allows for all size differences between people in terms of clearance, reach, eye height, leg room, etc. | 1 | 2 | 3 | 4 | 5 |
| 66. | The job involves excessive reaching (reverse scored). | 1 | 2 | 3 | 4 | 5 |

*Physical Demands*

| | | | | | |
|---|---|---|---|---|---|
| 67. | The job requires a great deal of muscular endurance. | 1 | 2 | 3 | 4 | 5 |
| 68. | The job requires a great deal of muscular strength. | 1 | 2 | 3 | 4 | 5 |
| 69. | The job requires a lot of physical effort. | 1 | 2 | 3 | 4 | 5 |

*Work Conditions*

| | | | | | |
|---|---|---|---|---|---|
| 70. | The workplace is free from excessive noise. | 1 | 2 | 3 | 4 | 5 |
| 71. | The climate at the workplace is comfortable in terms of temperature and humidity. | 1 | 2 | 3 | 4 | 5 |
| 72. | The job has a low risk of accident. | 1 | 2 | 3 | 4 | 5 |
| 73. | The job takes place in an environment free from health hazards (e.g., chemicals, fumes). | 1 | 2 | 3 | 4 | 5 |
| 74. | The job occurs in a clean environment. | 1 | 2 | 3 | 4 | 5 |

*Equipment Use*

| | | | | | |
|---|---|---|---|---|---|
| 75. | The job involves the use of a variety of different equipment. | 1 | 2 | 3 | 4 | 5 |
| 76. | The job involves the use of complex equipment or technology. | 1 | 2 | 3 | 4 | 5 |
| 77. | A lot of time was required to learn the equipment used on the job. | 1 | 2 | 3 | 4 | 5 |

*Source*: Table adapted from Morgeson and Humphrey (2006). See supporting reference for reliability and validity information. Scores for each approach are calculated by averaging applicable items.

with the peoples' on the job experience (Humphrey et al., 2007; Grant and Parker, 2009; Grant, 2007, 2008). Social characteristics consider the broader social environment in which the work is done as a component of workers' job experience. Social characteristics include social support (broadly refers to the support employees receive from others at work), interdependence (the interconnections of the tasks, sequencing, and impact of an employee's job with the jobs of others), interaction outside the organization, and feedback from others. Some of these social characteristics were originally encompassed within the motivational approach to job design (e.g., interdependence and feedback from others). By separating the social work characteristics from the task and knowledge characteristics, the WDQ allows job designers to focus specifically on the design of the interpersonal aspects of the work. Managers often have to address these aspects of work design in a different manner than they do with task and knowledge aspects. Subsequent meta-analytic evidence suggests that social characteristics addressed in the WDQ explain incremental variance above and beyond that explained by motivational characteristics (Humphrey et al., 2007).

### 2.2.2 Design Recommendations

Table 2 provides a list of statements that describe recommendations for the motivational approach. It suggests a job should allow a worker autonomy to make decisions about how and when tasks are to be done. A worker should feel his or her work is important to the overall mission of the organization or department. This

is often done by allowing a worker to perform a larger unit of work or to perform an entire piece of work from beginning to end. Feedback on job performance should be given to workers from the task itself as well as from the supervisor and others. Workers should be able to use a variety of skills and to personally grow on the job. This approach also considers the social, or people/interaction, aspects of the job; jobs should have opportunities for participation, communication, and recognition. Finally, other human resource systems should contribute to the motivating atmosphere, such as adequate pay, promotion, and job security systems.

### 2.2.3 Advantages and Disadvantages

The goal of this approach is to enhance psychological meaningfulness of jobs, thus influencing a variety of attitudinal and behavioral outcomes. Table 1 summarizes some of the advantages and disadvantages found in research. Jobs designed according to the motivational approach have more satisfied, motivated, and involved employees who tend to have higher performance and lower absenteeism. Customer service may be improved, because employees take more pride in work and can catch their own errors by performing a larger part of the work. Social impact, social worth, and mere social contact have been shown to have a positive influence on workers' performance (Grant, 2008; Grant et al., 2007). In a field experiment with community recreation center lifeguards, Grant (2008) demonstrated that task significance operated through their perceptions of social impact and social worth to influence job dedication and helping behavior. As an answer to the rapidly

changing nature of work, researchers have begun to study work design characteristics that could stimulate employee proactivity. While some characteristics are already embedded within the current models of job design, the approach has led to a few additional characteristics that could prove beneficial in a dynamic work environment. Specifically, both ambiguity and accountability have been suggested to influence employees' proactive behaviors (Grant and Parker, 2009; Staw and Boettger, 1990). In terms of disadvantages, jobs too high on the motivational approach require more training, have greater skill and ability requirements for staffing, and may require higher compensation. Overly motivating jobs may also be so stimulating that workers become predisposed to mental overload, fatigue, errors, and occupational stress.

## 2.3 Perceptual/Motor Job Design Approach

### 2.3.1 Historical Development

This approach draws on a scientific discipline which goes by many names, including human factors, human factors engineering, human engineering, man–machine systems engineering, and engineering psychology. It developed from a number of other disciplines, primarily experimental psychology, but also industrial engineering (Meister, 1971). Within experimental psychology, job design recommendations draw heavily from knowledge of human skilled performance (Welford, 1976) and the analysis of humans as information processors (see Chapters 2.4). The main concern of this approach is efficient and safe utilization of humans in human–machine systems, with emphasis on selection, design, and arrangement of system components to take account of both human abilities and limitations (Pearson, 1971). It is more concerned with equipment than psychology and with human abilities than engineering.

This approach received public attention with the Three Mile Island incident where it was concluded that the control room operator job in the nuclear power plant may have placed too many demands on the operator in an emergency situation, thus predisposing errors of judgment (Campion and Thayer, 1987). Government regulations issued since then require nuclear plants to consider "human factors" in their design (U.S. Nuclear Regulatory Commission, 1981). The primary emphasis of this approach is on perceptual and motor abilities of people. (See Chapters 20–22 for more information on equipment design).

The contextual characteristics of the WDQ reflect the physical and environmental contexts within which work is performed. This was an aspect initially described in the MJDQ and subsequently elaborated upon in the WDQ. Many of the contextual characteristics of job design encompass the perceptual–motor and biological/physiological (described in Section 2.4) approaches to job design as addressed in the MJDQ. Contextual characteristics include ergonomics, physical demands, work conditions, and equipment use. The WDQ's discrimination between contextual characteristics and other forms of motivational characteristics (i.e., task, knowledge, and social characteristics) allows managers to

focus specifically on the aspects of work that can produce worker strain or hazardous working conditions while still assessing the motivating aspects of the work. Meta-analytic evidence suggests that the contextual work characteristics addressed in the WDQ explain incremental variance above and beyond that explained by the motivational characteristics (Humphrey et al., 2007).

### 2.3.2 Design Recommendations

Table 2 provides a list of statements describing important recommendations of the perceptual/motor approach. They refer to either the equipment or environment and to information-processing requirements. Their thrust is to consider mental abilities and limitations of humans such that the attention and concentration requirements of the job do not exceed the abilities of the least capable potential worker. Focus is on the limits of the least capable worker because this approach is concerned with the effectiveness of the total system, which is no better than its "weakest link." Jobs should be designed to limit the amount of information workers must pay attention to and remember. Lighting levels should be appropriate, displays and controls should be logical and clear, workplaces should be well laid out and safe, and equipment should be easy to use. (See Chapters 56–61 for more information on human factors applications.)

### 2.3.3 Advantages and Disadvantages

The goals of this approach are to enhance reliability, safety, and positive user reactions. Table 1 summarizes advantages and disadvantages found in research. Jobs designed according to the perceptual/motor approach have lower errors and accidents. Like the mechanistic approach, it reduces mental ability requirements of the job; thus employees may be less stressed and mentally fatigued. It may also create some efficiencies, such as reduced training time and staffing requirements. On the other hand, costs from the excessive use of the perceptual/motor approach can include low satisfaction, low motivation, and boredom due to inadequate mental stimulation. This problem is exacerbated by the fact that designs based on the least capable worker essentially lower a job's mental requirements.

## 2.4 Biological Job Design Approach

### 2.4.1 Historical Development

This approach and the perceptual/motor approach share a joint concern for proper person–machine fit. The major difference is that this approach is more oriented toward biological considerations and stems from such disciplines as work physiology (see Chapter 10), biomechanics (i.e., study of body movements, see Chapter 9), and anthropometry (i.e., study of body sizes, see Chapters 8 and 23). Although many specialists probably practice both approaches together, as is reflected in many texts in the area (Konz, 1983), a split does exist between Americans who are more psychologically oriented and use the title "human factors engineer" and Europeans who are more physiologically oriented and use the title "ergonomist" (Chapanis, 1970). Like the

perceptual–motor approach, the biological approach is concerned with the design of equipment and workplaces as well as the design of tasks (Grandjean, 1980).

### 2.4.2 Design Recommendations

Table 2 lists important recommendations from the biological approach. This approach tries to design jobs to reduce physical demands to avoid exceeding people's physical capabilities and limitations. Jobs should not require excessive strength and lifting, and, again, abilities of the least physically able potential worker set the maximum level. Chairs should be designed for good postural support. Excessive wrist movement should be reduced by redesigning tasks and equipment. Noise, temperature, and atmosphere should be controlled within reasonable limits. Proper work–rest schedules should be provided so employees can recuperate from the physical demands.

### 2.4.3 Advantages and Disadvantages

The goals of this approach are to maintain employees' comfort and physical well-being. Table 1 summarizes some advantages and disadvantages observed in research. Jobs designed according to this approach require less physical effort, result in less fatigue, and create fewer injuries and aches and pains than jobs low on this approach. Occupational illnesses, such as lower back pain and carpal tunnel syndrome, are fewer on jobs designed with this approach. There may be lower absenteeism and higher job satisfaction on jobs which are not physically arduous. However, a direct cost of this approach may be the expense of changes in equipment or job environments needed to implement the recommendations. At the extreme, costs may include jobs with so few physical demands that workers become drowsy or lethargic, thus reducing performance. Clearly, extremes of physical activity and inactivity should be avoided, and an optimal level of physical activity should be developed.

## 3 TEAM DESIGN APPROACH

### 3.1 Historical Development

An alternative to designing work around individual jobs is to design work for teams of workers. Teams can vary a great deal in how they are designed and can conceivably incorporate elements from any of the job design approaches discussed. However, the focus here is on the self-managing, autonomous type of team design approach, which is gaining considerable popularity in organizations and substantial research attention today (Guzzo and Shea, 1992; Hoerr, 1989; Ilgen et al., 2005; Sundstrom et al., 1990; Swezey and Salas, 1992; Campion et al., 1996; Parker, 2003; LePine et al., 2008). Autonomous work teams derive their conceptual basis from motivational job design and from sociotechnical systems theory, which in turn reflect social and organizational psychology and organizational behavior (Cummings, 1978; Davis, 1971; Davis and Valfer, 1965; Morgeson and Campion, 2003). The Hawthorne studies (Homans, 1950) and European experiments with autonomous work groups (Kelly, 1982; Pasmore

et al., 1982) called attention to the benefits of applying work teams in situations other than sports and military settings. Although enthusiasm for the use of teams had waned in the 1960s and 1970s due to research discovering some disadvantages of teams (Buys, 1978; Zander, 1979), the 1980s brought a resurgence of interest in the use of work teams and it has become an extremely popular work design in organizations today (Hackman, 2002; Hoerr, 1989; Ilgen et al., 2005; Sundstrom et al., 1990). This renewed interest may be due to the cost advantages of having fewer supervisors with self-managed teams or the apparent logic of the benefits of teamwork.

### 3.2 Design Recommendations

Teams can vary in the degree of authority and autonomy they have (Banker et al., 1996). For example, manager-led teams have responsibility only for the execution of their work. Management designs the work, designs the teams, and provides an organizational context for the teams. However, in autonomous work teams, or self-managing teams, team members design and monitor their own work and performance. They may also design their own team structure (e.g., delineating interrelationships among members) and composition (e.g., selecting members). In such self-designing teams, management is only responsible for the teams' organizational context (Hackman, 1987). Although team design could incorporate elements of either mechanistic or motivational approaches to design, narrow and simplistic mechanistically designed jobs would be less consistent with other suggested aspects of the team approach to design than motivationally designed jobs. Mechanistically designed jobs would not allow an organization to gain as much of the advantages from placing workers in teams.

Figure 1 and Table 4 provide important recommendations from the self-managing team design approach. Many of the advantages of work teams depend on how teams are designed and supported by their organization. According to the theory behind self-managing team design, decision making and responsibility should be pushed down to the team members (Hackman, 1987). If management is willing to follow this philosophy, teams can provide several additional advantages. By pushing decision making down to the team and requiring consensus, the organization will find greater acceptance, understanding, and ownership of decisions (Porter et al., 1987). The perceived autonomy resulting from making work decisions should be both satisfying and motivating. Thus, this approach tries to design teams so they have a high degree of self-management and all team members participate in decision making.

The team design approach also suggests that the set of tasks assigned to a team should provide a whole and meaningful piece of work (i.e., have task identity as in the motivational approach to job design). This allows team members to see how their work contributes to a whole product or process, which might not be possible with individuals working alone. This can give workers a better idea of the significance of their work and create greater identification with the finished product or service. If team workers rotate among a variety of subtasks and cross-train on different operations,

**Themes/characteristics**                    **Effectiveness criteria**

**Job design**
- Self-management
- Participation
- Task variety
- Task significance
- Task identity

**Interdependence**
- Task interdependence
- Goal interdependence
- Interdependent feedback and rewards

**Composition**
- Heterogeneity
- Flexibility
- Relative size
- Preference for teamwork

**Context**
- Training
- Managerial support
- Communication/cooperation between teams

**Process**
- Potency
- Social support
- Workload sharing
- Communication/cooperation within the team

Productivity

Satisfaction

Manager judgments

**Figure 1** Characteristics related to team effectiveness.

workers should also perceive greater variety in the work (Campion et al., 1994b).

Interdependent tasks, goals, feedback, and rewards should be provided to create feelings of team interdependence among members and focus on the team as the unit of performance, rather than on the individual. It is suggested that team members should be heterogeneous in terms of areas of expertise and background so their varied knowledge, skills, and abilities (KSAs) complement one another. Teams also need adequate training, managerial support, and organizational resources to carry out their tasks. Managers should encourage positive group processes including open communication and cooperation within and between work groups, supportiveness and sharing of the workload among team members, and development of positive team spirit and confidence in the team's ability to perform effectively.

### 3.3 Advantages and Disadvantages

Table 5 summarizes advantages and disadvantages of team design relative to individual job design. To begin with, teams designed so members have heterogeneity of KSAs can help team members learn by working with others who have different KSAs. Cross-training on different tasks can occur, and the workforce can become more flexible (Goodman et al., 1986). Teams with heterogeneous KSAs also allow for synergistic combinations of ideas and abilities not possible with individuals working alone, and such teams have generally shown higher performance, especially when task requirements are diverse (Goodman et al., 1986; Shaw, 1983).

Social support can be especially important when teams face difficult decisions and deal with difficult psychological aspects of tasks, such as in military

**Table 4 Team Design Measure**

*Instructions:* This questionnaire consists of statements about your team and how your team functions as a group. Please indicate the extent to which each statement describes your team by circling a number to the right of each statement.

*Please use the following scale*:

| | |
|---|---|
| **(5)** | Strongly agree |
| **(4)** | Agree |
| **(3)** | Neither agree nor disagree |
| **(2)** | Disagree |
| **(1)** | Strongly disagree |
| **( )** | Leave blank if do not know or not applicable |

*Self-Management*

1. The members of my team are responsible for determining the methods, procedures, and schedules with which the work gets done.    1 2 3 4 5
2. My team rather than my manager decides who does what tasks within the team.    1 2 3 4 5
3. Most work-related decisions are made by the members of my team rather than by my manager.    1 2 3 4 5

*Participation*

4. As a member of a team, I have a real say in how the team carries out its work.    1 2 3 4 5
5. Most members of my team get a chance to participate in decision making.    1 2 3 4 5
6. My team is designed to let everyone participate in decision making.    1 2 3 4 5

*Task Variety*

7. Most members of my team get a chance to learn the different tasks the team performs.    1 2 3 4 5
8. Most everyone on my team gets a chance to do the more interesting tasks.    1 2 3 4 5
9. Task assignments often change from day to day to meet the workload needs of the team.    1 2 3 4 5

*Task Significance (Importance)*

10. The work performed by my team is important to the customers in my area.    1 2 3 4 5
11. My team makes an important contribution to serving the company's customers.    1 2 3 4 5
12. My team helps me feel that my work is important to the company.    1 2 3 4 5

*Task Identity (Mission)*

13. The team concept allows all the work on a given product to be completed by the same set of people.    1 2 3 4 5
14. My team is responsible for all aspects of a product for its area.    1 2 3 4 5
15. My team is responsible for its own unique area or segment of the business.    1 2 3 4 5

*Task Interdependence (Interdependence)*

16. I cannot accomplish my tasks without information or materials from other members of my team.    1 2 3 4 5
17. Other members of my team depend on me for information or materials needed to perform their tasks.    1 2 3 4 5
18. Within my team, jobs performed by team members are related to one another.    1 2 3 4 5

*Goal Interdependence (Goals)*

19. My work goals come directly from the goals of my team.    1 2 3 4 5
20. My work activities on any given day are determined by my team's goals for that day.    1 2 3 4 5
21. I do very few activities on my job that are not related to the goals of my team.    1 2 3 4 5

*Interdependent Feedback and Rewards (Feedback and Rewards)*

22. Feedback about how well I am doing my job comes primarily from information about how well the entire team is doing.    1 2 3 4 5
23. My performance evaluation is strongly influenced by how well my team performs.    1 2 3 4 5
24. Many rewards from my job (pay, promotion, etc.) are determined in large part by my contributions as a team member.    1 2 3 4 5

*(continued overleaf)*

**Table 4  (*continued*)**

| | *Heterogeneity (Membership)* | | | | | |
|---|---|---|---|---|---|---|
| 25. | The members of my team vary widely in their areas of expertise. | 1 | 2 | 3 | 4 | 5 |
| 26. | The members of my team have a variety of different backgrounds and experiences. | 1 | 2 | 3 | 4 | 5 |
| 27. | The members of my team have skills and abilities that complement each other. | 1 | 2 | 3 | 4 | 5 |
| | *Flexibility (Member Flexibility)* | | | | | |
| 28. | Most members of my team know each other's jobs. | 1 | 2 | 3 | 4 | 5 |
| 29. | It is easy for the members of my team to fill in for one another. | 1 | 2 | 3 | 4 | 5 |
| 30. | My team is very flexible in terms of membership. | 1 | 2 | 3 | 4 | 5 |
| | *Relative Size (Size)* | | | | | |
| 31. | The number of people in my team is too small for the work to be accomplished. (Reverse scored) | 1 | 2 | 3 | 4 | 5 |
| | *Preference for Team Work (Team Work Preferences)* | | | | | |
| 32. | If given the choice, I would prefer to work as part of a team rather than work alone. | 1 | 2 | 3 | 4 | 5 |
| 33. | I find that working as a member of a team increases my ability to perform effectively. | 1 | 2 | 3 | 4 | 5 |
| 34. | I generally prefer to work as part of a team. | 1 | 2 | 3 | 4 | 5 |
| | *Training* | | | | | |
| 35. | The company provides adequate technical training for my team. | 1 | 2 | 3 | 4 | 5 |
| 36. | The company provides adequate quality and customer service training for my team. | 1 | 2 | 3 | 4 | 5 |
| 37. | The company provides adequate team skills training for my team (communication, organization, interpersonal, etc.). | 1 | 2 | 3 | 4 | 5 |
| | *Managerial Support* | | | | | |
| 38. | Higher management in the company supports the concept of teams. | 1 | 2 | 3 | 4 | 5 |
| 39. | My manager supports the concept of teams. | 1 | 2 | 3 | 4 | 5 |
| | *Communication/Cooperation between Work Groups* | | | | | |
| 40. | I frequently talk to other people in the company besides the people on my team. | 1 | 2 | 3 | 4 | 5 |
| 41. | There is little competition between my team and other teams in the company. | 1 | 2 | 3 | 4 | 5 |
| 42. | Teams in the company cooperate to get the work done. | 1 | 2 | 3 | 4 | 5 |
| | *Potency (Spirit)* | | | | | |
| 43. | Members of my team have great confidence that the team can perform effectively. | 1 | 2 | 3 | 4 | 5 |
| 44. | My team can take on nearly any task and complete it. | 1 | 2 | 3 | 4 | 5 |
| 45. | My team has a lot of team spirit. | 1 | 2 | 3 | 4 | 5 |
| | *Social Support* | | | | | |
| 46. | Being in my team gives me the opportunity to work in a team and provide support to other team members. | 1 | 2 | 3 | 4 | 5 |
| 47. | My team increases my opportunities for positive social interaction. | 1 | 2 | 3 | 4 | 5 |
| 48. | Members of my team help each other out at work when needed. | 1 | 2 | 3 | 4 | 5 |
| | *Workload Sharing (Sharing the Work)* | | | | | |
| 49. | Everyone on my team does their fair share of the work. | 1 | 2 | 3 | 4 | 5 |
| 50. | No one in my team depends on other team members to do the work for them. | 1 | 2 | 3 | 4 | 5 |
| 51. | Nearly all the members of my team contribute equally to the work. | 1 | 2 | 3 | 4 | 5 |
| | *Communication/Cooperation within the Work Group* | | | | | |
| 52. | Members of my team are very willing to share information with other team members about our work. | 1 | 2 | 3 | 4 | 5 |
| 53. | Teams enhance the communications among people working on the same product. | 1 | 2 | 3 | 4 | 5 |
| 54. | Members of my team cooperate to get the work done. | 1 | 2 | 3 | 4 | 5 |

*Source*: Table adapted from Campion et al. (1993). See reference and related research (Campion et al., 1996) for reliability and validity information. Scores for each team characteristic are calculated by averaging applicable items.

**Table 5 Advantages and Disadvantages of Work Teams**

| Advantages | Disadvantages |
|---|---|
| • Team members learn from one another | • Lack of compatibility of some individuals with team work |
| • Possibility of greater workforce flexibility with cross-training | • Additional need to select workers to fit team as well as job |
| • Opportunity for synergistic combinations of ideas and abilities | • Possibility some members will experience less motivating jobs |
| • New approaches to tasks may be discovered | • Possible incompatibility with cultural, organizational, or labor–management norms |
| • Social facilitation and arousal | |
| • Social support for difficult tasks and situations | • Increased competition and conflict between teams |
| • Increased communication and information exchange between team members | • More time consuming due to socializing, coordination losses, and need for consensus |
| • Greater cooperation among team members | • Inhibition of creativity and decision-making processes; possibility of groupthink |
| • Beneficial for interdependent work flows | |
| • Greater acceptance and understanding of decisions when team makes decisions | • Less powerful evaluation and rewards; social loafing or free riding may occur |
| • Greater autonomy, variety, identity, significance, and feedback possible for workers | • Less flexibility in cases of replacement, turnover, or transfer |
| • Commitment to the team may stimulate performance and attendance | |

*Source*: Adapted from Campion and Medsker (1992).

squads, medical teams, or police units (Campion and Medsker, 1992). In addition, the simple presence of others can be psychologically arousing. Research has shown that such arousal can have a positive effect on performance when the task is well learned (Zajonc, 1965) and when other team members are perceived as evaluating the performer (Harkins, 1987; Porter et al., 1987). With routine jobs, this arousal effect may counteract boredom and performance decrements (Cartwright, 1968).

Another advantage of teams is that they can increase information exchanged between members through proximity and shared tasks (McGrath, 1984). Increased cooperation and communication within teams can be particularly useful when workers' jobs are highly interrelated, such as when workers whose tasks come later in the process must depend on the performance of workers whose tasks come earlier or when workers exchange work back and forth among themselves (Mintzberg, 1979; Thompson, 1967).

In addition, if teams are rewarded for team effort, rather than individual effort, members will have an incentive to cooperate with one another (Leventhal, 1976). The desire to maintain power by controlling information may be reduced. More experienced workers may be more willing to train the less experienced when they are not in competition with them. Team design and rewards can also be helpful in situations where it is difficult to measure individual performance or where workers mistrust supervisors' assessments of performance (Milkovich and Newman, 1993).

Finally, teams can be beneficial if team members develop a feeling of commitment and loyalty to their team (Cartwright, 1968). For workers who do not develop high commitment to their organization or management and who do not become highly involved in their job, work teams can provide a source of commitment. That is, members may feel responsible to attend work, cooperate with others, and perform well because of commitment to their work team, even though they are not strongly committed to the organization or the work itself.

Thus, designing work around teams can provide several advantages to organizations and their workers. Unfortunately, there are also disadvantages to using work teams and situations in which individual-level design is preferable to team design. For example, some individuals may dislike team work and may not have necessary interpersonal skills or desire to work in a team. When selecting team members, one has the additional requirement of selecting workers to fit the team as well as the job. (Section 4.3 provides more information on the selection of team members; see also Chapter 16 for general information on personnel selection.)

Individuals can experience less autonomy and less personal identification when working on a team. Designing work around teams does not guarantee workers greater variety, significance, and identity. If members within the team do not rotate among tasks or if some members are assigned exclusively to less desirable tasks, not all members will benefit from team design. Members can still have fractionated, demotivating jobs.

Team work can also be incompatible with cultural norms. The United States has a very individualistic culture (Hofstede, 1980). Applying team methods that have been successful in collectivistic societies like Japan may be problematic in the United States. In addition,

organizational norms and labor–management relations may be incompatible with team design, making its use more difficult.

Some advantages of team design can create disadvantages as well. First, though team rewards can increase communication and cooperation and reduce competition within a team, they may cause greater competition and reduced communication between teams. If members identify too strongly with a team, they may not realize when behaviors that benefit the team detract from organizational goals and create conflicts detrimental to productivity. Increased communication within teams may not always be task relevant either. Teams may spend work time socializing. Team decision making can take longer than individual decision making, and the need for coordination within teams can be time consuming.

Decision making and creativity can also be inhibited by team processes. When teams become highly cohesive, they may become so alike in their views that they develop "groupthink" (Janis, 1972). When groupthink occurs, teams tend to underestimate their competition, fail to adequately critique fellow team members' suggestions, not appraise alternatives adequately, and fail to work out contingency plans. In addition, team pressures distort judgments. Decisions may be based more on persuasiveness of dominant individuals or the power of majorities, rather than on the quality of decisions. Research has found a tendency for team judgments to be more extreme than the average of individual members' predecision judgments (Janis, 1972; McGrath, 1984; Morgeson and Campion, 1997). Although evidence shows highly cohesive teams are more satisfied with their teams, cohesiveness is not necessarily related to high productivity. Whether cohesiveness is related to performance depends on a team's norms and goals. If a team's norm is to be productive, cohesiveness will enhance productivity; however, if the norm is not one of commitment to productivity, cohesiveness can have a negative influence (Zajonc, 1965).

The use of teams and team-level rewards can also decrease the motivating power of evaluation and reward systems. If team members are not evaluated for individual performance, do not believe their output can be distinguished from the team's, or do not perceive a link between their personal performance and outcomes, social loafing (Harkins, 1987) can occur. In such situations, teams do not perform up to the potential expected from combining individual efforts.

Finally, teams may be less flexible in some respects because they are more difficult to move or transfer as a unit than individuals (Sundstrom et al., 1990). Turnover, replacements, and employee transfers may disrupt teams. And members may not readily accept new members.

Thus, whether work teams are advantageous depends to a great extent on the composition, structure, reward systems, environment, and task of the team. Table 6 presents questions that can help determine whether work should be designed around teams rather than individuals.

**Table 6  When to Design Jobs around Work Teams**

| | |
|---|---|
| 1. | Are workers' tasks highly interdependent or could they be made to be so? Would this interdependence enhance efficiency or quality? |
| 2. | Do the tasks require a variety of knowledge, skills, abilities such that combining individuals with different backgrounds would make a difference in performance? |
| 3. | Is cross-training desired? Would breadth of skills and workforce flexibility be essential to the organization? |
| 4. | Could increased arousal, motivation, and effort to perform make a difference in effectiveness? |
| 5. | Can social support help workers deal with job stresses? |
| 6. | Could increased communication and information exchange improve performance rather than interfere? |
| 7. | Could increased cooperation aid performance? |
| 8. | Are individual evaluation and rewards difficult or impossible to make or are they mistrusted by workers? |
| 9. | Could common measures of performance be developed and used? |
| 10. | Is it technically possible to group tasks in a meaningful, efficient way? |
| 11. | Would individuals be willing to work in teams? |
| 12. | Does the labor force have the interpersonal skills needed to work in teams? |
| 13. | Would team members have the capacity and willingness to be trained in interpersonal and technical skills required for team work? |
| 14. | Would team work be compatible with cultural norms, organizational policies, and leadership styles? |
| 15. | Would labor-management relations be favorable to team job design? |
| 16. | Would the amount of time taken to reach decisions, consensus, and coordination not be detrimental to performance? |
| 17. | Can turnover be kept to a minimum? |
| 18. | Can teams be defined as a meaningful unit of the organization with identifiable inputs, outputs, and buffer areas which give them a separate identity from other teams? |
| 19. | Would members share common resources, facilities, or equipment? |
| 20. | Would top management support team job design? |

*Source*: Table adapted from Campion and Medsker (1992). Affirmative answers support the use of team job design.

The more questions answered in the affirmative, the more likely teams are to be beneficial. If one chooses to design work around teams, suggestions for designing effective teams are presented in Section 4.3.

## 4   IMPLEMENTATION ADVICE FOR JOB AND TEAM DESIGN

### 4.1   General Implementation Advice

#### 4.1.1   Procedures to Follow

There are several general philosophies that are helpful when designing or redesigning jobs or teams:

1. As noted previously, designs are not inalterable or dictated by technology. There is some discretion in the design of all work situations and considerable discretion in most.
2. There is no single best design; there are simply better and worse designs depending on one's design perspective.
3. Design is iterative and evolutionary and should continue to change and improve over time.
4. Participation of workers affected generally improves the quality of the resulting design and acceptance of suggested changes.
5. The process of the project, or how it is conducted, is important in terms of involvement of all interested parties, consideration of alternative motivations, and awareness of territorial boundaries.

***Procedures for the Initial Design of Jobs or Teams***   In consideration of process aspects of design, Davis and Wacker (1982) suggest four steps:

1. *Form a steering committee.* This committee usually consists of a team of high-level executives who have a direct stake in the new jobs or teams. The purposes of the committee are to (a) bring into focus the project's objective, (b) provide resources and support for the project, (c) help gain cooperation of all parties affected, and (d) oversee and guide the project.
2. *Form a design task force.* The task force may include engineers, managers, job or team design experts, architects, specialists, and others with relevant knowledge or responsibility. The task force is to gather data, generate and evaluate design alternatives, and help implement recommended designs.
3. *Develop a philosophy statement.* The first goal of the task force is to develop a philosophy statement to guide decisions involved in the project. The philosophy statement is developed with input from the steering committee and may include the project's purposes, organization's strategic goals, assumptions about workers and the nature of work, and process considerations.
4. *Proceed in an evolutionary manner.* Jobs should not be overspecified. With considerable input from eventual job holders or team members, the work design will continue to change and improve over time.

According to Davis and Wacker (1982), the process of redesigning existing jobs is much the same as designing original jobs with two additions. First, existing job incumbents must be involved. Second, more attention needs to be given to implementation issues. Those involved in the implementation must feel ownership of and commitment to the change and believe the redesign represents their own interests.

***Potential Steps to Follow***   Along with the steps discussed above, a redesign project should also include the following five steps:

1. *Measuring Design of Existing job or Teams.* The questionnaire methodology and other analysis tools described in Section 5 may be used to measure current jobs or teams.
2. *Diagnosing Potential Design Problems.* Based on data collected in step 1, the current design is analyzed for potential problems. The task force and employee involvement are important. Focused team meetings are a useful vehicle for identifying and evaluating problems.
3. *Determining Job or Team Design Changes.* Changes will be guided by project goals, problems identified in step 2, and one or more of the approaches to work design. Often several potential changes are generated and evaluated. Evaluation of alternative changes may involve consideration of advantages and disadvantages identified in previous research (see Table 1) and opinions of engineers, managers, and employees.
4. *Making Design Changes.* Implementation plans should be developed in detail along with back-up plans in case there are difficulties with the new design. Communication and training are keys to implementation. Changes might also be pilot tested before widespread implementation.
5. *Conducting Follow-Up Evaluation.* Evaluating the new design after implementation is probably the most neglected part of the process in most applications. The evaluation might include the collection of design measurements on the redesigned jobs/teams using the same instruments as in step 1. Evaluation may also be conducted on outcomes, such as employee satisfaction, error rates, and training time (Table 1). Scientifically valid evaluations require experimental research strategies with control groups. Such studies may not always be possible in organizations, but often quasi-experimental and other field research designs are possible (Cook and Campbell, 1979). Finally, the need for adjustments is identified through the follow-up evaluation. (For examples of evaluations, see Section 5.8 and Campion and McClelland, 1991, 1993.)

### 4.1.2 Individual Differences Among Workers

It is a common observation that not all employees respond the same to the same job. Some people on a job have high satisfaction, whereas others on the same job have low satisfaction. Clearly, there are individual differences in how people respond to work. Considerable research has looked at individual differences in reaction to the motivational design approach. It has been found that some people respond more positively than others to highly motivational work. These differences are generally viewed as differences in needs for personal growth and development (Hackman and Oldham, 1980).

Using the broader notion of preferences/tolerances for types of work, the consideration of individual differences has been expanded to all four approaches to job design (Campion, 1988; Campion and McClelland, 1991) and to the team design approach (Campion et al., 1993; Campion et al., 1996). Table 7 provides scales that can be used to determine job incumbents' preferences/tolerances. These scales can be administered in the same manner as the questionnaire measures of job and team design discussed in Section 5.

Although consideration of individual differences is encouraged, there are often limits to which such differences can be accommodated. Jobs or teams may have

---

**Table 7  Preferences/Tolerances for the Design Approaches**

*Instructions:* Indicate the extent to which each statement is descriptive of your preferences and tolerances for types of work on the scale below. Circle answers to the right of each statement.
*Please use the following scale*:

| | |
|---|---|
| **(5)** | Strongly agree |
| **(4)** | Agree |
| **(3)** | Neither agree nor disagree |
| **(2)** | Disagree |
| **(1)** | Strongly disagree |
| **( )** | Leave blank if do not know or not applicable |

*Preferences/Tolerances for Mechanistic Design*

| | | | | | | |
|---|---|---|---|---|---|---|
| 1. | I have a high tolerance for routine work. | 1 | 2 | 3 | 4 | 5 |
| 2. | I prefer to work on one task at a time. | 1 | 2 | 3 | 4 | 5 |
| 3. | I have a high tolerance for repetitive work. | 1 | 2 | 3 | 4 | 5 |
| 4. | I prefer work that is easy to learn. | 1 | 2 | 3 | 4 | 5 |

*Preferences/Tolerances for Motivational Design*

| | | | | | | |
|---|---|---|---|---|---|---|
| 5. | I prefer highly challenging work that taxes my skills and abilities. | 1 | 2 | 3 | 4 | 5 |
| 6. | I have a high tolerance for mentally demanding work. | 1 | 2 | 3 | 4 | 5 |
| 7. | I prefer work that gives a great amount of feedback as to how I am doing. | 1 | 2 | 3 | 4 | 5 |
| 8. | I prefer work that regularly requires the learning of new skills. | 1 | 2 | 3 | 4 | 5 |
| 9. | I prefer work that requires me to develop my own methods, procedures, goals, and schedules. | 1 | 2 | 3 | 4 | 5 |
| 10. | I prefer work that has a great amount of variety in duties and responsibilities. | 1 | 2 | 3 | 4 | 5 |

*Preferences/Tolerances for Perceptual/Motor Design*

| | | | | | | |
|---|---|---|---|---|---|---|
| 11. | I prefer work that is very fast paced and stimulating. | 1 | 2 | 3 | 4 | 5 |
| 12. | I have a high tolerance for stressful work. | 1 | 2 | 3 | 4 | 5 |
| 13. | I have a high tolerance for complicated work. | 1 | 2 | 3 | 4 | 5 |
| 14. | I have a high tolerance for work where there are frequently too many things to do at one time. | 1 | 2 | 3 | 4 | 5 |

*Preferences/Tolerances for Biological Design*

| | | | | | | |
|---|---|---|---|---|---|---|
| 15. | I have a high tolerance for physically demanding work. | 1 | 2 | 3 | 4 | 5 |
| 16. | I have a fairly high tolerance for hot, noisy, or dirty work. | 1 | 2 | 3 | 4 | 5 |
| 17. | I prefer work that gives me some physical exercise. | 1 | 2 | 3 | 4 | 5 |
| 18. | I prefer work that gives me some opportunities to use my muscles. | 1 | 2 | 3 | 4 | 5 |

*Preferences/Tolerances for Team Work*

| | | | | | | |
|---|---|---|---|---|---|---|
| 19. | If given the choice, I would prefer to work as part of a team rather than work alone. | 1 | 2 | 3 | 4 | 5 |
| 20. | I find that working as a member of a team increases my ability to perform effectively. | 1 | 2 | 3 | 4 | 5 |
| 21. | I generally prefer to work as part of a team. | 1 | 2 | 3 | 4 | 5 |

*Source*: Table adapted from Campion (1988) and Campion et al. (1993).
*Note*: See reference for reliability and validity information. Scores for each preference/tolerance are calculated by averaging applicable items. Interpretations differ slightly across the scales. For the mechanistic and motivational designs, higher scores suggest more favorable reactions from incumbents to well-designed jobs. For the perceptual/motor and biological approaches, higher scores suggest less unfavorable reactions from incumbents to poorly designed jobs.

to be designed for people who are not yet known or who differ in their preferences. Fortunately, although evidence indicates individual differences moderate reactions to the motivational approach (Fried and Ferris, 1987), the differences are of degree but not direction. That is, some people respond more positively than others to motivational work, but few respond negatively. It is likely that this also applies to the other design approaches.

### 4.1.3 Some Basic Choices

Hackman and Oldham (1980) have provided five strategic choices that relate to implementing job redesign. The note that little research exists indicating the exact consequences of each choice, and correct choices may differ by organization. The basic choices are:

1. *Individual versus Team Designs for Work*. An initial decision is to either enrich individual jobs or create teams. This also includes consideration of whether any redesign should be undertaken and its likelihood of success.
2. *Theory Based versus Intuitive Changes*. This choice was basically defined as the motivational (theory) approach versus no particular (atheoretical) approach. In the present chapter, this choice may be better framed as choosing among the four approaches to job design. However, as argued earlier, consideration of only one approach may lead to some costs or additional benefits being ignored.
3. *Tailored versus Broadside Installation*. This choice is between tailoring changes to individuals and making the changes for all in a given job.
4. *Participative versus Top-Down Change Processes*. The most common orientation is that participative is best. However, costs of participation include the time involved and incumbents' possible lack of a broad knowledge of the business.
5. *Consultation versus Collaboration with Stakeholders*. The effects of job design changes often extend far beyond the individual incumbent and department. For example, a job's output may be an input to a job elsewhere in the organization. The presence of a union also requires additional collaboration. Depending on considerations, participation of stakeholders may range from no involvement through consultation to full collaboration.

### 4.1.4 Overcoming Resistance to Change in Redesign Projects

Resistance to change can be a problem in any project involving major changes (Morgeson et al., 1997). Failure rates of new technology implementations demonstrate a need to give more attention to the human aspects of change projects. This concern has also been reflected in the area of participatory ergonomics, which encourages the use of participatory techniques when undertaking an ergonomic intervention (Wilson and Haines, 1997). It has been estimated that between 50 and 75% of

newly implemented manufacturing technologies in the United States have failed, with a disregard for human and organizational issues considered to be a bigger cause for the failures than technical problems (Majchrzak, 1988; Turnage, 1990). The number one obstacle to implementation was considered to be human resistance to change (Hyer, 1984).

Based on the work of Majchrzak (1988), Gallagher and Knight (1986), and Turnage (1990), guidelines for reducing resistance to change include the following:

1. *Involve workers in planning the change*. Workers should be informed of changes in advance and involved in the process of diagnosing current problems and developing solutions. Resistance is decreased if participants feel the project is their own and not imposed from outside and if the project is adopted by consensus.
2. *Top management should strongly support the change*. If workers feel management is not strongly committed, they are less likely to take the project seriously.
3. *Create change consistent with worker needs and existing values*. Resistance is less if change is seen to reduce present burdens, offer interesting experience, and not threaten worker autonomy or security or be inconsistent with other goals and values in the organization. Workers need to see the advantages to them of the change. Resistance is less if proponents of change can empathize with opponents (recognize valid objections and relieve unnecessary fears).
4. *Create an environment of open, supportive communication*. Resistance will be lessened if participants experience support and have trust in each other. Resistance can be reduced if misunderstandings and conflicts are expected as natural to the innovation process. Provision should be made for clarification.
5. *Allow for flexibility*. Resistance is reduced if the project is kept open to revision and reconsideration with experience.

## 4.2 Implementation Advice for Job Design and Redesign

### 4.2.1 Methods for Combining Tasks

In many cases, designing jobs is largely a function of combining tasks. Some guidance can be gained by extrapolating from specific design recommendations in Table 2. For example, variety in the motivational approach can be increased by simply combining different tasks in the same job. Conversely, specialization from the mechanistic approach can be increased by only including very similar tasks in the same job. It is also possible when designing jobs to first generate alternative task combinations, then evaluate them using the design approaches in Table 2.

A small amount of research within the motivational approach has focused explicitly on predicting relationships between combinations of tasks and the design of resulting jobs (Wong, 1989; Wong and Campion, 1991). This research suggests that a job's motivational quality
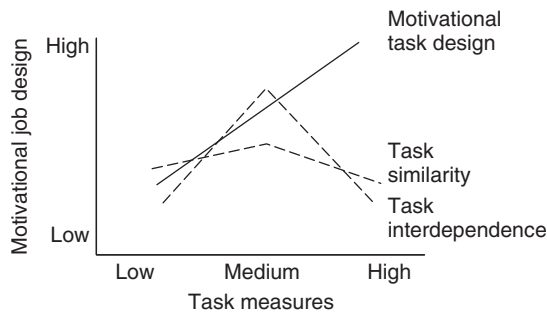
**Figure 2** Effects of task design, interdependence, and similarity on motivational job design.

is a function of three task-level variables, as illustrated in Figure 2.

1. *Task Design*. The higher the motivational quality of individual tasks, the higher the motivational quality of a job. Table 2 can be used to evaluate individual tasks; then motivational scores for individual tasks can be summed together. Summing is recommended rather than averaging because both the motivational quality of the tasks and the number of tasks are important in determining a job's motivational quality (Globerson and Crossman, 1976).

2. *Task Interdependence*. Interdependence among tasks has been shown to be positively related to motivational value up to some moderate point; beyond that point increasing interdependence has been shown to lead to lower motivational value. Thus, for motivational jobs, the total amount of interdependence among tasks should be kept at a moderate level. Both complete independence and excessively high interdependence should be avoided. Table 8 contains the dimension of task interdependence and provides a questionnaire to measure it. Table 8 can be used to judge the interdependence of each pair of tasks that are being evaluated for inclusion in a job.

3. *Task Similarity*. Similarity among tasks may be the oldest rule of job design, but beyond a moderate level, it tends to decrease a job's motivational value. Thus, to design motivational jobs, high levels of similarity should be avoided. Similarity at the task pair level can be judged in much the same manner as interdependence by using dimensions in Table 8 (see the note to Table 8).

### 4.2.2 Trade-Offs among Job Design Approaches

Although one should strive to construct jobs that are well designed on all the approaches, it is clear design approaches conflict. As Table 1 illustrates, benefits of some approaches are costs of others. No one approach satisfies all outcomes. The greatest potential conflicts are between the motivational and the mechanistic and perceptual/motor approaches. They produce nearly opposite outcomes. The mechanistic and perceptual/motor approaches recommend jobs that are simple, safe, and reliable, with minimal mental demands on workers. The motivational approach encourages more complicated and stimulating jobs, with greater mental demands. The team approach is consistent with the motivational approach and therefore also may conflict with the mechanistic and perceptual/motor approaches.

Because of these conflicts, trade-offs may be necessary. Major trade-offs will be in the mental demands created by the alternative design strategies. Making jobs more mentally demanding increases the likelihood of achieving workers' goals of satisfaction and motivation but decreases the chances of reaching the organization's goals of reduced training, staffing costs, and errors. Which trade-offs will be made depends on outcomes one prefers to maximize. Generally, a compromise may be optimal.

Trade-offs may not always be needed, however. Jobs can often be improved on one approach while still maintaining their quality on other approaches. For example, in one redesign study, the motivational approach was applied to clerical jobs to improve employee satisfaction and customer service (Campion and McClelland, 1991). Expected benefits occurred along with some expected costs (e.g., increased training and compensation requirements), but not all potential costs occurred (e.g., quality and efficiency did not decrease).

In another redesign study, Morgeson and Campion (2002) sought to increase both satisfaction and efficiency in jobs at a pharmaceutical company. They found that when jobs were designed to increase only satisfaction or only efficiency, the common trade-offs were present (e.g., increased or decreased satisfaction, training requirements). When jobs were designed to increase both satisfaction and efficiency, however, these trade-offs were reduced. They suggested that a work design process that explicitly considers both motivational and mechanistic aspects of work is key to avoiding the trade-offs.

Another strategy for minimizing trade-offs is to avoid design decisions that influence the mental demands of jobs. An example of this is to enhance motivational design by focusing on social aspects (e.g., communication, participation, recognition, feedback). These design features can be raised without incurring costs of increased mental demands. Moreover, many of these features are under the direct control of managers.

The independence of the biological approach provides another opportunity to improve design without incurring trade-offs with other approaches. One can reduce physical demands without affecting mental demands of a job. Of course, the cost of equipment may need to be considered.

Adverse effects of trade-offs can often be reduced by avoiding designs that are extremely high or low on any approach. Or, alternatively, one might require minimum acceptable levels on each approach. Knowing all approaches and their corresponding outcomes will

**Table 8 Dimensions of Task Interdependence**

*Instructions:* Indicate the extent to which each statement is descriptive of the pair of tasks using the scale below. Circle answers to the right of each statement. Scores are calcualted by averaging applicable items.
*Please use the following scale*:

| | |
|---|---|
| **(5)** | Strongly agree |
| **(4)** | Agree |
| **(3)** | Neither agree nor disagree |
| **(2)** | Disagree |
| **(1)** | Strongly disagree |
| **( )** | Leave blank if do not know or not applicable |

*Inputs of the Tasks*

1. Materials/supplies: One task obtains, stores, or prepares the materials or supplies necessary to perform the other task.  1 2 3 4 5
2. Information: One task obtains or generates information for the other task.  1 2 3 4 5
3. Product/service: One task stores, implements, or handles the products or services produced by the other task.  1 2 3 4 5

*Processes of the Tasks*

4. Input–output relationship: The products (or outputs) of one task are the supplies (or inputs) necessary to perform the other task.  1 2 3 4 5
5. Method and procedure: One task plans the procedures or work methods for the other task.  1 2 3 4 5
6. Scheduling: One task schedules the activities of the other task.  1 2 3 4 5
7. Supervision: One task reviews or checks the quality of products or services produced by the other task.  1 2 3 4 5
8. Sequencing: One task needs to be performed before the other task.  1 2 3 4 5
9. Time sharing: Some of the work activities of the two tasks must be performed at the same time.  1 2 3 4 5
10. Support service: The purpose of one task is to support or otherwise help the other task get performed.  1 2 3 4 5
11. Tools/equipment: One task produces or maintains the tools or equipment used by the other task.  1 2 3 4 5

*Outputs of the Tasks*

12. Goal: One task can only be accomplished when the other task is properly performed.  1 2 3 4 5
13. Performance: How well one task is performed has a great impact on how well the other task can be performed.  1 2 3 4 5
14. Quality: The quality of the product or service produced by one task depends on how well the other task is performed.  1 2 3 4 5

*Source*: Table adapted from Wong and Campion (1991). See reference and Wong (1989) for reliability and validity information.
*Note*: The task similarity measure contains 10 comparable items (excluding items 4, 6, 8, 9, and 14 and including an item on customer/client). Scores for each dimension are calculated by averaging applicable items.

help one make more informed decisions and avoid unanticipated consequences.

### 4.2.3 Other Implementation Advice for Job Design and Redesign

Griffin (1982) provides advice geared toward managers considering a job redesign intervention in their area. He notes that managers may also rely on consultants, task forces, or informal discussion groups. Griffin suggests nine steps:

1. Recognition of a need for change
2. Selection of job redesign as a potential intervention
3. Diagnosis of the work system and content on the following factors:
    a. Existing jobs
    b. Existing workforce
    c. Technology
    d. Organization design
    e. Leader behaviors
    f. Team and social processes
4. Cost/benefit analysis of proposed changes
5. Go/no-go decision
6. Establishment of a strategy for redesign
7. Implementation of the job changes
8. Implementation of any needed supplemental changes
9. Evaluation of the redesigned jobs

Recently, researchers have begun to study the manner in which employees are proactive actors in the job design process. Employees can either actively craft or change their jobs or they can negotiate idiosyncratic deals that alter the design of their work (Wrzesniewski and Dutton, 2001; Grant and Parker, 2009; Rousseau et al., 2006; Hornung et al., 2008). An idiosyncratic deal is a formal agreement that an employee and his or her manager or organization come to regarding the individual's work which creates a difference in the characteristics of the employee's work from the characteristics of the work of employees in a similar position. These types of deals represent formal individualized work design arrangements.

Job crafting differs from traditional job design as it describes the changes that employees make to their jobs. While traditional job design is implemented by a manager or an organization, job crafting refers to the informal changes to task or social characteristics that employees make to their work. The implications of these behaviors are that employees will informally change the design of their work. Being informal, they often go undetected and can be difficult for a manger to control. Managers should both recognize that these changes occur and design employees' work with the understanding that the design of the work can and probably will be altered to some degree by the employee.

## 4.3 Implementation Advice for Team Design

### 4.3.1 Deciding on Team Composition

Research encourages heterogeneous teams in terms of skills, personality, and attitudes because it increases the range of competencies in teams (Gladstein, 1984) and is related to effectiveness (Campion et al., 1996). However, homogeneity is preferred if team morale is the main criterion, and heterogeneous attributes must be complementary if they are to contribute to effectiveness. Heterogeneity for its own sake is unlikely to enhance effectiveness (Campion et al., 1993). Another composition characteristic of effective teams is whether members have flexible job assignments (Campion et al., 1993; Sundstrom et al., 1990). If members can perform different jobs, effectiveness is enhanced because they can fill in as needed.

A third important aspect of composition is team size. Evidence suggests the importance of optimally matching team size to team tasks to achieve high performance and satisfaction. Teams need to be large enough to accomplish work assigned to them but may be dysfunctional when too large due to heightened coordination needs (O'Reilly and Roberts, 1977; Steiner, 1972) or increased social loafing (McGrath, 1984; Wicker et al., 1976). Thus, groups should be staffed to the smallest number needed to do the work (Goodman et al., 1986; Hackman, 1987; Sundstrom et al., 1990).

### 4.3.2 Selecting Team Members

With team design, interpersonal demands appear to be much greater than with traditional individual-based job design (Lawler, 1986). A team-based setting highlights the importance of employees being capable of interacting in an effective manner with peers because the amount of interpersonal interactions required is higher in teams (Stevens and Campion, 1994a,b, 1999). Team effectiveness can depend heavily on members' "interpersonal competence" or their ability to successfully maintain healthy working relationships and react to others with respect for their viewpoints (Perkins and Abramis, 1990). There is a greater need for team members to be capable of effective interpersonal communication, collaborative problem solving, and conflict management (Stevens and Campion, 1994a,b, 1999).

The process of employment selection for team members places greater stress on adequately evaluating interpersonal competence than is normally required in the selection of workers for individual jobs. To create a selection instrument for evaluating potential team members' ability to work successfully in teams, Stevens and Campion (1994a,b) reviewed literature in areas of sociotechnical systems theory (e.g., Cummings, 1978; Wall et al., 1986), organizational behavior (e.g., Hackman, 1987; Shea and Guzzo, 1987; Sundstrom et al., 1990), industrial engineering (e.g., Davis and Wacker, 1987; Majchrzak, 1988), and social psychology (e.g., McGrath, 1984; Steiner, 1972) to identify relevant KSAs. Table 9 shows the 14 KSAs identified as important for teamwork.

These KSAs have been used to develop a 35-item, multiple-choice employment test which was validated in two studies to determine how highly related it was to team members' job performance (Stevens and Campion, 1999). The job performance of team members in two different companies was rated by both supervisors and co-workers. Correlations between the test and job performance ratings were significantly high, with some correlations exceeding 0.50. The test was also able to add to the ability to predict job performance beyond that provided by a large battery of traditional employment aptitude tests. Thus, these findings provide support for the value of the teamwork KSAs and a selection test based on them (Stevens and Campion, 1994a). Table 10 shows some example items from the test.

Aside from written tests, there may be other ways teamwork KSAs could be measured for purposes of selection. For example, interviews may be especially suited to measuring interpersonal attributes (e.g., Posthuma et al., 2002). There is evidence that a structured interview specifically designed to measure social (i.e., nontechnical) KSAs can have validity with job performance and predict incrementally beyond traditional employment tests (Campion et al., 1994a).

Assessment center techniques might also lend themselves to measuring teamwork KSAs. Group exercises have been used to measure leadership and other social skills with good success (Gaugler et al., 1987). It is likely that existing team exercises, such as group problem-solving tasks, could also be modified to score teamwork KSAs.

Selection techniques using biodata may be another way to measure teamwork KSAs. Many items in biodata instruments reflect previous life experiences of a social nature, and recruiters interpret biodata information on applications and resumes as reflecting attributes such as interpersonal skills (Brown and Campion, 1994).

**Table 9  Knowledge, Skill, and Ability (KSA) Requirements for Teamwork**

I.   Interpersonal KSAs
 A.   Conflict Resolution KSAs
  1.   The KSA to recognize and encourage desirable but discourage undesirable team conflict.
  2.   The KSA to recognize the type and source of conflict confronting the team and to implement an appropriate conflict resolution strategy.
  3.   The KSA to employ an integrative (win–win) negotiation strategy rather than the traditional distributive (win–lose) strategy.
 B.   Collaborative Problem-Solving KSAs
  4.   The KSA to identify situations requiring participative group problem solving and to utilize the proper degree and type of participation.
  5.   The KSA to recognize the obstacles to collaborative group problem solving and implement appropriate corrective actions.
 C.   Communication KSAs
  6.   The KSA to understand communication networks and to utilize decentralized networks to enhance communication where possible.
  7.   The KSA to communicate openly and supportively, that is, to send messages which are (a) behavior or event oriented, (b) congruent, (c) validating, (d) conjunctive, and (e) owned.
  8.   The KSA to listen nonevaluatively and to appropriately use active listing techniques.
  9.   The KSA to maximize consonance between nonverbal and verbal messages and to recognize and interpret the nonverbal messages of others.
  10.  The KSA to engage in ritual greetings and small talk and a recognition of their importance.
II.  Self-Management KSAs
 D.   Goal Setting and Performance Management KSAs
  11.  The KSA to help establish specific, challenging, and accepted team goals.
  12.  The KSA to monitor, evaluate, and provide feedback on both overall team performance and individual team member performance.
 E.   Planning and Task Coordination KSAs
  13.  The KSA to coordinate and synchronize activities, information, and task interdependencies between team members.
  14.  The KSA to help establish task and role expectations of individual team members and to ensure proper balancing of workload in the team.

**Table 10  Example Items from the Teamwork KSA Test**

1.  Suppose you find yourself in an argument with several co workers who should do a very disagreeable but routine task. Which of the following would likely be the most effective way to resolve this situation?
 A.   Have your supervisor decide, because this would avoid any personal bias.
 *B.  Arrange for a rotating schedule so everyone shares the chore.
 C.   Let the workers who show up earliest choose on a first come, first served basis.
 D.   Randomly assign a person to do the task and don't change it.
2.  Your team wants to improve the quality and flow of the conversations among its members. Your team should:
 *A.  Use comments that build upon and connect to what others have said.
 B.   Set up a specific order for everyone to speak and then follow it.
 C.   Let team members with more to say determine the direction and topic of conversation.
 D.   Do all of the above.
3.  Suppose you are presented with the following types of goals. You are asked to pick one for your team to work on. Which would you choose?
 A.   An easy goal to ensure the team reaches it, thus creating a feeling of success.
 B.   A goal of average difficulty so the team will be somewhat challenged but successful without too much effort.
 *C.  A difficult and challenging goal that will stretch the team to perform at a high level but one that is attainable so that effort will not be seen as futile.
 D.   A very difficult or even impossible goal so that even if the team falls short, it will at least have a very high target to aim for.

*Note*: asterisk indicates correct answer.

A biodata measure developed to focus on teamwork KSAs might include items on teamwork in previous jobs, team experiences in school (e.g., college clubs, class projects), and recreational activities of a team nature (e.g., sports teams and social groups).

### 4.3.3  Designing the Teams' Jobs

This aspect of team design involves team characteristics derived from the motivational job design approach. The main distinction is in level of application rather than content (Campion and Medsker, 1992; Shea and Guzzo, 1987; Wall et al., 1986). All the job characteristics of the motivational approach to job design can be applied to team design.

One such characteristic is self-management, which is the team-level analogy to autonomy at the individual job level. It is central to many definitions of effective work

teams (e.g., Cummings, 1978, 1981; Hackman, 1987). A related characteristic is participation. Regardless of management involvement in decision making, teams can still be distinguished in terms of the degree to which all members are allowed to participate in decisions (McGrath, 1984, Porter et al., 1987). Self-management and participation are presumed to enhance effectiveness by increasing members' sense of responsibility and ownership of the work. These characteristics may also enhance decision quality by increasing relevant information and by putting decisions as near as possible to the point of operational problems and uncertainties.

Other important characteristics are task variety, task significance, and task identity. Variety motivates by allowing members to use different skills (Hackman, 1987) and by allowing both interesting and dull tasks to be shared among members (Davis and Wacker, 1987). Task significance refers to the perceived significance of the consequences of the team's work, for either others inside the organization or its customers. Task identity (Hackman, 1987), or task differentiation (Cummings, 1978), refers to the degree to which the team completes a whole and meaningful piece of work. These suggested characteristics of team design have been found to be positively related to team productivity, team member satisfaction, and managers' and employees' judgments of their teams' performance (Campion et al., 1993, 1995).

### 4.3.4 Developing Interdependent Relations

Interdependence is often the reason teams are formed (Mintzberg, 1979) and is a defining characteristic of teams (Salas et al., 1992; Wall et al., 1986). Interdependence has been found to be related to team members' satisfaction and team productivity and effectiveness (Campion et al., 1993, 1995).

One form of interdependence is task interdependence. Team members interact and depend on one another to accomplish their work. Interdependence varies across teams, depending on whether the work flow in a team is pooled, sequential, or reciprocal (Thompson, 1967). Interdependence among tasks in the same job (Wong and Campion, 1991) or between jobs (Kiggundu, 1983) has been related to increased motivation. It can also increase team effectiveness because it enhances the sense of responsibility for others' work (Kiggundu, 1983) or because it enhances the reward value of a team's accomplishments (Shea and Guzzo, 1987).

Another form of interdependence is goal interdependence. Goal setting is a well-documented, individual-level performance improvement technique (Locke and Latham, 1990). A clearly defined mission or purpose is considered to be critical to team effectiveness (Campion et al., 1993, 1995; Davis and Wacker, 1987; Hackman, 1987; Sundstrom et al., 1990). Its importance has also been shown in empirical studies on teams (e.g., Buller and Bell, 1986; Woodman and Sherwood, 1980). Not only should goals exist for teams, but individual members' goals must be linked to team goals to be maximally effective.

Finally, interdependent feedback and rewards have also been found to be important for team effectiveness and team member satisfaction (Campion et al., 1993, 1995). Individual feedback and rewards should be linked to a team's performance in order to motivate team-oriented behavior. This characteristic is recognized in many theoretical treatments (e.g., Hackman, 1987; Leventhal, 1976; Steiner, 1972; Sundstrom et al., 1990) and research studies (e.g., Pasmore et al., 1982; Wall et al., 1986).

### 4.3.5 Creating the Organizational Context

Organizational context and resources are considered in all recent models of work team effectiveness (e.g., Guzzo and Shea, 1992; Hackman, 1987). One important aspect of context and resources for teams is adequate training. Training is an extensively researched determinant of team performance (for reviews see Dyer, 1984; Salas et al., 1992), and training is included in most interventions (e.g., Pasmore et al., 1982; Wall et al., 1986). Training is related to team members' satisfaction and managers' and employees' judgments of their teams' effectiveness (Campion et al., 1993, 1995).

Training content often includes team philosophy, group decision making, and interpersonal skills as well as technical knowledge. Many team-building interventions focus on aspects of team functioning that are related to the teamwork KSAs shown in Table 9. A recent review of this literature divided such interventions into four approaches (Tannenbaum et al., 1992)—goal setting, interpersonal, role, and problem solving—which are similar to the teamwork KSA categories. Thus, these interventions could be viewed as training programs on teamwork KSAs. Reviews indicate that the evidence for the effectiveness of this training appears positive despite the methodological limitations that plague this research (Buller and Bell, 1986; Tannenbaum et al., 1992; Woodman and Sherwood, 1980). It appears that workers can be trained in teamwork KSAs. (See Chapter 16 for more information on team training.)

Regarding how such training should be conducted, there is substantial guidance on training teams in the human factors and military literatures (Dyer, 1984; Salas et al., 1992; Swezey and Salas, 1992). Because these topics are thoroughly addressed in the cited sources, they will not be reviewed here.

Managers of teams also need to be trained in teamwork KSAs, regardless of whether the teams are manager led or self-managed. The KSAs are needed for interacting with employee teams and for participating on management teams. It has been noted that managers of teams, especially autonomous work teams, need to develop their employees (Cummings, 1978; Hackman and Oldham, 1980; Manz and Sims, 1987). Thus, training must ensure not only that managers possess teamwork KSAs but also that they know how to train employees on these KSAs.

Managerial support is another contextual characteristic (Morgeson, 2005). Management controls resources (e.g., material and information) required to make team functioning possible (Shea and Guzzo, 1987), and an organization's culture and top management must support

the use of teams (Sundstrom et al., 1992). Teaching facilitative leadership to managers is often a feature of team interventions (Pasmore et al., 1982). Finally, communication and cooperation between teams are contextual characteristics because they are often the responsibility of managers. Supervising team boundaries (Cummings, 1978) and externally integrating teams with the rest of the organization (Sundstrom et al., 1990) enhance effectiveness. Research indicates that managerial support and communication and cooperation between work teams are related to team productivity and effectiveness and to team members' satisfaction with their work (Campion et al., 1993, 1995).

### 4.3.6  Developing Effective Team Process

Process describes those things that go on in the group that influence effectiveness. One process characteristic is potency, or the belief of a team that it can be effective (Guzzo and Shea, 1992; Shea and Guzzo, 1987). It is similar to the lay-term "team spirit." Hackman (1987) argues that groups with high potency are more committed and willing to work hard for the group, and evidence indicates that potency is highly related to team members' satisfaction with work, team productivity, and members' and managers' judgments of their teams' effectiveness (Campion et al., 1993, 1995).

Another process characteristic found to be related to team satisfaction, productivity, and effectiveness is social support (Campion et al., 1993, 1995). Effectiveness can be enhanced when members help each other and have positive social interactions. Like social facilitation (Harkins, 1987; Zajonc, 1965), social support can be arousing and may enhance effectiveness by sustaining effort on mundane tasks.

Another process characteristic related to satisfaction, productivity, and effectiveness is workload sharing (Campion et al.1993, 1995). Workload sharing enhances effectiveness by preventing social loafing or free riding (Harkins, 1987). To enhance sharing, group members should believe their individual performance can be distinguished from the group's and that there is a link between their performance and outcomes.

Finally, communication and cooperation within the work group are also important to team effectiveness, productivity, and satisfaction (Campion et al. 1993, 1995). Management should help teams foster open communication, supportiveness, and discussions of strategy. Informal, rather than formal, communication channels and mechanisms of control should be promoted to ease coordination (Bass and Klubeck, 1952; Majchrzak, 1988). Managers should encourage self-evaluation, self-observation, self-reinforcement, self-management, and self-goal setting by teams. Self-criticism for purposes of recrimination should be discouraged (Manz and Sims, 1987).

Recent meta-analytic evidence suggests that numerous team processes are related to both team performance and member satisfaction (LePine et al., 2008). Many of these processes can be grouped into three categories. Transition processes are team actions that occur after one team task has ended and before the next begins and include actions such as mission analysis, goal specification, and strategy formulation/planning. Action processes are team activities that occur during the completion of a task. The four types of action processes include monitoring progress toward goals, systems monitoring (assessing resources and environmental factors that could influence goal accomplishment), team monitoring and backup behavior (team members assisting each other in their individual tasks), and coordination. Finally, team activities geared toward maintaining the team's interpersonal relationships are called interpersonal processes and include conflict management, motivating/confidence building, and affect management (e.g., emotional balance, togetherness, and coping with demands/frustrations). The results suggest that there are specific team processes that occur at different stages of task completion, and the occurrence, or lack thereof, of these processes has an impact on both the teams' performance and team members' satisfaction (LePine et al. 2008).

## 5  MEASUREMENT AND EVALUATION OF JOB AND TEAM DESIGN

The purpose of an evaluation study for either a job or team design is to provide an objective evaluation of success and to create a tracking and feedback system to make adjustments during the course of the design project. An evaluation study can provide objective data to make informed decisions, help tailor the process to the organization, and give those affected by the design or redesign an opportunity to provide input (see Morgeson and Campion, 2002). An evaluation study should include measures that describe the characteristics of the jobs or teams so that it can be determined whether or not jobs or teams ended up having the characteristics they were intended to have. An evaluation study should also include measures of effectiveness outcomes an organization hoped to achieve with a design project. Measures of effectiveness could include such *subjective* outcomes as employee job satisfaction or employee, manager, or customer perceptions of effectiveness. Measures of effectiveness should also include *objective* outcomes such as cost, productivity, rework/scrap, turnover, accident rates, or absenteeism. Additional information on measurement and evaluation of such outcomes can be found in Part 6 of this handbook.

### 5.1  Using Questionnaires to Measure Job and Team Design

One way to measure job or team design is by using questionnaires or checklists. This method of measuring job or team design is highlighted because it has been used widely in research on job design, especially on the motivational approach. More importantly, questionnaires are a very inexpensive, easy, and flexible way to measure work design characteristics. Moreover, they gather information from job experts, such as incumbents, supervisors, and engineers and other analysts.

Several questionnaires exist for measuring the motivational approach to job design (Hackman and Oldham, 1980; Sims et al., 1976), but only one questionnaire, the

*multimethod job design questionnaire*, measures characteristics for all four approaches to job design. This questionnaire (presented in Table 2) evaluates the quality of a job's characteristics based on each of the four approaches. The *team design measure* (presented in Table 4) evaluates the quality of work design based on the team approach.

Questionnaires can be administered in a variety of ways. Employees can complete them individually at their convenience at their work station or some other designated area or they can complete them in a group setting. Group administration allows greater standardization of instructions and provides the opportunity to answer questions and clarify ambiguities. Managers and engineers can also complete the questionnaires either individually or in a group session. Engineers and analysts usually find that observation of the work site, examination of the equipment and procedures, and discussions with any incumbents or managers are important methods of gaining information on the work before completing the questionnaires.

Scoring for each job design approach or for each team characteristic on the questionnaires is usually accomplished by simply averaging the applicable items. Then scores from different incumbents, managers, or engineers describing the same job or team are combined by averaging. Multiple items and multiple respondents are used to improve the reliability and accuracy of the results. The implicit assumption is that slight differences among respondents are to be expected because of legitimate differences in viewpoint. However, absolute differences in scores should be examined on an item-by-item basis, and large discrepancies (e.g., more than one point) should be discussed to clarify possible differences in interpretation. It may be useful to discuss each item until a consensus rating is reached.

The higher the score on a particular job design scale or work team characteristic scale, the better the quality of the design in terms of that approach or characteristic. Likewise, the higher the score on a particular item, the better the design is on that dimension. How high a score is needed or necessary cannot be stated in isolation. Some jobs or teams are naturally higher or lower on the various approaches, and there may be limits to the potential of some jobs. The scores have most value in comparing different jobs, teams, or design approaches, rather than evaluating the absolute level of the quality of a job or team design. However, a simple rule of thumb is that if the score for an approach is smaller than 3, the job or team is poorly designed on that approach and it should be reconsidered. Even if the average score on an approach is greater than 3, examine any individual dimension scores that are at 2 or 1.

**Uses of Questionnaires in Different Contexts**:

1. *Designing New Jobs or Teams*. When jobs or teams do not yet exist, the questionnaire is used to evaluate proposed job or team descriptions, work stations, equipment, and so on. In this role, it often serves as a simple design checklist. Additional administrations of the questionnaire in later months or years can be used to assess the longer term effects of the job or team design.

2. *Redesigning Existing Jobs or Teams or Switching from Job to Team Design*. When jobs or teams already exist, there is a much greater wealth of information. Questionnaires can be completed by incumbents, managers, and engineers. Questionnaires can be used to measure design both before and after changes are made to compare the redesign with the previous design approach. A premeasure before the redesign can be used as a baseline measurement against which to compare a postmeasure conducted right after the redesign implementation. A follow-up measure can be used in later months or years to assess the long-term difference between the previous design approach and the new approach. If other sites or plants with the same types of jobs or teams are not immediately included in the redesign but are maintained with the older design approach, they can be used as a comparison or "control group" to enable analysts to draw even stronger conclusions about the effectiveness of the redesign. Such a control group allows one to control for the possibilities that changes in effectiveness were not due to the redesign but were in fact due to some other causes such as increases in workers' knowledge and skills with the passage of time, changes in workers' economic environment (i.e., job security, wages, etc.), or workers trying to give socially desirable responses to questionnaire items.

3. *Diagnosing Problem Job or Team Designs*. *When* problems occur, regardless of the apparent source of the problem, the job or team design questionnaires can be used as a diagnostic device to determine if any problems exist with the design of the jobs or teams.

## 5.2  Choosing Sources of Data

1. *Incumbents*. Incumbents are probably the best source of information for existing jobs or teams. Having input can enhance the likelihood that changes will be accepted, and involvement in such decisions can enhance feelings of participation, thus increasing motivational job design in itself (see item 22 of the motivational scale in Table 2). One should include a large number of incumbents for each job or team because there can be slight differences in perceptions of the same job or team due to individual differences (discussed in Section 4.1). Evidence suggests that one should include at least five incumbents for each job or team, but more are preferable (Campion, 1988; Campion and McClelland, 1991; Campion et al., 1993, 1995).

2. *Managers or Supervisors*. First-level managers or supervisors may be the next most knowledgeable persons about an existing work design. They may also provide information on jobs or teams under development. Some differences in perceptions of the same job or team will exist

among managers, so multiple managers should be used.

3. *Engineers or Analysts*. Engineers may be the only source of information if the jobs or teams are not yet developed. But also for existing jobs or teams, an outside perspective of an engineer, analyst, or consultant may provide a more objective viewpoint. Again, there can be differences among engineers, so several should evaluate each job or team.

It is desirable to get multiple inputs and perspectives from different sources in order to get the most reliable and accurate picture of the results of the job or team design.

## 5.3 Long-Term Effects and Potential Biases

It is important to recognize that some effects of job or team design may not be immediate, others may not be long lasting, and still others may not be obvious. Initially, when jobs or teams are designed, or right after they are redesigned, there may be a short-term period of positive attitudes (often called a "honeymoon effect"). As the legendary Hawthorne studies indicated, changes in jobs or increased attention paid to workers tends to create novel stimulation and positive attitudes (Mayo, 1933). Such transitory elevations in affect should not be mistaken for long-term improvements in satisfaction, as they may wear off over time. In fact, with time, employees may realize their work is now more complex and should be paid higher compensation (Campion and Berger, 1990).

Costs which are likely to lag in time also include stress and fatigue, which may take a while to build up if mental demands have been increased excessively. Boredom may take a while to set in if mental demands have been overly decreased. In terms of lagged benefits, productivity and quality are likely to improve with practice and learning on the new job or team. And some benefits, like reduced turnover, simply take time to estimate accurately.

Benefits which may potentially dissipate with time include satisfaction, especially if the elevated satisfaction is a function of novelty rather than basic changes to the motivating value of the work. Short-term increases in productivity due to heightened effort rather than better design may not last. Costs which may dissipate include training requirements and staffing difficulties. Once jobs are staffed and everyone is trained, these costs disappear until turnover occurs. So these costs will not go away completely, but they may be less after initial start-up. Dissipating heightened satisfaction but long-term increases in productivity were observed in a motivational job redesign study conducted by Griffin (1991). These are only examples to illustrate how dissipating and lagged effects might occur. A more detailed example of long-term effects is given in Section 5.8.

A potential bias which may confuse the proper evaluation of benefits and costs is spillover. Laboratory research has shown that the job satisfaction of employees can bias perceptions of the motivational value of their jobs (O'Reilly et al., 1980). Likewise, the level of morale in the organization can have a spillover effect onto employees' perceptions of job or team design. If morale is particularly high, it may have an elevating effect on how employees or analysts view the jobs or teams; conversely, low morale may have a depressing effect on views. The term *morale* refers to the general level of job satisfaction across employees, and it may be a function of many factors, including management, working conditions, and wages. Another factor which has an especially strong effect on employee reactions to work design changes is *employment security*. Obviously, employee enthusiasm for work design changes will be negative if they view them as potentially decreasing their job security. Every effort should be made to eliminate these fears. The best method of addressing these effects is to be attentive to their potential existence and to conduct longitudinal evaluations of job and team design.

In addition to questionnaires, many other analytical tools are useful for work design. The disciplines that contributed the different approaches to work design have also contributed different techniques for analyzing tasks, jobs, and processes for design and redesign purposes. These techniques include job analysis methods created by specialists in industrial psychology, variance analysis methods created by specialists in sociotechnical design, time-and-motion analysis methods created by specialists in industrial engineering, and linkage analysis methods created by specialists in human factors. This section briefly describes a few of these techniques to illustrate the range of options. The reader is referred to the citations for detail on how to use the techniques.

## 5.4 Job Analysis

Job analysis can be broadly defined as a number of systematic techniques for collecting and making judgments about job information (Morgeson and Campion, 1997, 2000). Information derived from job analysis can be used to aid in recruitment and selection decisions, determine training and development needs, develop performance appraisal systems, and evaluate jobs for compensation as well as analyze tasks and jobs for job design. Job analysis may also focus on tasks, worker characteristics, worker functions, work fields, working conditions, tools and methods, products and services, and so on. Job analysis data can come from job incumbents, supervisors, and analysts who specialize in the analysis of jobs. Data may also be provided by higher management levels or subordinates in some cases.

Considerable literature has been published on the topic of job analysis (Ash et al., 1983; Dierdorff and Wilson, 2003; Gael, 1983; Harvey, 1991; Morgeson and Campion, 1997; Morgeson et al., 2004; Peterson et al., 2001; U.S. Department of Labor, 1972). Some of the more typical methods of analysis are briefly described below:

1. *Conferences and Interviews*. Conferences or interviews with job experts, such as incumbents and supervisors, are often the first step. During such meetings, information collected typically includes job duties and tasks, KSAs, and other worker characteristics.

2. *Questionnaires*. Questionnaires are used to collect information efficiently from a large number of people. Questionnaires require considerable prior knowledge of the job to form the basis of the items (e.g., primary tasks). Often this information is first collected through conferences and interviews, and then the questionnaire is constructed and used to collect judgments about the job (e.g., importance and time spent on each task). Some standardized questionnaires have been developed which can be applied to all jobs to collect basic information on tasks and requirements. Examples of standardized questionnaires are the position analysis questionnaire (McCormick et al., 1972) and the Occupational Information Network (O*NET; Peterson et al., 2001).

3. *Inventories*. Inventories are much like questionnaires, except they are simpler in format. They are usually simple checklists where the job expert checks whether a task is performed or an attribute is required.

4. *Critical Incidents*. This form of job analysis focuses only on aspects of worker behavior which are especially effective or ineffective.

5. *Work Observation and Activity Sampling*. Quite often job analysis includes the actual observation of work performed. More sophisticated technologies involve statistical sampling of work activities.

6. *Diaries*. Sometimes it is useful or necessary to collect data by having the employee keep a diary of activities on his or her job.

7. *Functional Job Analysis*. Task statements can be written in a standardized fashion. Functional job analysis suggests how to write task statements (e.g., start with a verb, be as simple and discrete as possible). It also involves rating jobs on the degree of data, people, and things requirements. This form of job analysis was developed by the U.S. Department of Labor and has been used to describe over 12,000 jobs as documented in the *Dictionary of Occupational Titles* (Fine and Wiley, 1971; U.S. Department of Labor, 1977).

Very limited research has been done to evaluate the practicality and quality of various job analysis methods for different purposes. But analysts seem to agree that combinations of methods are preferable to single methods (Levine et al., 1983; Morgeson and Campion, 1997).

Current approaches to job analysis do not give much attention to analyzing teams. For example, the *Dictionary of Occupational Titles* (U.S. Department of Labor, 1972) considers "people" requirements of jobs but does not address specific teamwork KSAs. Likewise, recent reviews of the literature mention some components of teamwork such as communication and coordination (e.g., Harvey, 1991) but give little attention to other teamwork KSAs. Thus, job analysis systems may need to be revised. The recent O*NET reflects a major new job analysis system designed to replace the *Dictionary of Titles* (Peterson et al., 2001). Although not explicitly addressing the issue of teamwork KSAs, it does contain a large number of worker attribute domains that may prove useful. Teamwork KSAs are more likely to emerge with conventional approaches to job analysis because of their unstructured nature (e.g., interviews), but structured approaches (e.g., questionnaires) will have to be modified to query about teamwork KSAs.

## 5.5 Other Approaches

Variance analysis is a tool used to identify areas of technological uncertainty in a production process (Davis and Wacker, 1982). It aids the organization in designing jobs to allow job holders to control the variability in their work. See Chapters 12 and 13 in this handbook for more information on task and workload analysis.

Industrial engineers have also created many techniques to help job designers visualize operations in order to improve efficiencies, which has led to the development of a considerable literature on the topic of time-and-motion analysis (e.g., Mundel, 1985; Niebel, 1988). Some of these techniques are *process charts* (graphically represent separate steps or events that occur during performance of a task or series of actions); *flow diagrams* (utilize drawings of an area or building in which an activity takes place and use lines, symbols, and notations to help designers visualize the physical layout of the work); *possibility guides* (tools for systematically listing all possible changes suggested for a particular activity or output, and examine the consequences of suggestions to aid in selecting the most feasible changes); and *network diagrams* (describe complex relationships, where a circle or square represents a "status," a partial or complete service, or substantive output; heavy lines represent "critical paths," which determine the minimum expected completion time for a project).

Linkage analysis is another technique used by human factors specialists to represent relationships (i.e., "links") between components (i.e., people or things) in a work system (Sanders and McCormick, 1987). Designers of physical work arrangements use tools (i.e., link tables, adjacency layout diagrams, and spatial operational sequences) to represent relationships between components in order to better understand how to arrange components to minimize the distance between frequent or important links.

## 5.6 Example of an Evaluation of a Job Design

Studies conducted by Campion and McClelland (1991, 1993) are described as an illustration of an evaluation of a job redesign project. They illustrate the value of considering an interdisciplinary perspective. The setting was a large financial services company. The units under study processed the paperwork in support of other units that sold the company's products. Jobs had been designed in a mechanistic manner such that individual employees prepared, sorted, coded, and computer input the paper flow.

The organization viewed the jobs as too mechanistically designed. Guided by the motivational approach, the project intended to enlarge jobs by combining existing

jobs in order to attain three objectives: (1) enhance motivation and satisfaction of employees, (2) increase incumbent feelings of ownership of the work, thus increasing customer service, and (3) maintain productivity in spite of potential lost efficiencies from the motivational approach. The consequences of all approaches to job design were considered. It was anticipated that the project would increase motivational consequences, decrease mechanistic and perceptual/motor consequences, and have no effect on biological consequences (Table 1).

The evaluation consisted of collecting detailed data on job design and a broad spectrum of potential benefits and costs of enlarged jobs. The research strategy involved comparing several varieties of enlarged jobs with each other and with unenlarged jobs. Questionnaire data were collected and focused team meetings were conducted with incumbents, managers, and analysts. The study was repeated at five different geographic sites.

Results indicated enlarged jobs had the benefits of more employee satisfaction, less boredom, better quality, and better customer service, but they also had the costs of slightly higher training, skill, and compensation requirements. Another finding was that all potential costs of enlarging jobs were not observed, suggesting that redesign can lead to benefits without incurring every cost in a one-to-one fashion.

In a two-year follow-up evaluation study, it was found that the costs and benefits of job enlargement changed substantially over time, depending on the type of enlargement. Task enlargement, which was the focus of the original study, had mostly long-term costs (e.g., lower satisfaction, efficiency, and customer service and more mental overload and errors). Conversely, knowledge enlargement, which emerged as a form of job design since the original study, had mostly benefits (e.g., higher satisfaction and customer service and lower overload and errors).

There are several important implications of the latter study. First, it illustrates that the long-term effects of job design changes can be different than the short-term effects. Second, it shows the classic distinction between enlargement and enrichment (Herzberg, 1966) in that simply adding more tasks did not improve the job, but adding more knowledge opportunities did. Third, it illustrates how the job design process is iterative. In this setting, the more favorable knowledge enlargement was discovered only after gaining experience with task enlargement. Fourth, as in the previous study, it shows that it is possible in some situations to gain benefits of job design without incurring all the potential costs, thus minimizing the trade-offs between the motivational and mechanistic approaches to job design.

## 5.7 Example of an Evaluation of a Team Design

Studies conducted by the authors and their colleagues are described here as an illustration of an evaluation of a team design project (Campion et al., 1993, 1995). They illustrate the use of multiple sources of data and multiple types of team effectiveness outcomes. The setting was the same financial services company as in the example job design evaluation above. Questionnaires based on Table 4 were administered to 391 clerical employees in 80 teams and 70 team managers in the first study (Campion et al., 1993) and to 357 professional workers in 60 teams (e.g., systems analysts, claims specialists, underwriters) and 93 managers in the second study (Campion et al., 1996) to measure teams' design characteristics. Thus, two sources of data were used, team members and team managers, to measure the team design characteristics.

In both studies, effectiveness outcomes included the organization's employee satisfaction survey, which had been administered at a different time than the team design characteristics questionnaire, and managers' judgments of teams' effectiveness, measured at the same time as the team design characteristics. In the first study, several months of records of team productivity were also used to measure effectiveness. Additional effectiveness measures in the second study were employees' judgments of their team's effectiveness, measured at the same time as the team design characteristics, managers' judgments of teams' effectiveness, measured a second time three months after the team design characteristics, and the average of team members' most recent performance ratings.

Results indicated that all of the team design characteristics had positive relationships with at least some of the outcomes. Relationships were strongest for process characteristics, followed by job design, context, interdependence, and composition characteristics (see Figure 1). Results also indicated that when teams were well designed according to the team design approach, they were higher on both employee satisfaction and team effectiveness ratings than less well designed teams.

Results were stronger when the team design characteristics data were from team members, rather than from the team managers. This illustrates the importance of collecting data from different sources to gain different perspectives on the results of a team design project. Collecting data from only a single source may lead one to draw different conclusions about a design project than if one obtains a broader picture of the team design results from multiple sources.

Results were also stronger when outcome measures came from employees (employee satisfaction, team member judgments of their teams), managers rating their own teams, or productivity records, than when they came from other managers or from performance appraisal ratings. This illustrates the use of different types of outcome measures to avoid drawing conclusions from overly limited data. This example also illustrates the use of separate data collection methods and times for collecting team design characteristics data versus team outcomes data. A single data collection method and time in which team design characteristics and outcomes are collected from the same source (e.g., team members only) on the same day can create an illusion of higher relationships between design characteristics and outcomes than really exist. Although it is more costly to use multiple sources, methods, and administration times, the ability to draw conclusions from the results is far stronger if one does.

# REFERENCES

Argyris, C. (1964), *Integrating the Individual and the Organization*, Wiley, New York.

Ash, R. A., Levine, E. L., and Sistrunk, F. (1983), "The Role of Jobs and Job-Based Methods in Personnel and Human Resources Management," in *Research in Personnel and Human Resources Management*, Vol. 1, K. M. Rowland and G. R. Ferris, Eds., JAI Press, Greenwich, CT.

Astrand, P. O., and Rodahl, K. (1977), *Textbook of Work Physiology: Physiological Bases of Exercise*, 2nd ed., McGraw-Hill, New York.

Babbage, C. (1835), "On the Economy of Machinery and Manufacturers," reprinted in *Design of Jobs*, 2nd ed., L. E. Davis and J. C. Taylor, Eds., Goodyear, Santa Monica, CA.

Banker, R. D., Field, J. M., Schroeder, R. G., and Sinha, K. K. (1996), "Impact of Work Teams on Manufacturing Performance: A Longitudinal Study," *Academy of Management Journal*, Vol. 39, pp. 867–890.

Bass, B. M., and Klubeck, S. (1952), "Effects of Seating Arrangements and Leaderless Team Discussions," *Journal of Abnormal and Social Psychology*, Vol. 47, pp. 724–727.

Blauner, R. (1964), *Alienation and Freedom*, University of Chicago Press, Chicago, IL.

Brown, B. K., and Campion, M. A. (1994), "Biodata Phenomenology: Recruiters' Perceptions and Use of Biographical Information in Personnel Selection," *Journal of Applied Psychology*, Vol. 79, pp. 897–908.

Buller, P. F., and Bell, C. H. (1986), "Effects of Team Building and Goal Setting on Productivity: A Field Experiment," *Academy of Management Journal*, Vol. 29, pp. 305–328.

Buys, C. J. (1978), "Humans Would Do Better without Groups," *Personality and Social Psychology Bulletin*, Vol. 4, pp. 123–125.

Campion, M. A. (1988), "Interdisciplinary Approaches to Job Design: A Constructive Replication with Extensions," *Journal of Applied Psychology*, Vol. 73, pp. 467–481.

Campion, M. A. (1989), "Ability Requirement Implications of Job Design: An Interdisciplinary Perspective," *Personnel Psychology*, Vol. 42, pp. 1–24.

Campion, M. A., and Berger, C. J. (1990), "Conceptual Integration and Empirical Test of Job Design and Compensation Relationships," *Personnel Psychology*, Vol. 43, pp. 525–554.

Campion, M. A., and McClelland, C. L. (1991), "Interdisciplinary Examination of the Costs and Benefits of Enlarged Jobs: A Job Design Quasi-Experiment," *Journal of Applied Psychology*, Vol. 76, pp. 186–198.

Campion, M. A., and McClelland, C. L. (1993), "Follow-up and Extension of the Interdisciplinary Costs and Benefits of Enlarged Jobs," *Journal of Applied Psychology*, Vol. 78, pp. 339–351.

Campion, M. A., and Medsker, G. J. (1992), "Job Design," in *Handbook of Industrial Engineering*, G. Salvendy, Eds., Wiley, New York.

Campion, M. A., and Stevens, M. J. (1991), "Neglected Questions in Job Design: How People Design Jobs, Influence of Training, and Task-Job Predictability," *Journal of Business and Psychology*, Vol. 6, pp. 169–191.

Campion, M. A., and Thayer, P. W. (1985), "Development and Field Evaluation of an Interdisciplinary Measure of Job Design," *Journal of Applied Psychology*, Vol. 70, pp. 29–43.

Campion, M. A., and Thayer, P. W. (1987), "Job Design: Approaches, Outcomes, and Trade-offs," *Organizational Dynamics*, Vol. 15, No. 3, pp. 66–79.

Campion, M. A., Medsker, G. J., and Higgs, A. C. (1993), "Relations between Work Group Characteristics and Effectiveness: Implications for Designing Effective Work Groups," *Personnel Psychology*, Vol. 46, pp. 823–850.

Campion, M. A., Campion, J. E., and Hudson, J. P. (1994a), "Structured Interviewing: A Note on Incremental Validity and Alternative Question Types," *Journal of Applied Psychology*, Vol. 79, pp. 998–1002.

Campion, M. A., Cheraskin, L., and Stevens, M. J. (1994b), "Job Rotation and Career Development: Career-Related Antecedents and Outcomes of Job Rotation," *Academy of Management Journal*, Vol. 37, pp. 1518–1542.

Campion, M. A., Papper, E. M., and Medsker, G. J. (1996), "Relations between Work Team Characteristics and Effectiveness: A Replication and Extension," *Personnel Psychology*, Vol. 49, pp. 429–452.

Caplan, R. D., Cobb, S., French, J. R. P., Van Harrison, R., and Pinneau, S. R. (1975), "Job Demands and Worker Health: Main Effects and Occupational Differences," HEW Publication No. NIOSH 75-160, U.S. Government Printing Office, Washington, DC.

Cartwright, D. (1968), "The Nature of Team Cohesiveness," in *Team Dynamics: Research and Theory*, D. Cartwright and A. Zander, Eds., 3rd ed., Harper & Row, New York.

Chapanis, A. (1970), "Relevance of Physiological and Psychological Criteria to Man-Machine Systems: The Present State of the Art," *Ergonomics*, Vol. 13, pp. 337–346.

Cook, T. D., and Campbell, D. T. (1979), *Quasi-Experimentation: Design & Analysis Issues for Field Settings*, Rand-McNally, Chicago, IL.

Cummings, T. G. (1978), "Self-Regulating Work Teams: A Sociotechnical Synthesis," *Academy of Management Review*, Vol. 3, pp. 625–634.

Cummings, T. G. (1981), "Designing Effective Work Groups," in *Handbook of Organization Design*, Vol. 2., P. C. Nystrom and W. H. Starbuck, Eds., Oxford University Press, New York.

Davis, L. E. (1957), "Toward a Theory of Job Design," *Journal of Industrial Engineering*, Vol. 8, pp. 305–309.

Davis, L. E. (1971), "The Coming Crisis for Production Management: Technology and Organization," *International Journal of Production Research*, Vol. 9, pp. 65–82.

Davis, L. E. (1982), "Organization Design," in *Handbook of Industrial Engineering*, G. Salvendy, Eds., Wiley, New York.

Davis, L. E., and Taylor, J. C. (1979), *Design of Jobs*, 2nd ed., Goodyear, Santa Monica, CA.

Davis, L. E., and Valfer, E. S. (1965), "Intervening Responses to Changes in Supervisor Job Design," *Occupational Psychology*, Vol. 39, pp. 171–189.

Davis, L. E., and Wacker, G. L. (1982), "Job Design," in *Handbook of Industrial Engineering*, G. Salvendy, Eds., Wiley, New York.

Davis, L. E., and Wacker, G. L. (1987), "Job Design," In *Handbook of Human Factors*, G. Salvendy, Eds., Wiley, New York.

Davis, L. E., Canter, R. R., and Hoffman, J. (1955), "Current Job Design Criteria," *Journal of Industrial Engineering*, Vol. 6, No. 2, pp. 5–8, 21–23.

Dierdorff, E. C., and Wilson, M. A. (2003), "A Meta-Analysis of Job Analysis Reliability," *Journal of Applied Psychology*, Vol. 88, pp. 635–646.

Dyer, J. (1984), "Team Research and Team Training: A State-of-the-Art Review," In *Human Factors Review*, F. A. Muckler, Ed., Human Factors Society, Santa Monica, CA.

Edwards, J. R., Scully, J. A., and Brtek, M. D. (1999), "The Measurement of Work: Hierarchical Representation of the Multimethod Job Design Questionnaire," *Personnel Psychology*, Vol. 52, pp. 305–334.

Edwards, J. R., Scully, J. A., and Brtek, M. D. (2000), "The Nature and Outcomes of Work: A Replication and Extension of Interdisciplinary Work-Design Research," *Journal of Applied Psychology*, Vol. 85, pp. 860–868.

Emory, F. E., and Trist, E. L. (1960), "Sociotechnical Systems," in *Management Sciences, Models, and Techniques*, Vol. 2, C. W. Churchman and M. Verhulst, Eds., Pergamon, London.

Fine, S. A., and Wiley, W. W. (1971), *An Introduction to Functional Job Analysis*, W. E. Upjohn Institute for Employment Research, Kalamazoo, MI.

Ford, R. N. (1969), *Motivation through the Work Itself*, American Management Association, New York.

Fried, Y., and Ferris, G. R. (1987), "The Validity of the Job Characteristics Model: A Review and Metaanalysis," *Personnel Psychology*, Vol. 40, pp. 287–322.

Gael, S. (1983), *Job Analysis: A Guide to Assessing Work Activities*, Jossey-Bass, San Francisco, CA.

Gallagher, C. C., and Knight, W. A. (1986), *Team Technology Production Methods in Manufacture*, Ellis Horwood, Chichester.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., and Benston, C. (1987), "Metaanalysis of Assessment Center Validity (Monograph)," *Journal of Applied Psychology*, Vol. 72, pp. 493–511.

Gilbreth, F. B. (1911), *Motion Study: A Method for Increasing the Efficiency of the Workman*, Van Nostrand, New York.

Gladstein, D. L. (1984), "Groups in Context: A Model of Task Group Effectiveness," *Administrative Science Quarterly*, Vol. 29, pp. 499–517.

Globerson, S., and Crossman, E. R. (1976), "Nonrepetitive Time: An Objective Index of Job Variety," *Organizational Behavior and Human Performance*, Vol. 17, pp. 231–240.

Goodman, P. S., Ravlin, E. C., and Argote, L. (1986), "Current Thinking about Teams: Setting the Stage for New Ideas," in *Designing Effective Work Teams*, P. S. Goodman and Associates, Eds., Jossey-Bass, San Francisco.

Grandjean, E. (1980), *Fitting the Tasks to the Man: An Ergonomic Approach*, Taylor & Francis, London.

Grant, A. M. (2007), "Relational Job Design and the Motivation to Make a Prosocial Difference," *Academy of Management Review*, Vol. 32, pp. 393–417.

Grant, A. M. (2008), "The Significance of Task Significance: Job Performance Effects, Relational Mechanisms, and Boundary Conditions," *Journal of Applied Psychology*, Vol. 93, pp. 108–124.

Grant, A. M., and Parker, S. K. (2009), "Redesigning Work Design Theories: The Rise of Relational and Proactive Perspectives," *The Academy of Management Annals*, Vol. 3, pp. 317–375.

Grant, A. M., Campbell, E. M., Chen, G., Cottone, K., Lapedis, D., and Lee, K. (2007), "Impact and the Art of Motivation Maintenance: The Effects of Contact with Beneficiaries on Persistence Behavior," *Organizational Behavior and Human Decision Processes*, Vol. 103, pp. 53–67.

Griffin, R. W. (1982), *Task Design: An Integrative Approach*, Scott-Foresman, Glenview, IL.

Griffin, R. W. (1991), "Effects of Work Redesign on Employee Perceptions, Attitudes, and Behaviors –A Long-Term Investigation," *Academy of Management Journal*, Vol. 34, pp. 425–435.

Guzzo, R. A., and Shea, G. P. (1992), "Group Performance and Intergroup Relations in Organizations," in *Handbook of Industrial and Organizational Psychology*, Vol. 3, M. D. Dunnette and L. M. Hough, Eds., Consulting Psychologists Press, Palo Alto, CA.

Hackman, J. R. (1987), "The Design of Work Teams," in *Handbook of Organizational Behavior*, J. Lorsch, Eds., Prentice-Hall, Englewood Cliffs, NJ.

Hackman, J. R. (2002), *Leading Teams: Setting the Stage for Great Performances*, Harvard Business School Press, Boston.

Hackman, J. R., and Oldham, G. R. (1980), *Work Redesign*, Addison-Wesley, Reading, MA.

Hammond, R. W. (1971), "The History and Development of Industrial Engineering," in *Industrial Engineering Handbook*, 3rd ed., H. B. Maynard, Eds., McGraw-Hill, New York.

Harkins, S. G. (1987), "Social Loafing and Social Facilitation," *Journal of Experimental Social Psychology*, Vol. 23, pp. 1–18.

Harvey, R. J. (1991), "Job Analysis," in *Handbook of Industrial and Organizational Psychology*, 2nd ed., Vol. 2, M. D. Dunnette and L. M. Hough, Eds., Consulting Psychologists Press, Palo Alto, CA.

Herzberg, F. (1966), *Work and the Nature of Man*, World, Cleveland, OH.

Hoerr, J. (1989), "The Payoff from Teamwork," *Business Week*, July 10, pp. 56–62.

Hofstede, G. (1980), *Culture's Consequences*, Sage, Beverly Hills, CA.

Homans, G. C. (1950), *The Human Group*, Harcourt, Brace, and World, New York.

Hoppock, R. (1935), *Job Satisfaction*, Harper and Row, New York.

Hornung, S., Rousseau, D. M., and Glaser, J. (2008), "Creating Flexible Work Arrangements through Idiosyncratic Deals," *Journal of Applied Psychology*, Vol. 93, pp. 655–664.

Humphrey, S. E., Nahrgang, J. D., and Morgeson, F. P. (2007), "Integrating Motivational, Social, and Contextual Work Design Features: A Meta-Analytic Summary and Theoretical Extension of the Work Design Literature," *Journal of Applied Psychology*, Vol. 92, pp. 1332–1356.

Hyer, N. L. (1984), "Management's Guide to Team Technology," in *Team Technology at Work*, N. L. Hyer, Ed., Society of Manufacturing Engineers, Dearborn, MI.

Ilgen, D. R., Hollenbeck, J. R., Johnson, M. D., and Jundt, D. (2005), "Teams in Organizations: From Input-Process-Output Models to IMOI Models," *Annual Review of Psychology*, Vol. 56, pp. 517–543.

Janis, I. L. (1972), *Victims of Groupthink*, Houghton-Mifflin, Boston.

Johansson, G., Aronsson, G., and Lindstrom, B. O. (1978), "Social Psychological and Neuroendocrine Stress Reactions in Highly Mechanized Work," *Ergonomics*, Vol. 21, pp. 583–599.

Kelly, J. (1982), *Scientific Management, Job Redesign, and Work Performance*, Academic, London.

Kiggundu, M. N. (1983), "Task Interdependence and Job Design: Test of a Theory," *Organizational Behavior and Human Performance*, Vol. 31, pp. 145–172.

Konz, S. (1983), *Work Design: Industrial Ergonomics*, 2nd ed., Grid, Columbus, OH.

Lawler, E. E. (1986), *High-Involvement Management: Participative Strategies for Improving Organizational Performance*, Jossey-Bass, San Francisco.

LePine, J. A., Piccolo, R. F., Jackson, C. L., Mathieu, J. E., and Saul, J. R. (2008), "A Meta-Analysis of Teamwork Processes: Tests of a Multidimensional Model and Relationships with Team Effectiveness Criteria," *Personnel Psychology*, Vol. 61, pp. 273–307.

Leventhal, G. S. (1976), "The Distribution of Rewards and Resources in Teams and Organizations," in *Advances in Experimental Social Psychology*, Vol. 9, L. Berkowitz and E. Walster, Eds., Academic, New York.

Levine, E. L., Ash, R. A., Hall, H., and Sistrunk, F. (1983), "Evaluation of Job Analysis Methods by Experienced Job Analysts," *Academy of Management Journal*, Vol. 26, pp. 339–348.

Locke, E. A., and Latham, G. P. (1990), *A Theory of Goal Setting and Task Performance*, Prentice Hall, Englewood Cliffs, NJ.

Majchrzak, A. (1988), *The Human Side of Factory Automation*, Jossey-Bass, San Francisco.

Manz, C. C., and Sims, H. P. (1987), "Leading Workers to Lead Themselves: The External Leadership of Self-Managing Work Teams," *Administrative Science Quarterly*, Vol. 32, pp. 106–129.

Mayo, E. (1933), *The Human Problems of an Industrial Civilization*, Macmillan, New York.

McCormick, E. J., Jeanneret, P. R., and Mecham, R. C. (1972), "A Study of Job Characteristics and Job Dimensions as Based on the Position Analysis Questionnaire (PAQ)," *Journal of Applied Psychology*, Vol. 56, pp. 347–368.

McGrath, J. E. (1984), *Teams: Interaction and Performance*, Prentice-Hall, Englewood Cliffs, NJ.

Meister, D. (1971), *Human Factors: Theory and Practice*, Wiley, New York.

Milkovich, G. T., and Newman, J. M. (1993), *Compensation*, 4th ed., Business Publications, Homewood, IL.

Mintzberg, H. (1979), *The Structuring of Organizations: A Synthesis of the Research*, Prentice-Hall, Englewood Cliffs, NJ.

Morgeson, F. P. (2005), "The External Leadership of Self-Managing Teams: Intervening in the Context of Novel and Disruptive Events," *Journal of Applied Psychology*, Vol. 90, pp. 497–508.

Morgeson, F. P., and Campion, M. A. (1997), "Social and Cognitive Sources of Potential Inaccuracy in Job Analysis," *Journal of Applied Psychology*, Vol. 82, pp. 627–655.

Morgeson, F. P., and Campion, M. A. (2000), "Accuracy in Job Analysis: Toward an Inference-Based Model," *Journal of Organizational Behavior*, Vol. 21, pp. 819–827.

Morgeson, F. P., and Campion, M. A. (2002), "Minimizing Tradeoffs When Redesigning Work: Evidence from a Longitudinal Quasi-Experiment," *Personnel Psychology*, Vol. 55, pp. 589–612.

Morgeson, F. P., and Campion, M. A. (2003), "Work Design," in *Handbook of Psychology: Industrial and Organizational Psychology*, Vol. 12, W. C. Borman, D. R. Ilgen, and R. J. Klimoski, Eds., Wiley, Hoboken, NJ, pp. 423–452.

Morgeson, F. P., and Humphrey, S. E. (2006), "The Work Design Questionnaire (WDQ): Developing and Validating a Comprehensive Measure for Assessing Job Design and the Nature of Work," *Journal of Applied Psychology*, Vol. 91, pp. 1321–1339.

Morgeson, F. P., Aiman-Smith, L. D., and Campion, M. A. (1997), "Implementing Work Teams: Recommendations from Organizational Behavior and Development Theories," in *Advances in Interdisciplinary Studies of Work Teams: Issues in the Implementation of Work Teams*, Vol. 4, M. Beyerlein, D. Johnson, and S. Beyerlein, Eds., JAI Press, Greenwich, CT, pp. 1–44.

Morgeson, F. P., Dierdorff, E. C., and Hmurovic, J. L. (2010), "Work Design In Situ: Understanding the Role of Occupational and Organizational Context," *Journal of Organizational Behavior*, Vol. 31, pp. 351–360.

Morgeson, F. P., Delaney-Klinger, K. A., Mayfield, M. S., Ferrara, P., and Campion, M. A. (2004), "Self-Presentation Processes in Job Analysis: A Field Experiment Investigating Inflation in Abilities, Tasks, and Competencies," *Journal of Applied Psychology*, Vol. 89, pp. 674–686.

Mundel, M. E. (1985), *Motion and Time Study: Improving Productivity*, 6th ed., Prentice-Hall, Englewood Cliffs, NJ.

Niebel, B. W. (1988), *Motion and Time Study*, 8th ed., Irwin, Homewood, IL.

O'Reilly, C., Parlette, G., and Bloom, J. (1980), "Perceptual Measures of Task Characteristics: The Biasing Effects of Differing Frames of Reference and Job Attitudes," *Academy of Management Journal*, Vol. 23, pp. 118–131.

O'Reilly, C. A., and Roberts, K. H. (1977), "Task Group Structure, Communication, and Effectiveness," *Journal of Applied Psychology*, Vol. 62, pp. 674–681.

Parker, S. K. (2003), "Longitudinal Effects of Lean Production on Employee Outcomes and the Mediating Role of Work Characteristics," *Journal of Applied Psychology*, Vol. 88, pp. 620–634.

Pasmore, W., Francis, C., and Haldeman, J. (1982), "Sociotechnical Systems: A North American Reflection on Empirical Studies of the Seventies," *Human Relations*, Vol. 35, pp. 1179–1204.

Pasmore, W. A. (1988), *Designing Effective Organizations: The Sociotechnical Systems Perspective*, Wiley, New York.

Pearson, R. G. (1971), "Human Factors Engineering," in *Industrial Engineering Handbook*, 3rd. ed., H. B. Maynard, Ed., McGraw-Hill, New York.

Perkins, A. L., and Abramis, D. J. (1990), "Midwest Federal Correctional Institution," in *Groups That Work* (*and Those That Don't*), J. R. Hackman, Ed., Jossey-Bass, San Francisco.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., Campion, M. A., Levin, K. Y., Mayfield, M. S., Morgeson, F. P., Pearlman, K., Gowing,

M. K., Lancaster, A., and Dye, D. (2001), "Understanding Work Using the Occupational Information Network (O*NET): Implications for Practice and Research," *Personnel Psychology*, Vol. 54, pp. 451–492.

Porter, L. W., Lawler, E. E., and Hackman, J. R. (1987), "Ways Teams Influence Individual Work Effectiveness," in *Motivation and Work Behavior*, 4th ed., R. M. Steers and L. W. Porter, Eds., McGraw-Hill, New York.

Posthuma, R. A., Morgeson, F. P., and Campion, M. A. (2002), "Beyond Employment Interview Validity: A Comprehensive Narrative Review of Recent Research and Trends over Time," *Personnel Psychology*, Vol. 55, pp. 1–81.

Rousseau, D. M. (1977), "Technological Differences in Job Characteristics, Employee Satisfaction, and Motivation: A Synthesis of Job Design Research and Sociotechnical Systems Theory," *Organizational Behavior and Human Performance*, Vol. 19, pp. 18–42.

Rousseau, D. M., Ho, V. T., and Greenberg, J. (2006), "I-Deals: Idiosyncratic Terms in Employment Relationships," *Academy of Management Review*, Vol. 31, pp. 977–994.

Salas, E., Dickinson, T. L., Converse, S. A., and Tannenbaum, S. I. (1992), "Toward an Understanding of Team Performance and Training," in *Teams: Their Training and Performance*, R. W. Swezey and E. Salas, Eds., Ablex, Norwood, NJ.

Salvendy, G., Ed. (1987), *Handbook of Human Factors*, Wiley, New York.

Salvendy, G., and Smith, M. J., Eds. (1981), *Machine Pacing and Occupational Stress*, Taylor & Francis, London.

Sanders, M. S., and McCormick, E. J. (1987), *Human Factors in Engineering and Design*, 6th ed., McGraw-Hill, New York.

Shaw, M. E. (1983), "Team Composition," in *Small Teams and Social Interaction*, Vol. 1, H. H. Blumberg, A. P. Hare, V. Kent, and M. Davies, Eds., Wiley, New York.

Shea, G. P., and Guzzo, R. A. (1987), "Teams as Human Resources," in *Research in Personnel and Human Resources*, Vol. 5., K. M. Rowland and G. R. Ferris, Eds., JAI Press, Greenwich, CT.

Sims, H. P., Szilagyi, A. D., and Keller, R. T. (1976), "The Measurement of Job Characteristics," *Academy of Management Journal*, Vol. 19, pp. 195–212.

Smith, A. (1776), *An Inquiry into the Nature and Causes of the Wealth of Nations*, reprinted by R. H. Campbell and A. S. Skinner, Eds., Liberty Classics, Indianapolis, IN.

Staw, B. M., and Boettger, R. D. (1990), "Task Revision: A Neglected Form of Work Performance," *Academy of Management Journal*, Vol. 33, pp. 534–559.

Steers, R. M., and Mowday, R. T. (1977), "The Motivational Properties of Tasks," *Academy of Management Review*, Vol. 2, pp. 645–658.

Steiner, I. D. (1972), *Group Process and Productivity*, Academic, New York.

Stevens, M. J., and Campion, M. A. (1994a), "Staffing Teams: Development and Validation of the Teamwork-KSA Test," paper presented at the Annual Meeting of the Society of Industrial and Organizational Psychology, Nashville, TN.

Stevens, M. J., and Campion, M. A. (1994b), "The Knowledge, Skill, and Ability Requirements for Teamwork: Implications for Human Resource Management," *Journal of Management*, Vol. 20, pp. 503–530.

Stevens, M. J., and Campion, M. A. (1999), "Staffing Work Teams: Development and Validation of a Selection Test for Teamwork Settings," *Journal of Management*, Vol. 25, pp. 207–228.

Sundstrom, E., DeMeuse, K. P., and Futrell, D. (1990), "Work Teams: Applications and Effectiveness," *American Psychologist*, Vol. 45, pp. 120–133.

Swezey, R. W., and Salas, E. (1992), *Teams: Their Training and Performance*, Ablex, Norwood, NJ.

Tannenbaum, S. I., Beard, R. L., and Salas, E. (1992), "Team Building and Its Influence on Team Effectiveness: An Examination of Conceptual and Empirical Developments," in *Issues, Theory, and Research in Industrial and Organizational Psychology*, K. Kelley, Eds., Elsevier, Amsterdam.

Taylor, F. W. (1911), *The Principles of Scientific Management*, Norton, New York.

Taylor, J. C. (1979), "Job Design Criteria Twenty Years Later," in *Design of Jobs*, 2nd ed., L. E. Davis and J. C. Taylor, Eds., Wiley, New York.

Thompson, J. D. (1967), *Organizations in Action*, McGraw-Hill, New York.

Tichauer, E. R. (1978), *The Biomechanical Basis of Ergonomics: Anatomy Applied to the Design of Work Situations*, Wiley, New York.

Turnage, J. J. (1990), "The Challenge of New Workplace Technology for Psychology," *American Psychologist*, Vol. 45, pp. 171–178.

Turner, A. N., and Lawrence, P. R. (1965), *Industrial Jobs and the Worker: An Investigation of Response to Task Attributes*, Harvard Graduate School of Business Administration, Boston, MA.

U.S. Department of Labor (1972), *Handbook for Analyzing Jobs*, U.S. Government Printing Office, Washington, DC.

U.S. Department of Labor (1977), *Dictionary of Occupational Titles*, 4th ed., U.S. Government Printing Office, Washington, DC.

U.S. Nuclear Regulatory Commission (1981), "Guidelines for Control Room Design Reviews," NUREG 0700, Nuclear Regulatory Commission, Washington, DC.

Vroom, V. H. (1964), *Work and Motivation*, Wiley, New York.

Walker, C. R., and Guest, R. H. (1952), *The Man on the Assembly Line*, Harvard University Press, Cambridge, MA.

Wall, T. B., Kemp, N. J., Jackson, P. R., and Clegg, C. W. (1986), "Outcomes of Autonomous Workgroups: A Long-Term Field Experiment," *Academy of Management Journal*, Vol. 29, pp. 280–304.

Warr, P., and Wall, T. (1975), *Work and Well-being*, Penguin, Harmondsworth, MD.

Weber, A., Fussler, C., O'Hanlon, J. F., Gierer, R., and Grandjean, E. (1980), "Psychophysiological Effects of Repetitive Tasks," *Ergonomics*, Vol. 23, pp. 1033–1046.

Welford, A. T. (1976), *Skilled Performance: Perceptual and Motor Skills*, Scott-Foresman, Glenview, IL.

Wicker, A., Kirmeyer, S. L., Hanson, L., and Alexander, D. (1976), "Effects of Manning Levels on Subjective Experiences, Performance, and Verbal Interaction in Groups," *Organizational Behavior and Human Performance*, Vol. 17, pp. 251–274.

Wilson, J. R., and Haines, H. M. (1997), "Participatory Ergonomics," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York.

Wong, C. S. (1989), "Task Interdependence: The Link between Task Design and Job Design," Doctoral Dissertation, Purdue University, West Lafayette, IN.

Wong, C. S., and Campion, M. A. (1991), "Development and Test of a Task Level Model of Job Design," *Journal of Applied Psychology*, Vol. 76, pp. 825–837.

Woodman, R. W., and Sherwood, J. J. (1980), "The Role of Team Development in Organizational Effectiveness. A Critical Review," *Psychological Bulletin*, Vol. 88, pp. 166–186.

Wrzesniewski, A., and Dutton, J. E. (2001), "Crafting a Job: Revisioning Employees as Active Crafters of Their Work," *Academy of Management Review*, Vol. 26, pp. 179–201.

Zajonc, R. B. (1965), "Social Facilitation," *Science*, Vol. 149, pp. 269–274.

Zander, A. (1979), "The Study of Group Behavior over Four Decades," *Journal of Applied Behavioral Science*, Vol. 15, pp. 272–282.

# CHAPTER **16**

# PERSONNEL RECRUITMENT, SELECTION, AND TURNOVER

**Jerry W. Hedge**
**RTI International**
**Research Triangle Park, North Carolina**

**Walter C. Borman**
**University of South Florida and Personnel Decisions Research Institutes, Inc.**
**Tampa, Florida**

**Dan Ispas**
**Illinois State University**
**Normal, Illinois**

## 1 INTRODUCTION

Major changes have taken place in the workplace over the last several decades and continue today. The globalization of numerous companies and industries, organizational downsizing and restructuring, expansion of information technology use at work, changes in work contracts, and increased use of alternative work strategies and schedules have transformed the nature of work in many organizations.

Technology alone changes the nature of work, and as Czaja (2001) has suggested, it will have a major impact on the future structure of the labor force, transforming the jobs that are available and how they are performed. Given the spreading use of technology in most occupations, it will likely create new jobs and opportunities for employment for some and eliminate jobs and create conditions of unemployment for other workers. It will also change the ways in which jobs are performed and alter job content and job demands.

The workforce itself is also changing, with a growing number of older workers, females, and dual-career couples. This means that organizations will need to

tailor their workplace policies to reflect a more diverse workforce.

More recently, economic conditions have also influenced how organizations approach attracting, hiring, and retaining employees. For example, in the United States alone, from January 2007 through December 2009, 6.9 million workers have been displaced from jobs they had held for at least three years (U.S. Department of Labor, 2010). Similar data from the Job Openings and Labor Turnover Survey, collected by the U. S Department of Labor, reflect the impact that a recession can have on the demand for labor and worker flows. Job openings (a measure of labor demand) and hires and separations (measures of worker flows) all declined during the 2007–2009 period and reached new lows in 2009 (deWolf and Klemmer, 2010).

These workforce changes are likely to result in different occupational and organizational structures in the future. A National Academy of Sciences report released in 2000 suggested that the nature of work is changing in ways that tend to blur the traditional distinctions between blue-collar and white-collar jobs (Committee

on Techniques for the Enhancement of Human Performance, 2000). For example, this committee suggested that blue-collar *production work* in many organizations is expanding to include more decision-making tasks than traditionally would have been part of a supervisory/managerial job. In addition, for some production workers, relatively narrow parameters of the job are giving way to broader involvement in work teams as well as interactions with external customers, clients, and patients. As team-based work structures have been used more widely, a number of studies have suggested that both cognitive and interactive skills are becoming more important in blue-collar jobs.

Technology is also having a significant impact on blue-collar jobs: for example, in some situations, replacing physical activity with mental and more abstract forms of responding. Generally, then, the implication of these changes is that information technology changes the mix of skills that are required, often creating jobs that require less sensory and physical skill and more "intellective" skills, such as abstract reasoning, inference, and cause–effect analysis (Committee on Techniques for the Enhancement of Human Performance, 2000).

For managerial jobs, the report suggested two interesting developments. First, at least lower level managers appear to be experiencing some loss in authority and control. Second, the need to communicate horizontally both within and across organizations may be becoming even more important than the supervision of an employee's work. There is also considerable discussion about the substantive content of managers' jobs, shifting toward the procurement and coordination of resources, toward coaching as opposed to commanding employees, and toward project management skills.

Within the service industry, the content of work is also evolving. First, a significant percentage of service jobs are becoming more routinized, in large measure because new information technologies enable greater centralization and control over work activities. Second, there is a tendency toward the blurring of sales and clerical jobs. Although the heterogeneity of work within specific service occupations appears to be increasing, this heterogeneity reflects, at least in part, the tendency to structure work differently according to market segments.

Interestingly, studies suggest a trend toward overall increase in technical skill requirements and cognitive complexity of service jobs. Although the initial impact of information technology involved a shift from manual to computer-mediated information processing, more recent applications involve the manipulation of a variety of software programs and databases. In addition, the rapid diffusion of access to the Internet has increased the potential for greater information-processing and cognitively complex activities. Finally, interpersonal interactions remain critical to service work, requiring skills in communications, problem solving, and negotiations.

With such changes becoming a more routine part of the work environment, to remain competitive, workers will likely need to upgrade their knowledge, skills, and abilities to avoid obsolescence, learning new systems and new activities at multiple points during their working lives. This may be particularly true for those workers who have been displaced from their jobs during economic downturns, only to find that those same jobs, once they are available again, may have different skill requirements from the old ones. In addition, over increased periods of unemployment, skills erode and behavior tends to change, leaving some people unqualified even for work they once did well (Peck, 2010). Because organizations increasingly operate in a wide and varied set of situations, cultures, and environments, not only may workers need to be more versatile and able to handle a wider variety of diverse and complex tasks, but employers will need to deal with an increasingly diverse workplace.

Recognizing that fundamentally new ways of thinking and acting will be necessary to meet the changing nature of work and worker requirements, organizations will be challenged to make wise and enduring decisions about how to attract, hire, and retain a skilled and motivated workforce. Historically, entry-level selection has centered on identifying skills important for performance early in a career. However, because finding and training workers in the future will be much more complex and costly than they are today, success on the job during and beyond the first few years will be increasingly important. The prediction of such long-term success indicators as retention and long-term performance will require the use of more complex sets of predictor variables that include such measures as personality, motivation, and vocational interest. To develop effective measures to predict long-term performance, it will be crucial to better understand the context of the workplace of the future, including the environmental, social, and group structural characteristics. Ultimately, combining the personal and organizational characteristics should lead to improved personnel selection models that go beyond the usual person–job relations, encouraging a closer look at theories of person–organization (PO) fit.

With the evolution of the workforce and the workplace in mind, the remainder of this chapter focuses on an examination of the evolving state of the science relevant to the future of worker recruitment, selection, and retention. We begin with a brief discussion of the activities required to identify the critical performance requirements of the job and the knowledge, skills, abilities, and other characteristics (KSAOs) that might be important to effective performance on the job.

## 2 JOB ANALYSIS AND JOB PERFORMANCE

### 2.1 Job Analysis

Job analysis identifies the critical performance requirements of the job and the KSAOs important to effective performance on the job; thus, it tells us what we should be looking for in a job candidate. Job analysis is a common activity with a well-defined methodology for conducting such analyses. As noted by Cascio (1995), terms such as *job element, task, job description*, and *job family* are well understood. Although job analysis continues to be an important first step in selection research and practice, the changing nature of work may suggest some movement in the future from a focus

on discrete job components to a more process-oriented approach.

A major development over the last decade or so in job analysis is a Department of Labor initiative to analyze virtually all jobs in the U.S. economy in order to build a database of occupational information (O*NET). This database may be used by organizations and individuals to help match people with jobs. The person–job (PJ) fit feature of the O*NET enables comparisons between personal attributes and targeted occupational requirements. There is also an organizational–characteristics component that facilitates PO matches. The hope is that O*NET will help unemployed workers and students entering the workforce to find more appropriate jobs and careers and employers to identify more highly qualified employees. These matches should be realized more systematically and with more precision than has been possible heretofore. An additional hope is that this initiative will encourage research that further advances the effectiveness of PJ matching, PO fit, and the science of personnel selection (Borman et al., 2003).

## 2.2 Job Performance Domain

A central construct of concern in work psychology is job performance, because performance criteria are often what we attempt to predict from our major interventions, including personnel selection, training, and job design. Traditionally, while most attention has focused on models related to predictors (e.g., models of cognitive ability, personality, and vocational interests), job performance models and research associated with them are beginning to foster more scientific understanding of criteria.

For example, Hunter (1983), using a path-analytic approach, found that cognitive ability has primarily a direct effect on individuals' acquisition of job knowledge. Job knowledge, in turn, influenced technical proficiency. Supervisory performance ratings were a function of both job knowledge and technical proficiency, with the job knowledge–ratings path coefficient three times as large as the technical proficiency–ratings coefficient. This line of research continued, with additional variables being added to the models. Schmidt et al. (1986) added job experience to the mix; they found that job experience had a direct effect on the acquisition of job knowledge and an indirect effect on task proficiency through job knowledge.

Later, Borman et al. (1991) included two personality variables, achievement and dependability, and behavioral indicators of achievement and dependability. The path model results showed that the personality variables had indirect effects on the supervisory performance ratings through their respective behavioral indicators. The best-fitting model also had paths from ability to acquisition of job knowledge, job knowledge to technical proficiency, and technical proficiency to the supervisory job performance ratings, arguably the most comprehensive measure of overall performance. Perhaps the most important result of this study was that the variance accounted for in the performance rating exogenous (dependent) variable increased substantially with the addition of personality and the behavioral indicators of personality beyond that found with previous models, including ability along with job knowledge and technical proficiency.

### 2.2.1 Task and Contextual Performance

Another useful way to divide the job performance domain has been according to task and contextual performance. Borman and Motowidlo (1993) argued that organization members may engage in activities that are not directly related to their main task functions but nonetheless are important for organizational effectiveness because they support the "organizational, social, and psychological context that serves as the critical catalyst for task activities and processes" (p. 71). Borman and colleagues have settled on a three-dimensional system: (1) personal support, (2) organizational support, and (3) conscientious initiative (Coleman and Borman, 2000; Borman et al., 2001). The notion is to characterize the citizenship performance construct according to the recipient or target of the behavior: other persons, the organization, and oneself, respectively.

Thus, job performance criteria have always been important in personnel selection research, but a recent trend has been to study job performance in its own right in an attempt to develop substantive models of performance. The vision has been to learn more about the nature of job performance, including its components and dimensions, so that performance itself, as well as predictor–performance links, will be better understood. Again, if we can make more progress in this direction, the cumulative evidence for individual predictor construct/performance construct relationships will continue to progress and the science of personnel recruitment, selection, and retention will be enhanced substantially.

## 3 PERSONNEL RECRUITMENT

Personnel recruitment is an important first step in the science and practice of matching employer needs with individual talent. Prior to the economic downturn that developed in the latter half of the current decade, labor markets in many countries had become increasingly tight, requiring organizations to compete to attract talented jobseekers and fill job openings. This convinced many organizations of the importance of recruitment and led many organizations to increase their recruiting efforts to attract prospective applicants. In spite of recent economic conditions, an increased awareness of the importance of the positive impact of recruitment continues today.

In terms of filling vacancies, organizations have a choice between internal and external recruitment strategies. According to Ployhart et al. (2006), internal recruitment is generally preferred by organizations because of its reduced costs: reduced recruitment, socialization, and training costs; increased probability of success as the employee has already been deemed successful by the organization's performance appraisal; and reduced start-up time because the employee is already familiar with the organization and its policies.

When it comes to external recruitment, organizations have choices with regard to the sources they use.

### 3.1 Recruitment Sources

Considerable research has focused on the effectiveness of various recruitment methods, that is, how the organization is making potential applicants aware of the job opening (Breaugh et al., 2008). Zottoli and Wanous (2000) reviewed studies conducted on recruitment methods and found that referrals (individuals referred by a current employee) and direct applicants (walk-ins) were associated with lower turnover. More recently, Yakubovich and Lup (2006) found that the performance level of the referrer can be important such that referrals from higher performing employees were not only viewed by human resources (HR) personnel as having better qualifications, but also scored higher on an objective selection measure.

### 3.2 Recruiter Effects

Research on the effects of the recruiter on recruitment outcomes has been summarized by Chapman et al. (2005). They found that recruiter personableness was a strong predictor of job pursuit intentions ($r = 0.50$). Similarly, recruiter competence, trustworthiness, and informativeness were related to applicants' attraction to the organization. No other recruiter characteristics appear to be important; however, results of the Chapman et al. (2005) meta-analysis are based on a small number of studies and thus more research is needed.

### 3.3 Internet Recruitment

The topic of Internet recruitment has become very popular in the recruitment literature in recent years. This is not surprising considering that HR practitioners see the Internet and websites as effective recruitment sources (Chapman and Webster, 2003). Websites allow organizations to reach larger pools of applicants; however, as a downside, they can lead to adverse impact and privacy issues (Stone et al., 2005) and can attract a large number of unqualified applicants. Website features perceived as important to applicants are aesthetics, content, and their purpose (Breaugh et al., 2008). Online job boards (such as Monster.com and CareerBuilder.com) are also widely used by organizations and job applicants. Jattuso and Sinar (2003) found that industry/position-specific job boards appear to result in higher quality applicants.

### 3.4 Summary

The last decade has witnessed an increase in attention to personnel recruitment strategies. Research suggests that the recruitment method(s) used plays a role, as do certain personal characteristics of the recruiter in recruitment success. In particular, applicant perceptions of the recruiter's competence, trustworthiness, personableness, and amount of information offered are important variables. In addition, given the ubiquitous nature of the Internet in the lives of most applicants and organizations, its use has quickly become a major tool in recruitment efforts and should continue for the foreseeable future.

## 4 PERSONNEL SELECTION

An understanding of current thinking in personnel selection requires a focus on predictor measurement—in particular, ability, personality, vocational interests, and biodata—as well as a discussion of an alternative model to the traditional PJ fit selection strategy: namely, PO fit.

### 4.1 Predictor Measurement

#### 4.1.1 Ability

Abilities are relatively stable individual differences that are related to performance on some set of tasks, problems, or other goal-oriented activities (Murphy, 1996). Another definition, offered by Carroll (1993), conceptualizes abilities as relatively enduring attributes of a person's capability for performing a particular range of tasks. Although the term *ability* is widely used in both the academic and applied literature, several other terms have been related loosely to abilities. For example, the term *competency* has been used to describe individual attributes associated with the quality of work performance. In practice, lists of competencies often include a mixture of knowledges, skills, abilities, motivation, beliefs, values, and interests (Fleishman et al., 1999).

Another term that is often confused in the literature with abilities is *skills*. Whereas abilities are general traits inferred from relationships among performances of individuals observed across a range of tasks, skills are more dependent on learning and represent the product of training on particular tasks. In general, skills are more situational, but the development of a skill is, to a large extent, predicted by a person's possession of relevant underlying abilities, usually mediated by the acquisition of the requisite knowledge. That is, these underlying abilities are related to the rate of acquisition and final levels of performance that a person can achieve in particular skill areas (Fleishman et al., 1999).

Ability tests usually measure mental or cognitive ability but may also measure other constructs, such as physical abilities. In the following sections we discuss recent developments and the most current topics in the areas of physical ability testing, cognitive ability, and practical intelligence.

***Physical Abilities*** One important area for selection into many jobs that require manual labor or other physical demands is the use of physical ability tests. Most physical ability tests are performance tests (i.e., not paper and pencil) that involve demonstration of attributes such as strength, cardiovascular fitness, or coordination. Although physical ability tests are reported to be used widely for selection (Hogan and Quigley, 1994), not much new information has been published in this area in the past few years. In one study, Blakley et al. (1994) provided evidence that isometric strength tests are valid predictors across a variety of different physically demanding jobs and females scored substantially lower than males on these isometric strength tests. In light of these findings, there is a recent and growing interest in reducing adverse impact through pretest preparation. Hogan and Quigley (1994)

demonstrated that participation in a physical training program can improve females' upper body strength and muscular endurance and participation in a pretest physical training program was significantly related to the likelihood of passing a firefighter physical ability test.

***Cognitive Ability*** In the early twentieth century, Cattell, Scott, Bingham, Viteles, and other applied psychologists used cognitive ability tests to lay the foundation for the current practice of personnel selection (Landy, 1993). Recent attention has focused on the usefulness of general cognitive ability *g* versus more specific cognitive abilities for predicting training and job performance and the contributions of information-processing models of cognitive abilities for learning more about ability constructs. First, there has been a debate concerning the "ubiquitousness" of the role of general cognitive ability, or *g*, in the prediction of training and job performance. Several studies have demonstrated that psychometric *g*, generally operationalized as the common variance in a battery of cognitive ability tests (e.g., the first principal component), accounts for the majority of the predictive power in the test battery and the remaining variance (often referred to in this research as "specific abilities") accounts for little or no additional variance in the criterion (e.g., Ree et al., 1994; Larson and Wolfe, 1995).

Other researchers have expressed concern related to the statistical model often used to define *g*. A general factor or *g* represents the correlations between specific ability tests, so specific abilities will, by definition, be correlated with the general factor. Thus, it could be argued that it is just as valid to enter specific abilities *first* and then say that *g* does not contribute beyond the prediction found with specific abilities alone (e.g., Murphy, 1996). In fact, Muchinsky (1993) found this to be the case for a sample of manufacturing jobs, where mechanical ability was the single best predictor of performance and an intelligence test had no incremental validity beyond the mechanical test alone.

We know very little about specific abilities when they are defined as the variance remaining once a general factor is extracted statistically. Interestingly, what little information is available suggests that these specific-ability components tend to be most strongly related to cognitive ability tests that have a large knowledge component (e.g., aviation information) (Olea and Ree, 1994). This is consistent with previous research showing that job knowledge tests tend to be slightly more valid than ability tests (Hunter and Hunter, 1984) and also with research demonstrating that job knowledge appears to mediate the relationship between abilities and job performance (e.g., Borman et al., 1993). Meta-analysis has demonstrated the generality of job knowledge tests as predictors of job performance (Dye et al., 1993). In addition, these authors found that the validity of job knowledge tests was moderated by job complexity and by job–test similarity, with validities significantly higher for studies involving high-complexity jobs and those with high job–test similarity.

***Practical Intelligence*** Sternberg and colleagues have attempted to broaden the discussion of general intelligence (Sternberg and Wagner, 1992). Based on a triarchic theory of intelligence, Sternberg (1985) suggested that practical intelligence and tacit knowledge play a role in job success. Practical intelligence is often described as the ability to respond effectively to practical problems or demands in situations that people commonly encounter in their jobs (Wagner and Sternberg, 1985; Sternberg et al., 1993). Conceptually, practical intelligence is distinct from cognitive ability. In fact, some research (Chan, 2001) has found measures of practical intelligence to be uncorrelated with traditional measures of cognitive ability.

Sternberg and his colleagues have repeatedly found significant correlations and some incremental validity (over general intelligence) for measures of tacit knowledge in predicting job performance or success (Sternberg et al., 1995). Tacit knowledge has been shown to be trainable and to differ in level according to relevant expertise. Certainly, tacit knowledge measures deal with content that is quite different from that found in traditional job knowledge tests (e.g., knowledge related to managing oneself and others).

***Summary*** A cumulation of 85 years of research demonstrates that if we want to hire people without previous experience in a job, the most valid predictor of future performance is general cognitive ability (Schmidt and Hunter, 1998). General cognitive ability measures have many advantages in personnel selection: (1) they show the highest validity for predicting training and job performance, (2) they may be used for all jobs from entry level to advanced, and (3) they are relatively inexpensive to administer. In addition, there is some evidence that measures of "practical intelligence" or tacit knowledge may under certain conditions provide incremental validity beyond general cognitive ability for predicting job performance (Sternberg et al., 1995). Finally, physical ability tests may be useful in predicting performance for jobs that are physically demanding.

### 4.1.2 Personality

Interest in personality stems from the desire to predict the motivational aspects of work behavior. Nevertheless, until recently, the prevalent view was that personality variables were a dead end for predicting job performance. Some of the factors fueling this belief were (1) the view that a person's behavior is not consistent across situations and thus that traits do not exist, (2) literature reviews concluding that personality variables lack predictive validity in selection contexts, and (3) concern about dishonest responding on personality inventories.

However, by the late 1980s, favorable opinions about personality regarding personnel selection began to grow (Hogan, 1991). Evidence accumulated to refute the notion that traits are not real (Kenrick and Funder, 1988) or stable (Conley, 1984). Research showed at least modest validity for some personality traits in predicting job performance (e.g., Barrick and Mount, 1991; McHenry et al., 1990; Ones et al., 1993). Further, evidence mounts that personality measures produce small, if any, differences between majority and protected

classes of people (Ones and Viswesvaran, 1998b) and that response distortion does not necessarily destroy criterion-related validity (e.g., Ones et al., 1996; Ones and Viswesvaran, 1998a).

Today's well-known, hierarchical, five-factor model (FFM) of personality (alternatively, "the Big Five") was first documented in 1961 by Tupes and Christal (see Tupes and Christal, 1992). The five factors were labeled *surgency*, *agreeableness*, *dependability*, *emotional stability*, and *culture*. Following Tupes and Christal, McCrae and Costa (1987) replicated a similar model of the FFM. In their version, *extraversion* is comprised of traits such as talkative, assertive, and active; *conscientiousness* includes traits such as organized, thorough, and reliable; *agreeableness* includes the traits kind, trusting, and warm; *neuroticism* includes traits such as nervous, moody, and temperamental; and *openness* incorporates such traits as imaginative, curious, and creative (Goldberg, 1992).

A large amount of evidence supports the generalizability and robustness of the Big Five. Others argue that the theoretical value and the practical usefulness of the Big Five factors are severely limited by their breadth *and* important variables are missing from the model. From an applied perspective, important questions pertain to the criterion-related validity of personality variables and the extent to which it matters whether we focus on broad factors or narrower facets.

Although the Big Five model of personality is not universally accepted, considerable research has been conducted using this framework. For example, Barrick et al. (2001) conducted a second-order meta-analysis including 11 meta-analyses of the relationship between Big Five personality dimensions and job performance. They included six performance criteria and five occupational groups. Conscientiousness was a valid predictor across all criteria and occupations ($\rho$ values of 0.19–0.26), with the highest overall validity of the Big Five dimensions. Emotional stability was predictive for four criteria and two occupational groups. Extraversion, agreeableness, and openness did not predict overall job performance, but each was predictive for some criteria and some occupations. Barrick et al.'s results echoed conclusions of previous research on this topic (e.g., Barrick and Mount, 1991; Hurtz and Donovan, 2000).

Despite the low to moderate magnitude of the Big Five's predictive validity, optimism regarding personality's usefulness in selection contexts remains high. One reason is that even a modestly predictive variable, if uncorrelated with other predictors, offers incremental validity. Personality variables tend to be unrelated to tests of general cognitive ability ($g$) and thus they have incremental validity over $g$ alone (Day and Silverman, 1989; Hogan and Hogan, 1989; McHenry et al., 1990; Schmitt et al., 1997). For example, conscientiousness produces gains in validity of 11–18 % compared to using $g$ alone (Salgado, 1998; Schmidt and Hunter, 1998). Salgado (1998) reported that emotional stability measures produced a 10 % increment in validity over $g$ for European civilian and military samples combined. For military samples alone, the incremental validity was 38 %.

**Compound Traits: Integrity, Adaptability, and Core Self-Evaluation** In contrast to considering links between narrow personality traits and individual job performance criteria, a very different approach for personality applications in personnel selection is the development of *compound traits*. Compound traits have the potential to show even stronger relationships with criteria because they are often constructed by identifying the criterion first and then selecting a heterogeneous group of variables expected to predict it (Hogan and Ones, 1997). Integrity, adaptability, and core self-evaluation are three compound traits that may be especially useful in a selection context.

Integrity consists of facets from all Big Five factors: mainly conscientiousness, agreeableness, and emotional stability (Ones et al., 1994). Integrity tests have several advantages as part of selection systems, including considerable appeal to employers. Because of the enormous costs of employee theft and other counterproductive behaviors (U.S. Department of Health and Human Services, 1997; Durhart, 2001), it is understandable that employers want to avoid hiring dishonest applicants. Also, paper-and-pencil integrity measures have grown in popularity since the 1988 Federal Polygraph Protection Act banned most preemployment uses of the polygraph test. Empirical evidence shows that integrity tests are better than any of the Big Five at predicting job performance ratings ($\rho = 0.41$) (Ones et al., 1993). Under some conditions, integrity tests also predict various counterproductive behaviors at work. Integrity tests are uncorrelated with tests of $g$ (Hogan and Hogan, 1989; Ones et al., 1993) and produce a 27 % increase in predictive validity over $g$ alone (Schmidt and Hunter, 1998).

Adaptive job performance has become increasingly important in today's workplace. Existing personality scales carrying the adaptability label tend to be narrowly focused on a particular aspect of adaptability (e.g., International Personality Item Pool, 2001) and would probably not be sufficient to predict the multiple dimensions of adaptive job performance. One would expect that someone who readily adapts on the job is likely to be patient, even tempered, and confident (facets of emotional stability); open to new ideas, values, and experiences (facets of openness); and determined to do what it takes to achieve goals (a facet of conscientiousness). It is possible that high levels of other aspects of conscientiousness, such as dutifulness and orderliness, are detrimental to adaptive performance because they involve overcommitment to established ways of functioning. Again, although speculative, these relationships are in line with the results of several different studies that linked personality variables to adaptive performance (e.g., Mumford et al., 1993; Le Pine et al., 2000).

Core self-evaluation (CSE) is a fundamental, global appraisal of oneself (Judge et al., 1998). We should note that, whereas some have called CSE a compound trait, Judge and his colleagues might not agree with this characterization. Erez and Judge (2001) described CSE as a single higher order factor explaining the association among four more specific traits: self-esteem, generalized self-efficacy, internal locus of control, and

emotional stability. A meta-analysis showed that CSE's constituent traits are impressively predictive of overall job performance (Judge and Bono, 2001). Estimated $\rho$ values corrected for sampling error and criterion unreliability ranged from 0.19 (emotional stability) to 0.26 (self-esteem).

CSE is even more strongly related to job satisfaction than it is to overall job performance. Predictive validity estimates ranged from 0.24 (emotional stability) to 0.45 (self-efficacy) (Judge and Bono, 2001). People with high CSE seem to seek out challenging jobs and apply a positive mindset to the perception of their jobs (Judge et al., 1998, 2000). The result is desirable job characteristics, both real and perceived, which contribute to high job satisfaction. The link between CSE and job satisfaction has important implications, because satisfied employees are more likely to stay on the job than are dissatisfied employees (Tett and Meyer, 1993; Harter et al., 2002). Low turnover rates are especially important in organizations that invest heavily in the training of new employees.

**Summary**  Over the past two decades, personality has enjoyed a well-deserved resurgence in research and applied use, aided in part by the wide acceptance of the Big Five model of personality. These broad traits offer incremental predictive validity over cognitive ability alone (Salgado, 1998; Schmidt and Hunter, 1998). Using personality predictors narrower than the Big Five shows promise for revealing stronger criterion-related validities, especially when the criteria are relatively specific. Compound traits may be especially important in organizational environments that are more dynamic and team oriented (Edwards and Morrison, 1994). Complex, heterogeneous predictors are needed to predict the complex, heterogeneous performance criteria that are likely in such an environment.

Recent research has not only shown the criterion-related validity of personality but also addressed certain persistent objections to personality testing. Adverse impact appears not to be a problem. Response distortion, although certainly problematic in selection settings, may be alleviated with targeted interventions, although it still may be problematic with small selection ratios (i.e., a relatively low ratio of selectees) (Rosse et al., 1998). Also, considerable evidence exists to suggest that personality is impressively stable over the entire life span (Roberts and Del Vecchio, 2000). Thus, we can be confident that personality variables used in selection will predict performance of selectees throughout their tenure with an organization.

### 4.1.3 Vocational Interest

Needs, drives, values, and interests are closely related motivational concepts that refer to the intentions or goals of a person's actions. Interests are generally thought of as the most specific and least abstract of these concepts (Hogan and Blake, 1996). Hogan and Blake pointed out that vocational interests have not often been studied in relation to other motivational constructs. Holland and Hough (1976) suggested that a likely reason for this lack of attention to theoretical links with other constructs is the early empirical successes of vocational interest inventories, predicting outcomes such as vocational choice. In a sense, there was little reason to relate to the rest of psychology because of these successes. Accordingly, many psychologists have regarded the area of vocational interest measurement as theoretically and conceptually barren (Strong, 1943).

The most obvious link between vocational interests and relevant criteria appears to be between vocational interest responses and occupational tenure and, by extension, job satisfaction. For the most part, the tenure relation has been confirmed. For example, regarding occupational tenure, Strong (1943) demonstrated that occupational membership could be predicted by vocational interest scores on the Strong Vocational Interest Blank administered between 5 and 18 years previously. This finding at least implies that persons suited for an occupation on the basis of their interests tend to gravitate to and stay in that occupation.

For job satisfaction, the relationships are more mixed, but at best, the vocational interest/job satisfaction correlations are moderate (0.31). Vocational interests are not usually thought of in a personnel selection context, and in fact, there are not many studies linking vocational interests and job performance. Those that do exist find a median validity for interest predictors against job performance around 0.20. Although this level of validity is not very high and the number of studies represented is not large, the correlation observed compares favorably to the validities of personality constructs (e.g., Barrick and Mount, 1991).

Analogous to personality measures, vocational interest inventories used for selection have serious potential problems with slanting of responses or faking. It has long been evident that people *can* fake interest inventories (e.g., Campbell, 1971). However, some research has shown that in an actual selection setting applicants may not slant their responses very much (e.g., Abrahams et al., 1971).

Although the personality and vocational interest may be closely linked conceptually, it is evident that inventories measuring the two constructs are quite different. Personality inventories present items that presumably reflect the respondent's tendency to act in a certain way in a particular situation. Vocational interest items elicit like–dislike responses to objects or activities. There has been a fair amount of empirical research correlating personality and vocational interest responses. Hogan and Blake (1996) summarized the findings of several studies linking personality and vocational interests at the level of the Big Five personality factors and the six Holland types. Although many of these correlations are significant, the magnitude of the relationships is not very large. Thus, it appears that personality constructs and vocational interest constructs have theoretically and conceptually reasonable and coherent relationships but that these linkages are relatively small. What does this mean for selection research? Overall, examining vocational interests separately as a predictor of job performance (*and* job satisfaction and attrition) seems warranted.

**Summary**  Vocational interest measures, similar to personality measures, tap motivation-related constructs.

Interests have substantial conceptual similarity to personality, but empirical links between the two sets of constructs are modest. Vocational interest measures are most often used in counseling settings and have been linked primarily to job satisfaction criteria. However, although not often used in a selection context, limited data suggest reasonable levels of validity. Accordingly, vocational interests may show some promise for predicting job performance.

### 4.1.4 Biodata

The primary principle behind the use of biodata is that the best predictor of future behavior is past behavior. In fact, biodata offer a number of advantages when used in personnel selection. Among the most significant is their power as a predictor across a number of work-related criteria. For example, in a meta-analytic review of over 85 years of personnel psychology research, Schmidt and Hunter (1998) reported mean biodata validity coefficients of 0.35 and 0.30 against job and training success, respectively. These findings support previous research reporting validities ranging from 0.30 to 0.40 between biodata and a range of criteria, such as turnover, absenteeism, job proficiency, and performance appraisal ratings (e.g., Asher, 1972; Reilly and Chao, 1982; Hunter and Hunter, 1984). Based on these meta-analytic results, researchers have concluded that biographical inventories have almost as high validities as cognitive ability tests (Reilly and Chao, 1982). In addition, research indicates that biodata show less adverse impact than that of cognitive ability tests (Wigdor and Garner, 1982). Importantly, the high predictability associated with biodata, the ease of administration of biodata instruments, the low cost, and the lack of adverse impact have led to the widespread use of biodata in both the public and private sectors (Farmer, 2001).

Mael (1991) reviewed certain ethical and legal concerns that have been raised about biodata. The first of these deals with the controllability of events. That is, there are actions that respondents choose to engage in (controllable events), whereas other events either are imposed upon them or happened to them (noncontrollable events). Despite the belief held by numerous biodata researchers that all events, whether or not controllable, have the potential to influence later behavior, some researchers (e.g., Stricker, 1987, 1988) argue that it is unethical to evaluate individuals based on events that are out of their control (e.g., parental behavior, socioeconomic status). As a result, some have either deleted all noncontrollable items from their biodata scales or created new measures with the exclusion of these items. A frequent consequence, however, is that using only controllable items reduces the validity of the biodata instrument (Mael, 1991).

Two other ethical and legal concerns that have been raised are equal accessibility and invasion of privacy. That is, some researchers (e.g., Stricker 1987, 1988) argue that items dealing with events not equally accessible to all individuals (e.g., playing varsity football) are inherently unfair and should not be included. Similarly, the current legal climate does not encourage the use of items perceived as personally invasive. Overall,

minimizing such issues as invasiveness might be encouraged. What should be especially avoided, however, is a reliance on subjective and less verifiable items that compromise the primary goal of retrieving relatively objective, historical data from applicants.

**Summary**   Biodata predictors are a powerful noncognitive alternative to cognitive ability tests that have shown significant promise as a predictor in selection. The principle relative to biodata is that past behaviors matter and should be taken into account when criteria such as performance, absenteeism, and other work-related outcomes are being predicted. In addition, efforts are currently under way to develop a more theoretical understanding of the constructs involved with biodata. Finally, ethical and legal concerns are being addressed in hopes of creating an acceptable compromise between high predictability and overall fairness. Thus, although there is still much to be done in understanding how past behaviors can be used in personnel selection, evidence suggests that enhancing biodata techniques seems like a step in the right direction. Some examples of published tests that measure many of the predictor constructs discussed in the preceding section are given in Table 1.

## 4.2   Person–Organization Fit

Conventional selection practices are geared toward hiring employees whose KSAOs provide the greatest fit with clearly defined requirements of specific jobs. The characteristics of the organization in which the jobs reside, those characteristics of the person relative to the organization as a whole, are rarely considered. The basic notion with PO fit is that a fit between personal attributes and characteristics of the target organization contributes to important positive individual and organizational outcomes.

### 4.2.1   Attraction–Selection–Attrition

Much of the recent interest in the concept of PO fit can be traced to the attraction–selection–attrition (ASA) framework proposed by Schneider (e.g., Schneider, 1987, 1989; Schneider et al., 2000). Schneider (1987) outlined a theoretical framework of organizational behavior based on the mechanism of person–environment fit that integrates both individual and organizational theories. It suggests that certain types of people are attracted to and prefer particular types of organizations; organizations formally and informally seek certain types of employees to join the organization; and attrition occurs when employees who do not fit a particular organization leave. Those people who stay with the organization, in turn, define the structure, processes, and culture of the organization.

Van Vianen (2000) argued that, although many aspects of organizational life may be influenced by the attitudes and personality of the employees in the organization, this does not necessarily require that the culture of a work setting originate in the characteristics of people. He suggested, instead, that cultural dimensions reflecting the human side of organizational life are more adaptable to characteristics of people, whereas cultural

**Table 1 Sampling of Tests That Measure Predictor Constructs**

| Test | Publisher | Construct/Focus |
|---|---|---|
| Achievement and Success Index | LIMRA | Biodata |
| Adaptability Test | Pearson Reid London House | Adaptability |
| Basic Skills Test | Psychological Services, Inc. | Cognitive abilities |
| California Psychological Inventory | CPP | Personality |
| Differential Aptitude Tests | Harcourt Assessment | Cognitive abilities |
| Employee Aptitude Survey | Harcourt Assessment | Cognitive abilities |
| Hogan Personality Inventory | Hogan Assessment Systems | Personality |
| Management Interest Inventory | SHL | Managerial interests |
| NEO Five-Factor Inventory | PAR | Personality |
| Myers–Briggs Type Indicator | CPP | Personality |
| MPTQ | Proctor & Gamble | Biodata |
| Physical Ability Test | FSI | Physical abilities |
| Self-Directed Search | PAR | Vocational interest |
| Strong Interest Inventory | CPP | Vocational interest |
| Watson–Glaser Critical Thinking | Harcourt Assessment | Cognitive abilities |

dimensions that reflect the production side of organizational life are more determined by organizational goals and the external environment.

Similarly, Schaubroeck et al. (1998) proposed that a more complex conceptualization of the ASA process that incorporates the distinction between occupational and organizational influences should be examined more closely. These researchers investigated the role of personality facets and PO fit and found that personality homogenization occurs differently and more strongly within particular occupational subgroups within an organization. Similarly, Haptonstahl and Buckley (2002) suggested that as work teams become more widely used in the corporate world, person–group (PG) fit becomes an increasingly relevant construct.

### 4.2.2 Toward an Expanded Model of Fit and a Broader Perspective of Selection

The research and theorizing reported in this section suggest that selection theory should consider making fit assessments based PJ fit, PO fit, and PG fit. Traditional selection theory considers PJ fit as the basis for selecting job applicants, with the primary predictor measures being KSAOs and the criterion targets being job proficiency and technical understanding of the job.

To include PO fit as a component of the selection process, one would evaluate applicants' needs, goals, and values. The assumptions here are that the greater the match between the needs of the applicant and organizational reward systems, the greater the willingness to perform for the organization, and the greater the match between a person's goals and values and an organization's expectations and culture, the greater the satisfaction and commitment.

Finally, at a more detailed level of fit, there is the expectation that suborganizational units such as groups may have different norms and values than the organization in which they are embedded. Thus, the degree of fit between an individual and a group may differ significantly from the fit between the person and the organization. PG fit has not received as much research attention as either PJ fit or PO fit, but it is clearly different from these other types of fit (Kristof, 1996; Borman et al., 1997).

Werbel and Gilliland (1999) suggested the following tenets about the three types of fit and employee selection: (1) the greater the technical job requirements, the greater the importance of PJ fit; (2) the more distinctive the organizational culture, the greater the need for PO fit; (3) the lengthier the career ladder associated with an entry-level job, the greater the importance of PO fit; (4) the more frequent the use of team-based systems within a work unit, the greater the importance of PG fit; and (5) the greater the work flexibility within the organization, the greater the importance of PO fit and PG fit.

### 4.2.3 Summary

When jobs and tasks are changing constantly, the process of matching a person to some fixed job requirements becomes less relevant. Whereas traditional selection models focused primarily on PJ fit, several have argued that with new organizational structures and ways of functioning individual–organizational fit and individual–group fit become more relevant concepts. In our judgment, selection models in the future should incorporate all three types of fit as appropriate for the target job and organization. We can conceive of a hybrid selection model where two or even all three types of fit are considered simultaneously in making selection decisions. For example, consider the possibility of a special multiple-hurdle application in which, in order to be hired, applicants must have above a level of fit for the initial job, the team to which they will first be assigned, *and* the target organization. *Or*, depending on the particular selection context for an organization, the three types of fit might be weighted differentially in selecting applicants. Obviously, the details of hybrid selection models such as these have yet

to be worked out. However, the notion of using more than one fit concept seems to hold promise for a more flexible and sophisticated approach to making selection decisions.

## 5  PERSONNEL TURNOVER

The retention of high-quality employees is a strategic priority for contemporary organizations, as turnover can impact firm performance (Glebbeek and Bax, 2004; Kacmar et al., 2006), quality of customer service (Schlesinger and Heskett, 1991), and turnover among remaining employees. In addition, according to Cascio (2000), other turnover costs include *separation costs* (e.g., costs of the exit interview, separation benefits, vacancy costs), *replacement costs* (costs of replacing the leaver: recruitment and selection), and *training costs* (orientation and training for new employees). Regarding a typology of turnover, we can distinguish between involuntary and voluntary turnover (Hom and Griffeth, 1995). Involuntary turnover refers to job separations initiated by the organization (such as firings and layoffs). Voluntary turnover refers to turnover initiated by the employee. A further differentiation of voluntary turnover can be made between functional (turnover of low performers) and dysfunctional (the turnover of high performers).

### 5.1  Reducing Voluntary Turnover

Efforts to reduce turnover can begin before employees enter the organization in the recruitment and personnel selection stages. In the recruitment stage, a useful tool for reducing turnover is the realistic job preview (RJP). RJPs are comprehensive profiles of both the positive and negative features of a job presented by the organization to prospective or new employees. RJPs are hypothesized to work by reducing the unrealistic expectations new employees may have. In a meta-analytic review, Philips (1998) found that RJPs were successful in reducing attrition from the recruitment process, fostering accurate initial expectations, and reducing all types of turnover once on the job, including voluntary turnover. In addition, RJPs were related to job performance, although the mean correlation was only 0.05. RJPs are considered most effective when given in the recruitment stage (Philips, 1998; Wanous, 1980), although they can be administered to new employees during the socialization stage as well.

Efforts to reduce turnover can also be made in the personnel selection stage. For example, in a rare longitudinal study using actual job applicants, Barrick and Zimmerman (2005) found that self-confidence and decisiveness and a biographical measure (consisting of tenure in the previous job, the number of family and friends in the organization, and whether the applicant was referred by a current employee of the organization) were predictive of voluntary turnover six months later. Meta-analytic reviews confirmed that weighted application blanks are predictive of employee turnover (Hom and Griffeth, 1995; Griffeth et al., 2000). Similarly, Barrick and Mount (1996) found that conscientiousness was negatively related to employee turnover. Carefully developed structured employment interviews have also been found to predict turnover, with a corrected correlation of 0.39 for an empirically based telephone administered interview (Schmidt and Rader, 1999).

A variety of interventions can be used to reduce turnover after employees have joined the organization. McEvoy and Cascio (1985) conducted a meta-analysis of the strategies used by organizations to reduce turnover. In addition to RJPs discussed above, they found that job enrichment strategies (such as increasing decision making, task variety, and autonomy) were successful in reducing turnover. In fact, they found that job enrichment strategies were about twice as effective as RJPs at reducing turnover. An intervention to improve retention in the U.S. Army involved a unit retention climate feedback system that surveyed unit members about their shared perceptions relevant to soldier retention decisions in the unit. Part of the system was a unit leadership feedback report based on the survey responses that provided actionable guidance for leaders to use to improve the unit's retention climate to in turn enhance reenlistement rates (Kubisiak et al., 2009). Turnover can also be reduced indirectly by increasing employees' job satisfaction and organizational commitment (Griffeth et al., 2000).

Research has begun to recognize that outside work factors (e.g., factors related to the community) can impact voluntary turnover. About 10 years ago, Mitchell et al. (2001) introduced job embeddedness as a construct focused on the factors that influence a person's staying on a job. Job embeddedness has three dimensions: links, fit, and sacrifice, which are relevant to both the organization and the community. Job embeddedness has been found to be predictive of voluntary turnover (Mitchell et al., 2001). More importantly, Mitchell et al. (2001) found that job embeddedness has incremental validity over measures of job satisfaction, organizational commitment, and perceived alternatives. Lee et al. (2004) distinguished between two major subdimensions of job embeddedness: on the job and off the job. They report that only off-the-job embeddedness (e.g., links to and fit with the community) was predictive of voluntary turnover.

### 5.2  Summary

Because of the financial costs and the potential loss of critical organizational knowledge when valued employees leave the organization, personnel turnover and retention are a critical focus for organizations. Research points to the importance of prospective employers offering a realistic job preview to applicants as a vital first step in strengthening the employee–employer relationship. Using proper predictor measures that allow identification of applicants with the necessary job-related attributes is also crucial. Beyond use of these recruitment and selection strategies, creating the right climate where employees feel they will be able to work productively and develop professionally will increase the chances that employees will choose to remain with the organization rather than seek employment elsewhere.

## 6 CONCLUSION

The world of work is in the midst of profound changes that will require new thinking about personnel selection. Workers will probably need to be more versatile, handle a wider variety of diverse and complex tasks, and have more sophisticated technological knowledge and skills. The aim of this chapter was to provide a review of the state of the science on personnel recruitment, selection, and turnover research and thinking. Accordingly, we reviewed research on performance criteria and provided a review of recruitment strategies; predictor space issues, including predictors that tap such domains as ability, personality, vocational interests, and biodata; selection issues related to PJ match and PO fit; and, finally, turnover and retention strategies.

Relative to criterion measurement, research has demonstrated that short-term, technical performance criteria are best predicted by general cognitive ability, whereas longer term criteria such as nontechnical job performance, retention, and promotion rates are better predicted by other measures, including personality, vocational interest, and motivation constructs. To select and retain the best possible applicants, it would seem critical to understand, develop, and evaluate multiple measures of short- and long-term performance as well as other indicators of organizational effectiveness, such as turnover/retention.

Recruitment emphasis has increased over the last decade, with research noting the importance of recruitment methods and recruiter characteristics. In addition, organizations are increasingly taking advantage of internet and Web-based recruitment tools for targeting qualified applicants. As use of these techniques continues to grow, associated research will accumulate to assess their effectiveness.

On the predictor side, advances in the last decade have shown that we can reliably measure personality, motivational, and interest facets of human behavior and that under certain conditions these can add substantially to our ability to predict turnover, retention, and job performance. The reemergence of personality and related volitional constructs as predictors is a positive sign in that this trend should result in a more complete mapping of the KSAO requirements for jobs and organizations, beyond general cognitive ability.

We would also recommend that organizations consider ways of expanding the predictor and criterion domains that result in selecting applicants with a greater chance of long-term career success, and when doing so, it will be important to extend the perspective to broader implementation issues that involve classification of personnel and PO fit. As organizational flexibility in effectively utilizing employees increasingly becomes an issue, the PO fit model may be more relevant compared to the traditional PJ match approach.

Finally, in order to combat the potential loss of critical organizational knowledge when valued employees leave the organization, personnel turnover and retention remains a critical focus for organizations. Use of effective recruitment and hiring strategies as well as creating the right climate where employees feel they will be able to work productively and develop professionally will help to embed individuals in the culture of the organization and increase the chances that they will remain.

## REFERENCES

Abrahams, N. M., Neumann, I., and Githens, W. H. (1971), "Faking Vocational Interests: Simulated versus Real-Life Motivation," *Personnel Psychology*, Vol. 24, pp. 5–12.

Asher, J. J. (1972), "The Biographical Item: Can It Be Improved?" *Personnel Psychology*, Vol. 25, pp. 251–269.

Barrick, M. R., and Mount, M. K. (1991), "The Big Five Personality Dimensions and Job Performance: A Meta-Analysis," *Personnel Psychology*, Vol. 44, pp. 1–26.

Barrick, M. R., and Mount, M. K. (1996), "Effects of Impression Management and Self-Deception on the Predictive Validity of Personality Constructs," *Journal of Applied Psychology*, Vol. 81, pp. 261–272.

Barrick, M. R., and Zimmerman, R. D. (2005), "Reducing Voluntary, Avoidable Turnover through Selection," *Journal of Applied Psychology*, Vol. 90, pp. 159–166.

Barrick, M. R., Mount, M. K., and Judge, T. A. (2001), "Personality and Performance at the Beginning of the New Millennium: What Do We Know and Where Do We Go Next?" *International Journal of Selection and Assessment*, Vol. 9, pp. 52–69.

Blakley, B. R., Quinones, M. A., Crawford, M. S., and Jago, I. A. (1994), "The Validity of Isometric Strength Tests," *Personnel Psychology*, Vol. 47, pp. 247–274.

Borman, W. C., and Motowidlo, S. M. (1993), "Expanding the Criterion Domain to Include Elements of Contextual Performance," in *Personnel Selection*, N. Schmitt and W. C. Borman, Eds., Jossey-Bass, San Francisco, pp. 71–98.

Borman, W. C., White, L. A., Pulakos, E. D., and Oppler, S. H. (1991), "Models of Supervisory Job Performance Ratings," *Journal of Applied Psychology*, Vol. 76, pp. 863–872.

Borman, W. C., Hanson, M. A., Oppler, S. H., Pulakos, E. D., and White, L. A. (1993), "The Role of Early Supervisory Experience in Supervisor Performance," *Journal of Applied Psychology*, Vol. 78, pp. 443–449.

Borman, W. C., Hanson, M. A., and Hedge, J. W. (1997), "Personnel Selection," in *Annual Review of Psychology*, Vol. 48, J. T. Spence, J. M. Darley, and D. J. Foss, Eds., Annual Reviews, Palo Alto, CA, pp. 299–337.

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., and Drasgow, F. (2001), "An Examination of the Comparative Reliability, Validity, and Accuracy of Performance Ratings Made Using Computerized Adaptive Rating Scales," *Journal of Applied Psychology*, Vol. 86, pp. 965–973.

Borman, W. C., Hedge, J. W., Ferstl, K., Kaufman, J. D., Farmer, W. L., and Bearden, R. (2003), "Current Directions and Issues in Personnel Selection and Classification," in *Research in Personnel and Human Resources Management*, Vol. 22, J. J. Martocchio and G. R. Ferris, Eds., JAI Press, Stamford, CT, pp. 287–355.

Breaugh, J. A., Macan, T. H., and Grambow, D. M. (2008), "Employee Recruitment: Current Knowledge and Directions for Future Research," in *International Review of Industrial and Organizational Psychology*, Vol. 23, G. P. Hodgkinson and J. K. Ford, Eds., Wiley, New York, pp. 45–82.

Campbell, D. P. (1971), *Handbook for the Strong Vocational Interest Blank*, Stanford University Press, Stanford, CA.

Carroll, J. B. (1993), *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*, Cambridge University Press, New York.

Cascio, W. F. (1995), "Whither Industrial and Organizational Psychology in a Changing World of Work?" *American Psychologist*, Vol. 50, pp. 928–939.

Cascio W. (2000), *Costing Human Resources: The Financial Impact of Behavior in Organizations*, South-Western Publishing, Cincinnati, OH.

Chan, D. (2001), "Practical Intelligence and Job Performance," paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Chapman, D. S., and Webster, J. (2003), "The Use of Technologies in the Recruiting, Screening, and Selection Processes for Job Candidates," *International Journal of Selection and Assessment*, Vol. 11, pp. 113–120.

Chapman, D. S., Uggerslev, K. L., Carroll, S. A., Piasentin, K. A., and Jones, D. A. (2005), "Applicant Attraction to Organizations and Job Choice: A Meta-Analytic Review of the Correlates of Recruiting Outcomes," *Journal of Applied Psychology*, Vol. 90, pp. 928–944.

Coleman, V. I., and Borman, W. C. (2000), "Investigating the Underlying Structure of the Citizenship Performance Domain," *Human Resource Management Review*, Vol. 10, pp. 25–44.

Committee on Techniques for the Enhancement of Human Performance: Occupational Analysis, Commission on Behavioral and Social Sciences and Education, National Research Council (2000), *The Changing Nature of Work: Implications for Occupational Analysis*, National Academy Press, Washington, DC.

Conley, J. J. (1984), "Longitudinal Consistency of Adult Personality: Self-Reported Psychological Characteristics across 45 Years," *Journal of Personality and Social Psychology*, Vol. 47, pp. 1325–1333.

Conway, J. M. (1999), "Distinguishing Contextual Performance from Task Performance for Managerial Jobs," *Journal of Applied Psychology*, Vol. 84, pp. 3–13.

Czaja, S. J. (2001), "Technological Change and the Older Worker," in *Handbook of the Psychology of Aging*, J. E. Birren and K. W. Schaie, Eds., Academic, San Diego, CA, pp. 547–568.

Czaja, S. J., and Moen, P. (2004), "Technology and Employment," in *Technology for Adaptive Aging: Report and Papers*, R. Pew and S. Van Hemel, Eds., National Academies Press, Washington, DC.

Day, D. V. and Silverman, S. B. (1989), "Personality and Job Performance: Evidence of Incremental Validity," *Personnel Psychology*, Vol. 42, pp. 25–36.

deWolf, M., and Klemmer, K. (2010, May). "Job Openings, Hires, and Separations Fall During the Recession," *Monthly Labor Review*, Vol. 133, pp. 36–44.

Drasgow, F., and Olson-Buchanon, J. B., Eds. (1999), *Innovations in Computerized Assessment*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Durhart, D. T. (2001), *Violence in the Workplace, 1993–1999*, Publication 190076, U.S. Department of Justice, Washington, DC.

Dye, D. A., Reck, M., and Murphy, M. A. (1993), "The Validity of Job Knowledge Measures," *International Journal of Selection and Assessment*, Vol. 1, pp. 153–517.

Edwards, J. E., and Morrison, R. F. (1994), "Selecting and Classifying Future Naval Officers: The Paradox of Greater Specialization in Broader Areas," in *Personnel Selection and Classification*, M. G. Rumsey, C. B. Walker, and J. H. Harris, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 69–84.

Erez, A., and Judge, T. A. (2001), "Relationship of Core Self-Evaluations to Goal Setting, Motivation, and Performance," *Journal of Applied Psychology*, Vol. 86, pp. 1270–1279.

Farmer, W. L. (2001), "A Brief Review of Biodata History, Research, and Applications," unpublished manuscript, Navy Personnel Research, Studies, and Technology, Millington, TN.

Fleishman, E. A., Constanza, D. P., and Marshall-Mies, J. (1999), "Abilities," in *An Occupational Information System for the 21st Century: The Development of O*NET*, N. G. Peterson, M. D. Mumford, W. C. Borman, R. P. Jeanneret, and E. A. Fleishman, Eds., American Psychological Association, Washington D.C.

Glebbeek, A. C., and Bax, E. H. (2004). "Is High Employee Turnover Really Harmful? An Empirical Test Using Company Records," *Academy of Management Journal*, Vol. 47, pp. 277–286.

Goldberg, L. R. (1992), "The Structure of Phenotypic Personality Traits (or the Magical Number Five, Plus or Minus Zero)," keynote address to the Sixth European Conference on Personality, Groningen, The Netherlands.

Griffeth, R.W., Hom, P.W., and Gaertner, S. (2000). "A Meta-Analysis of Antecedents and Correlates of Employee Turnover: Update, Moderator Tests, and Research Implications for the Millennium," *Journal of Management*, Vol. 26, pp. 463–488.

Haptonstahl, D. E., and Buckley, M. R. (2002), "Applicant Fit: A Three-Dimensional Investigation of Recruiter Perceptions," paper presented at the Annual Meeting of the Society for Industrial and Organizational Psychology, Toronto, Ontario, Canada.

Harter, J. K., Schmidt, F. L., and Hayes, T. L. (2002), "Business-Unit-Level Relationship between Employee Satisfaction, Employee Engagement, and Business Outcomes: A Meta-Analysis," *Journal of Applied Psychology*, Vol. 87, pp. 268–279.

Hogan, J., and Hogan, R. (1989), "How to Measure Employee Reliability," *Journal of Applied Psychology*, Vol. 74, pp. 273–279.

Hogan, J., and Ones, D. S. (1997), "Conscientiousness and Integrity at Work," in *Handbook of Personality Psychology*, R. Hogan, J. A. Johnson, and S. R. Briggs, Eds., Academic, San Diego, CA, pp. 849–870.

Hogan, J., and Quigley, A. (1994), "Effects of Preparing for Physical Ability Tests," *Public Personnel Management*, Vol. 23, pp. 85–104.

Hogan, R. (1991), "Personality and Personality Measurement," in *Handbook of Industrial and Organizational Psychology*, 2nd ed., Vol. 2, M. D. Dunnette and L. M. Hough, Eds., Consulting Psychologists Press, Palo Alto, CA, pp. 873–919.

Hogan, R., and Blake, R. J. (1996), "Vocational Interests: Matching Self-Concept with the Work Environment," in *Individual Differences and Behavior in Organizations*, K. R. Murphy, Ed., Jossey-Bass, San Francisco, pp. 89–144.

Holland, J. L., and Hough, L. M. (1976), "Vocational Preferences," in *Handbook of Industrial and Organizational Psychology*, M. D. Dunnette, Ed., Rand McNally, Skokie, IL, pp. 521–570.

Hom, P. W., Griffeth, R. W. (1995), *Employee Turnover*, South-Western College Publishing, Cincinnati, OH.

Hunter, J. E. (1983), "A Causal Analysis of Cognitive Ability, Job Knowledge, Job Performance and Supervisory Ratings," in *Performance Measurement and Theory*, F. Landy, S. Zedeck, and J. Cleveland, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 257–266.

Hunter, J. E., and Hunter, R. F. (1984), "Validity and Utility of Alternative Predictors of Job Performance," *Psychological Bulletin*, Vol. 96, pp. 72–98.

Hurtz, G. M., and Donovan, J. J. (2000), "Personality and Job Performance: The Big Five Revisited," *Journal of Applied Psychology*, Vol. 85, pp. 869–879.

International Personality Item Pool (2001), "A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences," available: http://www.ipip.ori.org/.

Jattuso, M. L., and Sinar, E. F. (2003). "Source Effects in Internet-Based Screening Procedures," *International Journal of Selection and Assessment*, Vol. 11, pp. 137–140.

Judge, T. A., and Bono, J. E. (2001), "Relationship of Core Self-Evaluations Traits—Self-Esteem, Generalized Self-Efficacy, Locus of Control, and Emotional Stability—With Job Satisfaction and Job Performance: A Meta-Analysis," *Journal of Applied Psychology*, Vol. 86, pp. 80–92.

Judge, T. A., Locke, E. A., Durham, C. C., and Kluger, A. N. (1998), "Dispositional Effects on Job and Life Satisfaction: The Role of Core Evaluations," *Journal of Applied Psychology*, Vol. 83, pp. 17–34.

Judge, T. A., Bono, J. E., and Locke, E. A. (2000), "Personality and Job Satisfaction: The Mediating Role of Job Characteristics," *Journal of Applied Psychology*, Vol. 85, pp. 237–249.

Kacmar, K. M., Andrews, M. C., Van Rooy, D. L., Steilberg, R. C., and Cerrone, S. (2006), "Sure Everyone Can be Replaced . . . but at What Cost? Turnover as a Predictor of Unit-Level Performance," *Academy of Management Journal*, Vol. 49, pp. 133–144.

Kenrick, D. T., and Funder, D. C. (1988), "Profiting from Controversy: Lessons from the Person- Situation Debate," *American Psychologist*, Vol. 43, pp. 23–34.

Kristof, A. L. (1996), "Person–Organization Fit: An Integrative Review of Its Conceptualizations, Measurement, and Implications," *Personnel Psychology*, Vol. 49, pp. 1–49.

Kubisiak, U. C., Young, M. C., Legree, P. J., and Tremble, T. R. (2009), *Understanding and Managing the Career Continuance of Enlisted Soldiers*, U.S. Army Research Institute for the Behavioral and Social Sciences, Arlington, VA.

Landy, F. J. (1993), "Early Influences on the Development of Industrial/Organizational Psychology," in *Explaining Applied Psychology: Origins and Critical Analyses*, T. K. Fagan and G. R. Vardenbos, Eds., American Psychological Association, Washington, DC, pp. 79–118.

Larson, G. E., and Wolfe, J. H. (1995), "Validity Results for g from an Expanded Test Base," *Intelligence*, Vol. 20, pp. 15–25.

Lee, T. W., Mitchell, T. R., Sablynski, C. J., Burton, J. P., and Holtom, B. C. (2004), "The Effects of Job Embeddedness on Organizational Citizenship, Job Performance, Volitional Absences, and Voluntary Turnover," *Academy of Management Journal*, Vol. 47, pp. 711–722.

Le Pine, J. A., Colquitt, J. A., and Erez, A. (2000), "Adaptability to Changing Task Contexts: Effects of General Cognitive Ability, Conscientiousness, and Openness to Experience," *Personnel Psychology*, Vol. 53, pp. 563–593.

Mael, F. A. (1991), "A Conceptual Rationale for the Domain and Attributes of Biodata Items," *Personnel Psychology*, Vol. 44, pp. 763–792.

McCrae, R. R., and Costa, P. T., Jr. (1987), "Validation of the Five-Factor Model of Personality across Instruments and Observers," *Journal of Personality and Social Psychology*, Vol. 52, pp. 81–90.

McEvoy, G. M., and Cascio, W. F. (1985), "Strategies for Reducing Employee Turnover: A Meta-Analysis," *Journal of Applied Psychology*, Vol. 70, pp. 342–353.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M., and Ashworth, S. (1990), "Project A Validity Results: The Relationship between Predictor and Criterion Domains," *Personnel Psychology*, Vol. 43, pp. 335–354.

Mitchell, T. R., Holtom, B. C., Lee. T. W., Sablynski, C., and Erez, M. (2001), "Why People Stay: Using Job Embeddedness to Predict Voluntary Turnover," *Academy of Management Journal*, Vol. 44, pp. 1102–1121.

Muchinsky, P. M. (1993), "Validation of Intelligence and Mechanical Aptitude Tests in Selecting Employees for Manufacturing Jobs," *Journal of Business Psychology*, Vol. 7, pp. 373–382.

Mumford, M. D., Baughman, W. A., Threlfall, K. V., Uhlman, C. E., and Constanza, D. P. (1993), "Personality, Adaptability, and Performance: Performance on Well-Defined and Ill-Defined Problem-Solving Tasks," *Human Performance*, Vol. 6, pp. 241–285.

Murphy, K. R. (1996), "Individual Differences and Behavior in Organizations: Much More Than g," in *Individual Differences and Behavior in Organizations*, K. Murphy, Ed., Jossey-Bass, San Francisco.

Olea, M. M., and Ree, M. J. (1994), "Predicting Pilot and Navigator Criteria: Not Much More Than g," *Journal of Applied Psychology*, Vol. 79, pp. 845–851.

Ones, D. S., and Viswesvaran, C. (1998a), "The Effects of Social Desirability and Faking on Personality and Integrity Assessment for Personnel Selection," *Human Performance*, Vol. 11, pp. 245–269.

Ones, D. S., and Viswesvaran, C. (1998b), "Gender, Age, and Race Differences on Overt Integrity Tests: Results across Four Large-Scale Job Applicant Data Sets," *Journal of Applied Psychology*, Vol. 83, pp. 35–42.

Ones, D. S., Viswesvaran, C., and Schmidt, F. L. (1993), "Comprehensive Meta-Analysis of Integrity Test Validities: Findings and Implications for Personnel Selection and Theories of Job Performance", *Journal of Applied Psychology* (*Monograph*), Vol. 78, pp. 679–703.

Ones, D. S., Schmidt, F. L., and Viswesvaran, C. (1994), "Do Broader Personality Variables Predict Job Performance with Higher Validity?" in *Personality and Job Performance: Big Five versus Specific Traits*, R. Page, Chair, symposium conducted at the 9th Annual Conference of the Society for Industrial and Organizational Psychology, Nashville, TN.

Ones, D. S., Viswesvaran, C., and Reiss, A. D. (1996), "Role of Social Desirability in Personality Testing for Personnel Selection: The Red Herring," *Journal of Applied Psychology*, Vol. 81, pp. 660–679.

Peck, D. (2010, March). "How a New Jobless Era Will Transform America," *The Atlantic Monthly*, available: http://www.theatlantic.com/doc/201003/jobless-america-future.

Phillips, J. M. (1998). "Effects of Realistic Job Previews on Multiple Organizational Outcomes: A Meta-Analysis," *Academy of Management Journal*, Vol. 41, pp. 673–690.

Ployhart, R. E., Schneider, B., and Schmitt, N. (2006), *Staffing Organizations: Contemporary Practice and Theory*, 3rd ed., Lawrence Erlbaum Associates, Mahwah, NJ.

Podsakoff, P. M., MacKenzie, S. B., Paine, J. B., and Bachrach, D. G. (2000), "Organizational Citizenship Behaviors: A Critical Review of the Theoretical and Empirical Literature and Suggestions for Future Research," *Journal of Management*, Vol. 26, pp. 513–563.

Ree, M. J., Earles, J. A., and Teachout, M. S. (1994), "Predicting Job Performance: Not Much More than g," *Journal of Applied Psychology*, Vol. 79, pp. 518–524.

Reilly, R. R., and Chao, G. T. (1982), "Validity and Fairness of Some Alternative Employee Selection Procedures," *Personnel Psychology*, Vol. 35, pp. 1–62.

Roberts, B. W., and Del Vecchio, W. F. (2000), "The Rank Order Consistency of Personality Traits from Childhood to Old Age: A Quantitative Review of Longitudinal Studies," *Psychological Bulletin*, Vol. 126, pp. 3–25.

Rosse, J. G., Stecher, M. D., Miller, J. L., and Levin, R. A. (1998), "The Impact of Response Distortion on Pre-employment Personality Testing and Hiring Decisions," *Journal of Applied Psychology*, Vol. 83, pp. 634–644.

Salgado, J. F. (1998), "Big Five Personality Dimensions and Job Performance in Army and Civil Occupations: A European Perspective," *Human Performance*, Vol. 11, pp. 271–288.

Schaubroeck, J., Ganster, D. C., and Jones, J. R. (1998), "Organization and Occupation Influences in the Attraction–Selection–Attrition Process," *Journal of Applied Psychology*, Vol. 83, pp. 869–891.

Schlesinger, L. A., and Heskett, J. L. (1991), "Breaking the Cycle of Failure in Services," *Sloan Management Review*, Vol. 32, pp. 17–28.

Schmidt, F. L., and Hunter, J. E. (1998), "The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings," *Psychological Bulletin*, Vol. 124, pp. 262–274.

Schmidt, F, L,, and Rader, M. (1999), "Exploring the Boundary Conditions for Interview Validity: Meta-Analytic Validity Findings for a New Interview Type," *Personnel Psychology*, Vol. 52, pp. 445–464.

Schmidt, F. L., Hunter, J. G., and Outerbridge, A. N. (1986), "Impact of Job Experience and Ability on Job Knowledge, Work Sample Performance, and Supervisory Ratings of Job Performance," *Journal of Applied Psychology*, Vol. 71, pp. 432–439.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., and Jennings, D. (1997), "Adverse Impact and Predictive Efficiency of Various Predictor Combinations," *Journal of Applied Psychology*, Vol. 82, pp. 719–730.

Schneider, B. (1987), "The People Make the Place," *Personnel Psychology*, Vol. 40, pp. 437–453.

Schneider, B. (1989), "$E = f (P, B)$: The Road to a Radical Approach to $P-E$ Fit," *Journal of Vocational Behavior*, Vol. 31, pp. 353–361.

Schneider, B., Smith, D. B., and Goldstein, H. W. (2000), "Toward a Person–Environment Psychology of Organizations," in *Person–Environment Psychology*, 2nd ed., W. B. Walsh, K. H. Craik, and R. H. Price, Eds., Lawrence Erlbaum Associates, Mahwah, NJ.

Sternberg, R. J. (1985), *Beyond IQ: A Triarchic Theory of Human Intelligence*, Cambridge University Press, New York.

Sternberg, R. J., and Wagner, R. K. (1992), "Tacit Knowledge: An Unspoken Key to Managerial Success," *Tacit Knowledge*, Vol. 1, pp. 5–13.

Sternberg, R. J., Wagner, R. K., and Okagaki, L. (1993), "Practical Intelligence: The Nature and Role of Tacit Knowledge in Work and at School," in *Advances in Lifespan Development*, H. Reese and J. Puckett, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 205–227.

Sternberg. R. J., Wagner, R. K., Williams, W. M., and Horvath, J. A. (1995), "Testing Common Sense," *American Psychologist*, Vol. 50, pp. 912–927.

Stone, D. L., Lukaszewski, K., and Isenhour, L. C. (2005), "e-Recruiting: Online Strategies for Attracting Talent," In H. Gueutal and D. L. Stone, Eds., *The Brave New World of EHR: Human Resources in the Digital Age*, Wiley, New York, pp. 22–53.

Stricker, L. J. (1987), "Developing a Biographical Measure to Assess Leadership Potential," paper presented at the Annual Meeting of the Military Testing Association, Ottawa, Ontario, Canada.

Stricker, L. J. (1988). "Assessing Leadership Potential at the Naval Academy with a Biographical Measure," paper presented at the Annual Meeting of the Military Testing Association, San Antonio, TX.

Strong, E. K. (1943), *Vocational Interests of Men and Women*, Stanford University Press, Stanford, CA.

Tett, R., and Meyer J. (1993), "Job Satisfaction, Organizational Commitment, Turnover Intention, and Turnover: Path Analysis Based on Meta-Analytic Findings," *Personnel Psychology*, Vol. 46, pp. 259–293.

Tupes, E. C., and Christal, R. E. (1992), "Recurrent Personality Factors Based on Trait Ratings," *Journal of Personality*, Vol. 60, pp. 225–251.

U.S. Department of Health and Human Services (1997), *An Analysis of Worker Drug Use and Workplace Policies and Programs*, U.S. Government Printing Office, Washington, DC.

U. S Department of Labor, Bureau of Labor Statistics (2010, August 26), Worker Displacement News Release (USDL-10-1174), available: http://www.bls.gov/news.release/archives/disp_08262010.htm.

Van Vianen, A. E. M. (2000), "Person–Organization Fit: The Match between Newcomers' and Recruiters' Preferences for Organizational Cultures," *Personnel Psychology*, Vol. 53, pp. 1–32.

Wagner, R. K., and Sternberg, R. J. (1985), "Practical Intelligence in Real-World Pursuits: The Role of Tacit Knowledge," *Journal of Personality and Social Psychology*, Vol. 49, pp. 436–458.

Wanous, J. P. (1980), *Organizational Entry: Recruitment, Selection, and Socialization of Newcomers*, Addison-Wesley, Reading, MA.

Werbel, J., and Gilliland, S. W. (1999), "Person–Environment Fit in the Selection Process," in *Research in Personnel and Human Resources Management*, Vol. 17, G. R. Ferris, Ed., JAI Press, Stanford, CT, pp. 209–243.

Wigdor, A. K., and Garner, W. R., Eds. (1982), *Ability Testing: Uses, Consequences, and Controversies*, Parts I and II, National Academy Press, Washington, DC.

Yakubovich, V., and Lup, D. (2006). "Stages of the Recruitment Process and the Referrer's Performance Effect," *Organization Science*, Vol. 17, pp. 710–723.

Zottoli, M. A., and Wanous, J. P. (2000), "Recruitment Source Research: Current Status and Future Directions," *Human Resource Management Review*, Vol. 10, pp. 353–383.

# CHAPTER 17

# DESIGN, DELIVERY, EVALUATION, AND TRANSFER OF TRAINING SYSTEMS

**Chris W. Coultas, Rebecca Grossman, and Eduardo Salas**
University of Central Florida
Orlando, Florida

## 1 INTRODUCTION

Training is a key component to the life of the modern organization. Various instructional strategies, learning options, and technologies have made the training industry larger and more diverse than ever. The American Society for Training and Development notes in their "State of the Industry Report" (Paradise and Patel, 2009) that organizations, on average, spent $1068 and allotted 36.3 h per employee for learning and development efforts in 2008. While this represents a 3.8% monetary decrease from 2007, due largely to the worldwide economic downturns, it is a 30% increase from 2003 to 2004, when organizations spent approximately $820 per employee (Sugrue and Rivera, 2005). These individual training expenses aggregate to an astounding $134.07 billion spent on employee training in 2008. Considering organizations invest so much in training, it follows that training should be designed in such a way as to maximize its effectiveness. Not only can proper training be a boon to organizations, but improper training can have severely negative consequences, and effective training can help mitigate these. At the most dramatic, improper (or lack of) training can lead to injury or death—as early as 2001, reports indicated the $131.7 billion was spent annually due to employee injuries and death (National Safety Council); these figures increased to $183.0 billion during 2008 (National Safety Council, 2010).

Training can, and does, differ greatly in regards to specific emphases (e.g., prevent accidents, improve service, create better products), and these generally vary depending on the organization. However, the universal goal of training is to improve the quality of the workforce so as to strengthen the bottom line of the organization. If this is to happen, training designers and providers must take advantage of the ever-growing field of training research (Aguinis and Kraiger, 2009; Salas and Cannon-Bowers, 2001) and apply it to the analysis, design, delivery, evaluation, and transfer of training systems. Therefore, this chapter will outline the relevant research on what constitutes an effective training system. To this end, we have updated the Salas and colleagues (2006b) "design, delivery, and evaluation of training systems" and incorporated a vast amount of recent work on training into the review (e.g., Kozlowski and Salas, 2010; Aguinis and Kraiger, 2009; Sitzmann et al., 2008; Burke and Hutchins, 2007).

## 1.1 Training Defined

While training has been shown to generally have positive effects on individual and organization levels, it is vital for an overview of the training literature to begin with a definition of what training is. At its most basic level, training can be defined as any systematic efforts to impart knowledge, skills, attitudes,[*] or other characteristics with the end goal being seeing improved performance. To achieve this broad goal, training must change some (or all) of the following characteristics of the trainee: knowledge, patterns of cognition, attitudes, motivation, and abilities. However, this cannot occur if training is designed and delivered haphazardly and with general disregard to the scientific principles of training. Training efforts must have a keen eye toward the science of training and learning—it must provide opportunities to not only *learn* the necessary knowledge, skills, attitudes, and other characteristics (KSAOs) but also practice and apply this learning and receive feedback regarding these attempts within the training (Salas and Cannon-Bowers, 2000a; Aguinis and Kraiger, 2009).

In today's rapidly expanding technological landscape, it is the tendency of individuals to nearly automatically accept the newest technology in training as the best or most appropriate. However, technological advances do not necessarily equate to psychological advances, and many times, training strategies independent of advanced technology have proven to be just as effective (if not more so) than their more advanced counterparts. This is discussed in greater detail later in the chapter, but this serves to illustrate the point that the elements of training have a very real scientific basis.

Chen and Klimoski (2007) recently reviewed the literature on training and development, and their qualitative findings should be encouraging to researchers and practitioners alike. They state that the current literature regarding training is, on the whole, scientifically rigorous and has "generated a large body of

knowledge pertaining to learning antecedents, processes, and outcomes" (p.188). This is a major improvement over earlier literature reviews describing then-current research on training as "nonempirical, nontheoretical, poorly written, and dull" (Campbell, 1971, p. 565). While research avenues in the field of training are by no means exhausted, theories, models, and frameworks provide effective guidelines for developing and implementing training programs (Salas and Cannon-Bowers, 2001). These theories and frameworks are described and referenced throughout this chapter.

## 2 SCIENCE OF TRAINING: THEORETICAL DEVELOPMENTS

It has been said that there is "nothing more practical than a good theory" (Lewin, 1951). This timeless quote illustrates the fact that, while some emphasize a divide between basic and applied research or between academia and industry, it is empirically grounded theory that enables practitioners to design, develop, and deliver training programs with confidence. Accordingly, we do not overlook the theoretical underpinnings that drive the science and practice of training. In the previous review by Salas and colleagues (2006b), we identified the transfer of training as a developing area of research we deemed important to include in our discussion of training. Prominent models of transfer and training effectiveness were depicted and described. Since that time, training scholars have made additional strides toward our understanding of the training process. The role of transfer continues to be emphasized in theoretical models and empirical studies. Transfer models have been supplemented and revised. Updated models of training effectiveness have also been proposed. In the following paragraphs, we briefly review the training theories we previously described and discuss theoretical developments that have since occurred.

If the ultimate goal of training is to see positive organizational change, trainees must be able to transfer what they have learned in the training environment and apply it to work within the organizational setting. Accordingly, it is vital for researchers and practitioners alike to understand under what contexts this transfer of training is likely to occur. Recent research continues to support Thayer and Teachout's (1995) widely accepted model of training transfer outlined in the previous review (Salas et al., 2006b). The factors they identified as maximizing transfer have been separately supported within the literature and include trainee: (1) reactions to previous training (Baldwin and Ford, 1988), (2) education (Mathieu et al., 1992), (3) pretraining self-efficacy (Ford et al., 1992), (4) ability (Ghiselli, 1966), (5) locus of control (Williams et al., 1991), (6) job involvement (Noe and Schmitt, 1986), and (7) career/job attitudes (Williams et al., 1991). Additionally, trainee reactions to the training/task at hand regarding overall likability (Kirkpatrick, 1976), perceived instrumentality of training (Velada et al., 2007), and expectancy that transfer can/will occur (i.e., self-efficacy, Latham, 1989; Tannenbaum et al., 1991) have all been shown to lead to

---

[*] Although much of the literature refers to the "A" of KSAOs as abilities (e.g., Goldstein, 1993), in this chapter we refer to the "A" as attitudes.

greater training transfer. Furthermore, organizational climate and transfer-enhancing activities can facilitate the transfer of training. In fact, a recent meta-analysis (Blume et al., 2010) showed that an organizational climate that supports training efforts is the most important aspect of the work environment in predicting training transfer. Organizational cues and consequences can encourage transfer. Transfer-enhancing activities such as goal setting, relapse prevention, self-management (Baldwin et al., 2009), and job aids can assist the trainee in applying learning in the long term. The final factor in the Thayer and Teachout (1995) model is results; transfer of training should be evaluated in the context of whether learned knowledge translates into workplace behavior and whether behavioral change leads to organizationally desirable results.

More recently, Burke and Hutchins (2008) proposed a model of transfer based on a major review of the training literature. The model, though simple, effectively summarizes much of the research on training, as it identifies personal (i.e., trainer/trainee), training (e.g., design, content), and environmental (i.e., organizational) characteristics as all having significant effects on training transfer. Furthermore, various aspects are seen as having greater impact at various times in the life cycle of training (e.g., trainer characteristics are more salient pretraining, organizational characteristics are more salient posttraining). See Figure 1 for the full model.

Recent research has emphasized that training is a multifaceted phenomenon, consisting of complex interactions between trainee, trainer, organization, and the

training itself. Bell and Kozlowski (2008) developed a model that reflects many of these interactions. Their model outlines the various interactions necessary to see trainees gain knowledge and transfer that knowledge to the workplace. While the model excludes some important aspects of the training experience, most notably organizational characteristics, it shows the impact that individual characteristics (i.e., cognitive ability, motivation, personality, etc.) can have on training design choices (i.e., error framing, exploratory learning) and how they interact to affect learning and training outcomes. Another major advantage of this model is that Bell and Kozlowski empirically verified their model using structural equation modeling. Their theoretical model is replicated in Figure 2, but a more complex version, complete with correlation indices, is included in Bell and Kozlowski (2008).

An incredibly in-depth examination of the factors and processes involved in training provided by Tannenbaum and colleagues (1993; Cannon-Bowers et al., 1995) was described in the previous review and remains relevant in current discussions of the training process (Figure 3). The framework considers training from a longitudinal, process-oriented perspective. Similarities between this model and the Burke and Hutchins (2008) model are apparent; however, this model is much more detailed. Additionally, this model identifies vital actions to ensure training success (e.g., training needs analysis).

Mathieu and Tesluk (2010) recently described a multilevel perspective of training effectiveness in which several levels of analysis are considered. In this view,
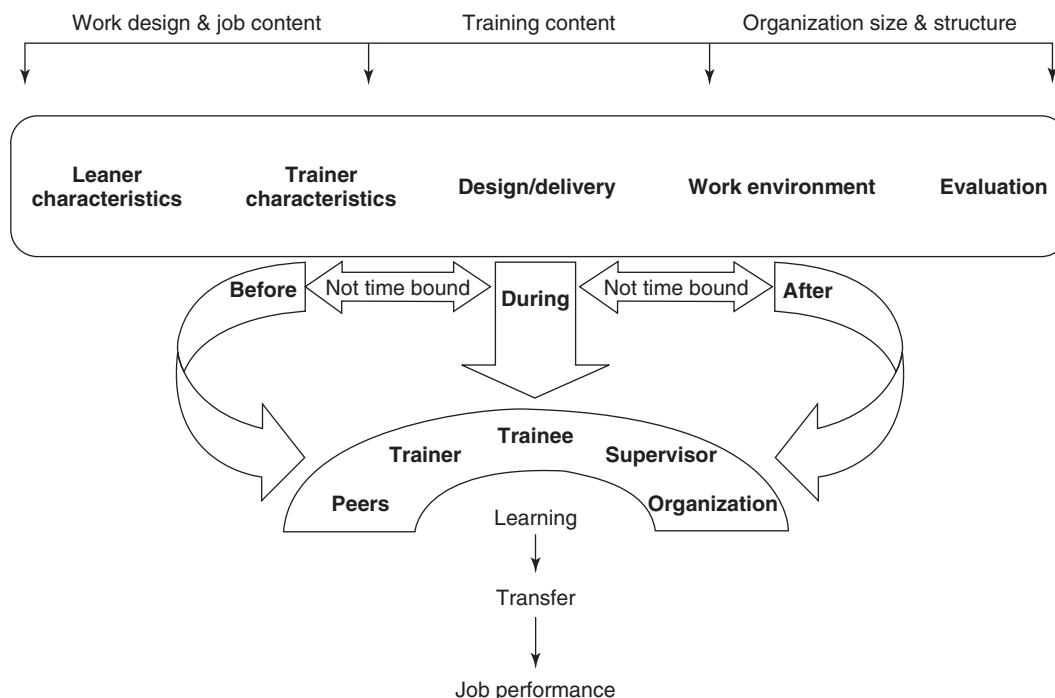


**Figure 1**   Transfer model. (From Burke and Hutchins, 2008.)
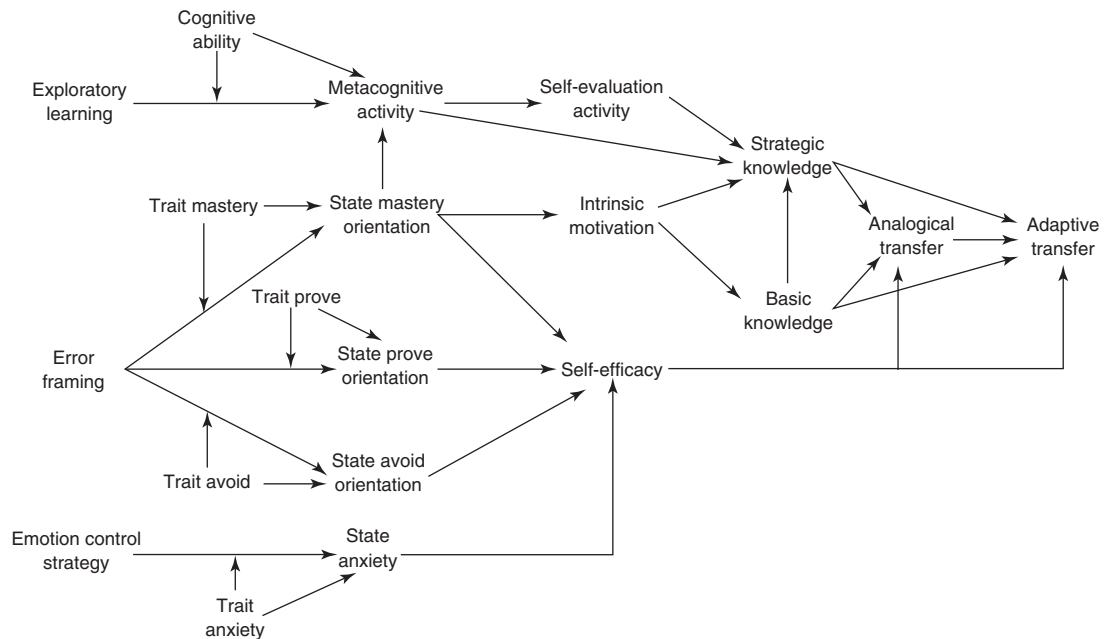
**Figure 2** Transfer model (From Bell and Kozlowski, 2008).
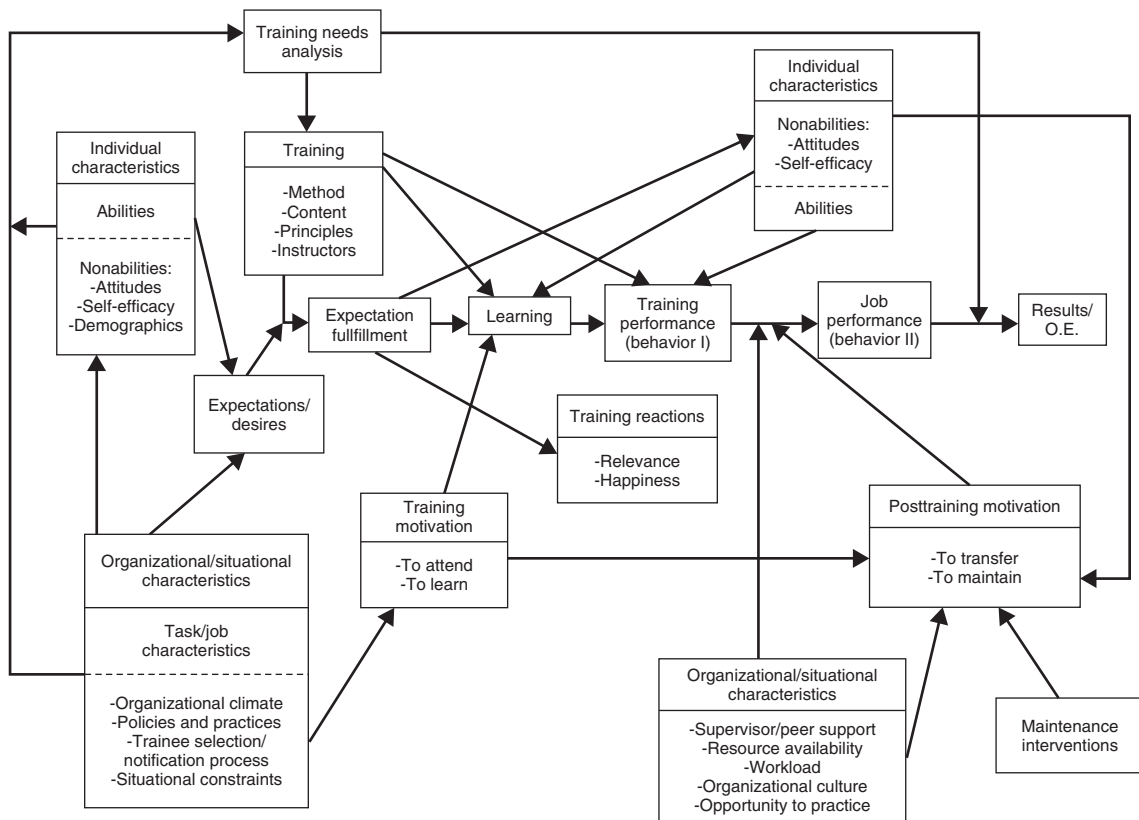


**Figure 3** Training effectiveness model. (From Tannenbaum et al., 1993; Cannon-Bowers et al., 1995.)

training outcomes result from the combined influences of various organizational levels (e.g., individuals, work-groups, departments). Further, unlike other multilevel approaches, the current framework incorporates both bottom-up (e.g., factors at the training level) and top-down (e.g., factors at the organizational level) influences on training effectiveness. Further, microlevel outcomes such as the specific KSAOs that are developed and macrolevel outcomes such as broader organizational outcomes are considered. Finally, the multilevel perspective put forth by Mathieu and Tesluk (2010) considers the combined effects of training and other human resource interventions (i.e., compilation approach) as well as the impact of organizational factors on individual-level training outcomes (i.e., cross-level approach). A visual representation of this approach is depicted in Figure 4.

Specific issues relevant to training, such as motivation (Colquitt et al., 2000; Chiaburu and Lindsay, 2008), individual characteristics and work environment (Tracey et al., 2001; Bell and Kozlowski, 2008), learner control (Sitzmann et al., 2008), training evaluation (Kraiger et al., 1993), and transfer of training (Quinones, 1997; Burke and Hutchins, 2007), have been studied extensively. The increasing reliance on teams in organizations has also led to a focus on team training. Kozlowski and colleagues (2000) explored how training and individual processes lead to team and organizational outcomes. Other models throughout the past decade or more have examined the factors of organizations, individuals, and training that can impact team motivation and performance (Tannenbaum et al., 1993). Salas and colleagues (2007) reviewed and integrated over 50 models on team effectiveness; the

| | Focal Level of Influences | |
|---|---|---|
| | **Specific Training Program** | **HRM system** |
| **Individual** | **Micro approach**<br><br>Influence of specific training program on targeted employee KSAOs | **Cross-level approach**<br><br>Impact of unit or organizational factors on effectiveness of individual level training outcomes<br><br>**_"Regional" illustration_**<br><br>• Senior executive support<br>• Organizational structure and systems<br>• Organizational culture<br>• Situational constrains<br><br>_Top-down influences_<br><br>Influences on participants' pretraining conditions (e.g., motivation, self-efficacy), learning, and transfer |
| **Organizational** | **Compilation Approach**<br><br>Combined influence of training and other HR systems on organizational outcomes<br><br>**_"Regional" illustration_**<br><br>Influences on human capital, social capital,organizational capital outcomes<br><br>_Bottom-up influences_<br><br>• Accounting for different forms of HR<br>• capital<br>• Vertical fit<br>• Horizontal fit<br>• Differentiation of HRM activities | **Macro Approach**<br><br>Relative influence of training system on organizational outcomes |

(Left margin label spanning rows: **Focal level of outcome**)

**Figure 4**   Multilevel model of training effectiveness. (From Mathieu and Tesluk, 2010.)

result was a comprehensive yet parsimonious framework of team effectiveness that can be used to tailor training.

Researchers and organizations have continued to benefit from developments in our knowledge of the training process and the factors that influence it. However, our knowledge is far from comprehensive. In our review of the literature, we noticed a heavy theoretical emphasis on training transfer. This is not a misplaced emphasis, for if transfer does not occur, then much of the purpose of training has been nullified. Despite the importance of transfer, it would be useful to see theoretical (and empirically grounded) universal models of the individual elements of training. These elements are discussed in the subsequent section.

## 2.1 Summary

Since Campell's (1971) review of the training literature, researchers have made great efforts toward solidifying the science of training. It is no longer a point of contention that training is atheoretical. Framework and models are more solid, and the field as a whole is much more empirical (Chen and Klimoski, 2007). While research is never complete, practioners can be confident that their efforts, when conducted in accordance with scientific findings, are founded on strong empirical evidence.

## 3 INSTRUCTIONAL SYSTEMS DEVELOPMENT MODEL

Our discussion of the design, delivery, transfer, and evaluation of training in this chapter is organized on the basis of a macrolevel systems approach. Specifically, we adopt the systems approach put forth by Goldstein (1993) in which four items serve as the basic foundation: (1) training design is iterative and hence feedback is continuously used to update and modify the training program; (2) complex interactions develop between training components (e.g., trainees, tools, instructional strategies); (3) a framework for reference to planning is provided; and (4) training systems are merely a small component of the overall organizational system and hence characteristics of the organization, task, and person should be considered throughout training design. Additionally, much of this chapter is guided by the instructional systems development (ISD) model (Branson, et al., 1975), which was established nearly four decades ago. Initially used in the development of U.S. military training programs, the ISD model continues to be used today and has been regarded as the most comprehensive training model available (Swezey and Llaneras, 1997). The model has also been the subject of criticism, however; thus we do not suggest that it is without flaws or superior to other models. Instead, we view it as a valuable framework for organizing the material we present in this chapter.

In the ISD model, occasionally referred to as the ADDIE model, training consists of five basic stages: analysis (A), design (D), development (D), implementation (I), and evaluation (E). Consistent with the previous review (Salas et al., 2006b), we also include a sixth step, namely transfer of training, as we continue to argue for its critical role in the training process. The process begins with the first phase, *training analysis*, in which the needs and goals of the organization are identified, and the overall plan for the design, delivery, and evaluation of the training is formulated. In the second phase, *training design*, learning objectives and performance measures are developed, and the progression of the training program is further mapped out. The training is fully developed in phase 3, *training development*, and weaknesses are identified and revised before the program is implemented. In the next phase, *training implementation*, final preparations are made (e.g., selecting a location where the training will be conducted), and the training is actually implemented. The training *evaluation phase* should follow, in which a multilevel approach is used to evaluate the effectiveness of the training. Finally, the *training transfer* phase should be implemented in order to determine the degree to which trained competencies are applied when the trainees return to the work environment. We utilize our knowledge of the science of training to describe each of these phases and form an integrated discussion of training as a whole.

While we adopt the ADDIE model of training and use it as the organizational framework for this review, other instructional design models exist (over 100; Chen, 2008). A few that bear mention are the rapid prototyping, R2D2, and 4C/ID models (Chen, 2008; Swaim, 2005; van Merriënboer and Kester, 2008). The rapid prototyping model is a form of ISD that developed in response to the breakneck pace of technological developments; the overall goal of this model is to utilize the key elements of ISD, but to do so much faster. This is accomplished by layering the different elements of instructional design. That is, even while the training designer is still analyzing training needs, he or she begins contructing and testing a prototype; when training objectives finally start to materialize, the training designer will still be developing and implementing trial runs of the training (Chen, 2008, Swaim, 2005). The "recursive, reflective, design, and development" model (R2D2; Chen, 2008) suggests a nonlinear approach to instructional design. While not as specific as the ADDIE model, it acknowledges a salient truth about training design: It never goes as planned. The R2D2 model of training design asserts that the process of reconsidering, rethinking, and revamping one's training is almost a neverending struggle throughout the training design and implementation; it is this amorphous, constantly evolving form of instructional design that delineates this from other theories of ISD (Chen, 2008). Finally, the 4C/ID (four components instructional design), discussed in the previous section, has implications for the development, design, and delivery of training systems (van Merriënboer and Kester, 2008). This holistic approach to training may be advantageous for trainers training for a complex set of tasks, as this model takes into account the complexity of the tasks, their interactions with each other, and the impact this may have on the learning process.

### 3.1 Summary

Though there are multiple models available in the literature, we argue for a systems approach and use the instructional design model to guide the organization of this chapter. Because no model is without flaws, the effective design, delivery, and evaluation of training may require a combination of several models.

## 4 TRAINING ANALYSIS

Before designing or implementing a training program, a training needs analysis is essential (Goldstein and Ford, 2002). Such an analysis includes where an organization needs training, what needs to be trained, and who needs the training (Goldstein, 1993). Training needs analysis will lead to (1) specifying learning objectives, (2) guiding training design and delivery, and (3) developing criterion. Needs assessment is vital in setting these training goals and determining the readiness of participants (Aguinis and Kraiger, 2009). But despite the foundational nature of training needs analysis, very little empirical analysis is available on its effectiveness, even at this juncture in the literature (Dierdorff and Surface, 2008; Tannenbaum and Yukl, 1992). Following is an explication of the three aspects of training analysis: organizational, job/task, and person.

### 4.1 Organizational Analysis

Organizational analysis is the first step in conducting training needs analysis. Various organizational aspects that may affect training delivery and/or transfer are considered through organizational analysis. This includes organizational climate/culture, norms, resources, constraints, support, et cetera (Festner and Gruber, 2008; Goldstein 1993). Another key element of organizational analysis is determining the "fit" of training objectives (see Section 5.1) with organizational objectives. For example, if an objective of training is to encourage appreciation for diversity, the organization must reflect that goal in day-to-day practice. A supportive organizational climate for transfer and application of trained KSAOs is vital for effective training; this concept is explored in further detail later (Bunch, 2007; Rouiller and Goldstein, 1993; Tracey et al., 1995). Considering that organizational factors can have a great impact on the implementability, continuity, and effectiveness of a training program, organizational analysis is a vital step in training needs analysis.

It has been nearly two decades since researchers began recognizing the importance of organizational analysis; this has led to several well-cited studies of the impact of the organization on training. Essentially, what has been found is that organizational climate (e.g., situational cues, consequences, support) is a strong predictor of training transfer (Lim and Morris, 2006; Rouiller and Goldstein, 1993; Tracey et al., 1995). For example, an organization implementing a sexual harassment training program that does not foster a climate of safety and openness will see little training effectiveness (Knapp and Kustis, 1996). But despite the clear importance of organizational factors on training transfer,

a recent meta-analysis by Arthur and colleagues (2003) showed inconsistent effects of organizational analysis on training effectiveness because there were *only four studies* that included this vital aspect of needs analysis.

### 4.2 Job/Task Analysis

After identifying organizational characteristics relevant to the intended training program, it is necessary to conduct a job/task analysis. The aim of this analysis is to identify characteristics of the actual task being trained for, so that the training program has specific, focused learning objectives (Goldstein, 1993). The first step is specifying the essential work functions of the job as well as requisite resources for job success. After outlining these generalities, specific tasks are identified. Furthermore, the contexts under which these tasks will be performed must be specified. Finally, task requirements and competencies (i.e., knowledge, skills, and attitudes) need to be defined. Often, this final step is the most difficult, because vague task requirements such as knowledge and attitudes are often difficult to observe in practice and thus are likely to be overlooked when designing training.

More complex tasks, however, may require an analysis of the cognitive demands (e.g., decision making, problem solving) of the tasks. These competencies are less observable and therefore require a specific cognitive task analysis to uncover them.

#### 4.2.1 Cognitive Task Analysis

Recent research has focused on trainees' acquisition and development of knowledge and how they come to understand processes and concepts (Zsambok and Klein, 1997; Schraagen et al., 2000). Additionally, research has acknowledged the importance of understanding subject matter experts' (SMEs') decision making in natural, complex settings (Gordon and Gill, 1997). A more targeted approach for identifying SME task cognitive processes is necessary, as SMEs' awareness of their skills tend to fade as they become more automated (Villachica and Stone, 2010). Cognitive task analysis (CTA) has emerged as the primary tool for understanding these processes (Salas and Klein, 2000). Various elicitation techniques (discussed subsequently) are used in CTA to specify cognitive processes involved in learning (trainees) as well as maximal performance (experts) (Villachica and Stone, 2010; Cooke, 1999; Klein and Militello, 2001).

It has been suggested that three criteria are essential to successful CTA (Klein and Militello, 2001). First, CTA must identify new information regarding trainees' patterns and strategies for learning and task success as well as other cognitive demands (e.g., cue patterns). This may be done through a variety of elicitation methods, such as structured and unstructured interviews, observing and coding of actual behaviors, or elicitation by critiquing. Elicitation by critiquing refers to having SMEs observe prerecorded scenarios and then critique the processes occurring in that recording (Miller et al., 2006). Second, these findings must be conveyed to training designers, who will incorporate them into the training design. The successful (read: impactful)

incorporation of CTA findings into training design is the final step of effective CTA. A potential prerequisite has been suggested for successful CTA: accurately identifying SMEs; this is a more complex undertaking than one might initially think (see Hoffman and Lintern, 2006). The steps for conducting a cognitive task analysis are summarized in Table 1.

### 4.3 Person Analysis

The final stage of training needs analysis is person analysis. Person analysis encompasses who needs training,

what they need to be trained on, and how effective training will be for individuals (Tannenbaum and Yukl, 1992; Goldstein, 1993). The foundation of person analysis is that not everyone needs the same training, and not everyone responds similarly to all training. Different job domains within organizations require different training, as do individuals with differing levels of expertise (Feldman, 1988). Even employees within the same job domain (i.e., management) operating at different levels tend to need training and view training differently due to varying job demands (Lim and Morris, 2006; Ford

**Table 1  Steps in Conducting Cognitive Task Analysis**

| Step | Guidelines |
|---|---|
| 1. Select experts. | • Consider how many SMEs would be appropriate.<br>• Select SMEs who have experience with domain.<br>• Select SMEs who have experience with task. |
| 2. Develop scenarios based on task analysis. | • Develop scenarios that are task relevant with problem statements.<br>• Pretest scenarios to determine if they are complete. |
| 3. Choose a knowledge elicitation method. | • Determine the information you are trying to get at. |
|    • Interviews | • Ask SMEs about cognitive processes needed to complete task scenarios. |
|    • Verbal protocols | • Require SMEs to think out loud as they complete task scenarios. |
|    • Observations | • Observe SMEs performing tasks. |
|    • Conceptual methods | • Develop inferences based on SME input and relatedness judgments. |
|    • Critiquing | • Have SMEs outline cognitive needs based on their observations of prerecorded scenarios. |
| 4. Implement chosen knowledge elicitation method. | • Decide how many sessions to be recorded.<br>• Obtain consent and provide task scenario to SMEs. |
|    • Interviews | • Provide each scenario to each SME.<br>• Ask SMEs relevant questions:<br>  What rules/strategies would you use to complete scenarios?<br>  What knowledge and cognitive skills would you use to complete scenarios?<br>  If A or B happens, what would they do?<br>• Keep a record of SME responses: use video, audio, pen and paper, etc. |
|    • Verbal protocols | • Provide each scenario to each SME.<br>• Require SMEs to think out loud as they complete task scenarios.<br>• Keep a record of SME responses: use video, audio, pen and paper, etc. |
|    • Observations | • Provide each scenario to each SME.<br>• Be as unobtrusive as possible, but do not be afraid to ask for clarification when necessary.<br>• Keep a record of SME responses: use video, audio, pen and paper, etc. |
|    • Conceptual methods | • Provide each scenario to each SME.<br>• Present pairs of tasks to experts.<br>• Ask SMEs for relatedness judgments regarding task pairs. |
|    • Critiquing | • Provide each scenario to each SME.<br>• Have SMEs critique performance observed in the scenario<br>• Ask SMEs to identify cognitive processes at play in scenarios |
| 5. Organize and analyze data. | • Keep a record of similarities between SMEs.<br>• Interview additional experts if questions arise from data.<br>• Identify the rules and strategies applied to the scenario.<br>• Identify knowledge and cognitive skills required.<br>• Generate a list of task requirements.<br>• Verify list with additional SMEs. |

*Source:* Adapted from Salas et al. (2006b)

and Noe, 1987). Person analysis identifies individuals' possession of the requisite competencies outlined in task analysis. As previously noted, knowledge and attitudes are also frequently considered competencies to be included in training needs analysis, and person analysis places emphases on these. Individual characteristics such as personality, self-efficacy, goal orientation, and prior training experience have all been shown to impact training effectiveness and should therefore be considered in training needs analysis (Rowold, 2007; Chiaburu and Lindsay, 2008; Roberson et al., 2009; Velada et al., 2007; Cheramie and Simmering, 2010).

One aspect of the person that often has bearing on what kind of training is used is individual learning styles. The concept of learning styles refers to the notion that individuals prefer specific instructional methods (e.g., visual, auditory, experiential) and that, when instructional strategies match learning styles, individuals learn significantly better. While this is by no means a new hypothesis, a recent review of the concept provides insights relevant to person analysis. In their review of the literature, Pashler et al. (2008) found that, as of late, research has not supported the learning styles hypothesis. However, they note that not all hypothesized learning styles have been studied, so the theory cannot be discounted entirely. The overall suggestion stemming from this review is that training designers need not be overly concerned with tailoring instructional strategies to learning styles, but simply need to use the instructional strategy that best fits the KSAOs being trained.

### 4.4 Summary

For training to be effective, training needs analysis is a necessity. If trainees do not meet the requirements for training (e.g., possess necessary competencies and attitudes), if training is essentially irrelevant to the actual job, or if aspects of the organization (e.g., culture, support) are not conducive to training, training will fail. Our suggestion: Thoroughly analyze training needs before designing any training system.

## 5  TRAINING DESIGN

Training design is the natural outcome of training needs analysis and is the next step in the ISD framework. Analysis-driven training design ensures that training is systematic and provides structure to the training program. Several things must be taken into account when designing training: development of training objectives, external factors relevant to the training program (i.e. individual/organizational characteristics and resources), selection of training strategies and methods, and developing specific program content.

### 5.1 Training Objectives

Training and learning objectives are developed in accordance with the training needs analysis. Objectives should be specific, measurable, and task relevant so that learning and performance can be evaluated posttraining. There are three components to training objectives: *performance*, *conditions*, and *standards* (Buckley and

Caple, 2007; Goldstein, 1993). *Performance* refers to both the end goal of training (terminal objective) and the requisite process behaviors (enabling and session objectives) to reach the end goal. Performance objectives can be laid out in a hierarchy such that session objectives (mundane behaviors) aggregate to enabling objectives, which are the major components of the terminal objective, or the end goal of training. *Conditions* also must be specified as to when and where these behaviors must be exhibited. *Standards* is the final component of training objectives; this refers to specifying what will be considered acceptable performance levels of the various objectives. Clearly defined objectives are able to effectively guide the training designers' choice of instructional strategy (or strategies). These strategies are selected according to their ability to promote objective-based behaviors and competencies. See Table 2 for guidance on developing training objectives.

### 5.2 Individual Characteristics

An important aspect of any training program is the individuals that receive it. Accordingly, the individual characteristics that participants bring to the training program, such as cognitive ability, self-efficacy, goal orientation, and motivation, must be considered when designing any training program. Each of these characteristics is discussed in the following.

#### 5.2.1 Cognitive Ability

Research has shown cognitive ability (i.e., $g$, or general mental ability) has clear and strong effects on the outcomes of training. Cognitive ability positively influences on-the-job knowledge acquisition (see Ree et al., 1995; Colquitt et al., 2000) and generally impacts trainees' ability to learn, retain, and transfer training material (Burke and Hutchins, 2007). Overall, individuals higher in cognitive ability are more likely to achieve training success on all fronts.

#### 5.2.2 Self-Efficacy

Self-efficacy, or the individual's belief in personal ability, is another individual characteristic that has exhibited strong effects on reactions to training (Mathieu et al., 1992), motivation within training (Chiaburu and Lindsay, 2008; Quinones, 1995), training performance (Tziner et al., 2007; Ford et al., 1997; Stevens and Gist, 1997), transfer of training (Velada et al., 2007), and application of training technology (Christoph et al., 1998). Self-efficacy has also been shown to mediate several important training relationships, such as between conscientiousness and learning (Martocchio and Judge, 1997) and training and adjustment (Saks, 1995), as well as many other individual characteristics and transfer, both analogical and adaptive (Bell and Kozlowski, 2008).

#### 5.2.3 Goal Orientation

Goal orientation, defined as the mental framework that determines behavior in different goal-oriented environments, has been heavily researched as to its

**Table 2  Steps in Developing Training Objectives**

| Step | Guidelines |
|---|---|
| 1. Review existing documents to determine job tasks and competencies required. | Examine your sources:<br>• Performance standards for the organization<br>• Essential task lists<br>• Past training objectives<br>• SMEs and instructors to mine their previous experiences |
| 2. Translate identified competencies into training objectives. | Include objectives that:<br>• Specify targeted behaviors<br>  • Use ''action'' verbs (e.g., ''provide,'' ''prepare,'' ''locate,'' and ''decide'')<br>  • Outline specific behavior(s) that demonstrate the appropriate skill or knowledge<br>  • Can be easily understood<br>• Clearly outline the conditions under which skills and behaviors should be seen<br>• Specify standards to which they will be held when behaviors or skills are performed or demonstrated<br>  • Make sure that standards are realistic<br>  • Make sure that standards are clear.<br>  • Make sure that standards are complete, accurate, timely, and performance rated |
| 3. Organize training objectives. | Make sure to categorize:<br>• General objectives that specify the end state that trainees should attain/strive for<br>• Specific objectives that identify the tasks that trainees must perform to meet the general objectives |
| 4. Implement training objectives. | Use training objectives to:<br>• Design exercise training events (e.g., scenarios)<br>  • Use events as opportunities to evaluate how well trainees exhibit training objectives<br>• Develop performance measurement tools (e.g., checklists)<br>• Brief trainees on training event |

*Source:* Originally published in Salas et al. (2006b).

impacts on learning and transfer (Kozlowski and Bell, 2006; Phillips and Gully, 1997; Ford et al., 1992). Early theorizing on the issue dichotomized goal orientation between either mastery or performance orientations (see Dweck and Leggett, 1988). Mastery orientation has been shown to impact knowledge-based learning outcomes (Ford et al., 1997), self-efficacy (Phillips and Gully, 1997), motivation and satisfaction (Orvis et al., 2009), as well as overall training effectiveness (Orvis et al., 2009; Klein et al., 2006; Tziner et al., 2007). On the other hand, performance orientation has exhibited many negative effects on training (Orvis et al., 2009; Tziner et al., 2007). Debate exists as to the nature of goal orientation, however. Research has explored whether goal orientation is a state or a trait (Stevens and Gist, 1997), a multidimensional construct (Elliot and Church, 1997; Vandewalle, 1997), or even mutually exclusive (Buttom et al., 1996). Regarding this last question, some researchers have encouraged another level of goal orientation, approach–avoid (see Elliot and McGregor, 2001). While studies on the simple mastery–performance dichotomy are much more plentiful, the mastery–approach orientation has been shown to be the learning orientation most strongly associated with positive training attitudes (Narayan and Steele-Johnson, 2007).

### 5.2.4 Motivation

Trainee motivation can be defined as the direction, effort, intensity, and persistence that individuals exert towards training-relevant behaviors and learning before, during, and after training (Naylor et al., 1980, as cited in Goldstein, 1993). Learning-oriented motivation has been shown to affect acquisition and retention of knowledge as well as willingness to participate (Tziner et al., 2007; Klein et al., 2006; Martocchio and Webster, 1992; Mathieu et al., 1992; Tannenbaum and Yukl, 1992; Quinones, 1995). Motivational levels can be impacted by individual (e.g., self-efficacy) and organizational (e.g., culture) characteristics as well as characteristics of the training itself (i.e., some training may be more interesting/motivating than others). Self-efficacy toward training and instrumentality of training have both been shown to positively impact training effectiveness (Rowold, 2007; Roberson et al., 2009). Training instrumentality, or the expected usefulness of training content, tends to predict motivation to transfer (Chiaburu and Lindsay, 2008; Noe, 1986); pretraining motivation has similarly been shown to predict positively training effectiveness and trainee reactions (Burke and Hutchins, 2007; Tannenbaum et al., 1991; Mathieu et al., 1992).

While the literature on training motivation has been criticized as lacking in conceptual precision and specificity (Salas and Cannon-Bowers, 2001), strides have been made toward identifying the underlying processes of trainee motivation. Colquitt and colleagues (2000) greatly bolstered this field with their meta-analysis on the topic which revealed that motivation to learn is impacted by individual (e.g., self-efficacy, cognitive ability) and situational (e.g., organizational climate, supervisory support) characteristics. These findings have

been further replicated in more recent empirical studies (Major et al., 2006; Nijman et al., 2006; Scaduto et al., 2008) as well as meta-analyses (Sitzmann et al., 2008). Burke and Hutchins (2007) note that motivation to transfer is impacted by motivation to learn, self-efficacy, utility reactions, and organizational climate. Furthermore, they note inconsistencies in the literature as to whether intrinsic or extrinsic motivation is more effective in encouraging transfer. The multifaceted nature of motivation thus demands attention when designing training.

### 5.3 Organizational Characteristics

While individual characteristics do indeed affect training outcomes, organizational characteristics (i.e., the organization to which training is relevant) also play a role in determining training outcomes. These may include organizational culture, policies and procedures, miscellaneous situational influences (e.g., improper equipment), and prepractice conditions (see Salas et al., 1995). These organizational characteristics must be accounted for when designing training.

#### 5.3.1 Organizational Culture

Since the 1980s, when the term *organizational culture* began to be discussed in the literature, research has suggested that organizational culture is vital to an organization's success (Guldenmund, 2000; Glendon and Stanton, 2000). Organizational culture has been defined as "a pattern of shared basic assumptions that the group learned as it solved its problems of external adaptation and internal integration" (Schein, 1992, p. 12). These assumptions and other aggregate behaviors and social norms must then be inculcated to trainees, such that employees will think and respond to situations similarly and harmoniously (Schein, 1992). Organizational leaders play a vital role in transmitting preexisting organizational cultures to new employees through the socialization process (Burke, 1997). For example, supervisory support of training has been shown to foster an organizational culture supportive of training and learning, which in turn positively impacts training outcomes (Nijman et al., 2006).

#### 5.3.2 Policies and Procedures

Out of an organization's policies and procedures arise its organizational culture. Broad requirements and expectations (e.g., job performance, interpersonal relations) comprise an organization's policies (Degani and Wiener, 1997), while more specific guidelines as to how to follow the policies comprise an organization's procedures. For example, if a company has a policy that emphasizes employees' punctuality, management may implement procedures to assist/ensure this policy is met (e.g., time card, consequences for tardiness). Training programs can be designed to inform trainees of these policies and procedures in such a way that they will more effectively perform them on the job. There can be issues, however, when implicit social norms are contraindicative of training goals (e.g., Hofmann and Stetzer, 1996). Continuing the punctuality example above, if there are social situations that allow (or rather, do not disallow) employees

to have excessive flexibility with their punctuality (e.g., having other employees punch time cards), the written policies, procedures, and training will be less effective. This is a real problem with real consequences that has been dramatically evinced within the oil industry (Wright, as cited in Hofmann et al., 1995). Unfortunately, organizations often do not engage in behaviors supportive of training before or after training when targeted behaviors are most likely to subside (Saks and Belcourt, 2006). If policies and procedures are to be enacted throughout an organization, policies and procedures must be implemented within management that support the development of lower level policies and procedures.

#### 5.3.3 Situational Influences

Several aspects of the training situation have been shown to have significant effects on trainees' motivation to learn as well as transfer of knowledge from training to the job: framing of training, work environment, and previous training experience. Framing of training typically refers to whether attendance is voluntary or mandatory; research has shown trainees respond more positively to voluntary framing (Baldwin and Magjuka, 1997). Training can also be framed from remedial/advanced or mastery/performance (recall the discussion on learning goal orientation). Using these frames can have important effects on reactions to training (Quinones, 1995, 1997; Kozlowski and Bell, 2006). Even doing something as simple as labeling training as an "opportunity" can have positive effects (Martocchio, 1992). The work environment and perceptions thereof can also have an effect on motivation and training transfer (e.g., Goldstein, 1993). Work environment can refer to organizational culture and supervisory support (as previously discussed); it can also be more tangible, such as lack of materials/information (e.g., Peters et al., 1985) or even lack of time to apply learning (Festner and Gruber, 2008; Lohman, 2009). Lastly, prior training experiences, especially negative ones, have been shown to negatively affect further training endeavors (Smith-Jentsch et al., 1996aa). These conditions must all be taken into account when designing a training program.

#### 5.3.4 Prepractice Conditions

Certain conditions have been shown to better prepare trainees for practice during training; these are termed prepractice conditions. Research has well evinced the positive impact practice has on skill acquisition, but not all practice is the same. However, due to the complex nature of the skill acquisition process, simple task exposure or repetition is not enough (Schmidt and Bjork, 1992; Shute and Gawlick, 1995; Eherenstein et al., 1997). Cannon-Bowers and colleagues (1998) provide a good review of conditions antecedent to enhanced utility and effectiveness of practice. This review suggests that prepractice interventions, such as preparatory information, advanced organizers, or metacognitive strategies, can help prepare trainees for and encourage learning within training. For example, Frey and colleagues (2007, as cited in Orvis et al., 2008) found that providing prepractice training on novel

instructional media increased performance and learning, especially for novel users.

## 5.4 Practice Opportunities

A vital aspect of any training program is the practice opportunities offered to trainees during training. Research has shown that practice under conditions (either simulated or real) that differ from the end-goal task will improve on-the-job performance by developing meaningful contexts and knowledge structures (i.e., mental models) (Satish and Streufert, 2002). Scripted practice scenarios ensure that the necessary competencies are being practiced and will allow for easier and better performance assessment. Beyond simply increasing performance, practice should also improve overall task-relevant learning. In a 2006 meta-analysis of training literature, Sitzmann and colleagues found that Web-based instruction led to more learning than classroom instruction, not because of the media itself but because it allowed learners more control and to *practice* the material at their own pace. That more practice opportunities lead to more learning is a finding replicated in other studies as well (Festner and Gruber, 2008; Goettl et al., 1996). Research has also been done on the scheduling of practice opportunities as well as introducing variations in practice difficulty (Schmidt and Bjork, 1992; Ghodsian et al., 1997). These studies have shown that varying the order of tasks during practice, providing less feedback (both in quality and quantity), and introducing alterations in the specifics of the tasks being practiced all led to enhanced retention and generalization, despite exhibiting initial decreases in task performance. This suggests that varying practice opportunities in such a fashion actually leads to deeper learning. Shute and Gawlick (1995) confirmed these findings for computer-based training. One thing to note: When the training medium is unfamiliar to the trainee (e.g., video games), introducing difficulty in practice will be less beneficial than for experienced users (Orvis et al., 2008), as they expend cognitive resources on learning the medium rather than the task.

## 5.5 Feedback

Constructive and timely feedback is important to the success of any training program. If not for feedback, trainees would have no metric by which to determine where improvements are needed or if performance is sufficient (Cannon-Bowers and Salas, 1997). Research has consistently shown that feedback leads to increases in learning and performance (Burke and Hutchins, 2007). However, effective feedback must meet three criteria. First, feedback should be task/training performance based but not critical of the person. Second, feedback should provide information and strategies on how to improve learning strategies so that performance expectations are met. Finally, feedback must be seen as meaningful at all applicable levels (i.e., individual and/or team). Feedback strategies may also vary depending on task complexity; "scaffolding" is a technique that involves gradually *reducing* the amount of feedback provided so as to encourage self-monitoring of errors (van Merrienboer et al., 2006). However, such techniques are

still in need of further research as to their effectiveness and generalizability (Burke and Hutchins, 2007).

## 5.6 Instructional Strategies

A key element of training design is the instructional strategy selected. Characteristics of the trainee and the organization must be considered so as to select the most appropriate strategy. There exist a myriad of instructional strategies that may be effective for training both individuals and teams; Table 3 summarizes many of these strategies.

Four basic guidelines should guide training designers seeking to develop effective training: (1) Information or concepts to be learned must be presented, (2) necessary KSAOs to be learned/acquired must be demonstrated, (3) opportunities for practices should be provided, and (4) feedback during and after training must be given to trainees. As is evident from Table 3, much research has been done toward investigating training strategies, but no single method is totally effective at meeting all training needs. Accordingly, research continues on how to present targeted information to trainees, based on factors such as organizational resources and training needs. Current research, then, seeks to develop and test validated, user-friendly, interesting, and interactive training tools with high return on investment (ROI) (e.g., Bretz and Thompsett, 1992; Baker et al., 1993; Steele-Johnson and Hyde, 1997).

As organizations continue to move into the future, several issues affecting training and training strategies will continue to increase in saliency: (1) reliance on workplace teams, (2) advances in technology, and (3) globalization. Therefore, the instructional strategies discussed here are organized accordingly. However, future training strategies cannot simply acknowledge and incorporate these factors while ignoring three universal training issues: (1) adaptability, (2) feedback, and (3) interactivity. *Adaptability* refers to trainees' acceptance of change and adoption of flexible knowledge structures (i.e., mental models); the breakneck advances in technology demand adaptability of today's employee. *Feedback* simply refers to the need for training programs to provide constructive information regarding performance to trainees; this allows them to adjust their efforts to better their performance. Training programs should increasingly incorporate *interactivity*; while this element has been shown to have some positive effects, it should not be applied blindly, as learner control inherent in interactive programs has been shown to have inconsistent effects (Aguinis & Kraiger, 2009). Recent reports have exhibited the upward trend in technology use in training—approximately 26% of learning hours in 2004 were technology based, and this increased to 32% as recently as 2007 (Paradise, 2008), but there clearly are more arenas where interactive training programs might be utilized.

### 5.6.1 Teams

Over the past half century, organizational theory and business practices have changed a great deal. A major facet of these changes has been the upwards trend of workplace teams and their use in health care,

**Table 3  Instructional Strategies**

| Strategy | Definition | References |
|---|---|---|
| Technology-based strategies | | |
| Simulation-based training and games | Provides practice opportunities through technology. Has the potential to create realistic scenarios by simulating terrain, events, and social interactions; has been applied to military, business, medical, cross-cultural, and research settings. Varies regarding fidelity, immersion, and cost. | Tannenbaum and Yukl (1992), Marks (2000), Bell et al. (2008) |
| Behavior modeling training | Presents the trainee with examples (positive and negative) of targeted behaviors. Can be done in a variety of settings but is often done electronically. Essentially a combination of simulation and role playing and varies as to the degree of interactivity. | Taylor et al. (2005) |
| Distance learning | Allows instructors and students physically or synchronously separated to interact. Utilizes technologies such as the Internet, CD/DVD-ROM, videoconferencing, and text messaging. Includes both electronic and nonelectronic forms of training. | Moe and Blodget (2000), Orvis et al. (2009) |
| E-learning | Same concept as distance learning but is strictly electronic and generally more interactive. Uses the Internet, text messaging, virtual classrooms, online collaboration, and a host of other technologies to quickly connect learner and instructor. Time-pressured employees are typically ideal candidates for this training. | Kaplan-Leiserson (2002) |
| Learner control | Allows trainees to exercise personal control over certain aspects of the training (e.g., content, delivery structure, pacing). A key component of many distance, e-learning, and other technology-based instructional strategies. | Wydra (1980), Brown and Ford (2002), Sitzmann et al. (2006) |
| Scenario-based training | Similar to simulation-based training, in that it provides complex and realistic environments. It is more targeted, however, as it embeds task-relevant events with behavioral triggers into the program. The program then monitors performance and provides feedback on the task and processes within the scenarios. | Fowlkes et al. (1998), Oser et al. (1999a,b), Salas et al. (2006) |
| Collaborative learning | Utilizes technology to enable groups to train together. This may happen with collocated or separated groups. Emphasizes group interaction and deemphasizes tasks. | Arthur et al. (1996, 1997), Marks et al. (2005) |
| Error training | Allows, encourages, induces, or guides trainees to make errors within the program. Trainees are shown the consequences of failure and provided with feedback. Especially necessary in situations when posttraining errors are likely or in highly dynamic and complex training areas. | Dormann and Frese (1994), Ivancic and Hesketh (1995), Keith and Frese (2008) |
| Stress exposure training | Informs trainees as to the potential stressors inherent in the targeted task as well as the relationship between stressors, trainee affect, and performance. Trainees also receive information on coping strategies. Training may also incorporate graduated exposure to actual stressors to desensitize trainees to potential stressors. | Johnston and Cannon-Bowers (1996), Driskell and Johnston, (1998), Driskell et al. (2008) |
| On-the-job training | Training occurs in the same environment in which actual task behavior will eventually be performed. Experts instruct novices, who then practice relevant tasks under supervision of the expert. Includes both apprenticeship and mentoring models. | Goldstein (1993), Ford et al. (1997), Munro (2009) |
| Team-based strategies | | |
| Team coordination training | Provides opportunities for teams to practice workload distribution, implicit and explicit communication, backing up behaviors, interpersonal relations, with the goal being maximal team coordination. This is becoming increasingly important as teams are on the rise even as face-to-face interactions are declining. | Bowers et al. (1998), Entin and Serfaty (1999), Serfaty et al. (1998) |

**Table 3 (continued)**

| Strategy | Definition | References |
|---|---|---|
| Cross-training | Rotates team members through the tasking of other team members. The goal is to provide team members with a better understanding of role requirements and responsibilities and to coordinate workload more effectively. Goal is improved shared mental models and understanding of technology usage across team members. A recent meta-analysis showed this strategy to be essentially ineffective. | Volpe et al. (1996), Salas et al. (1997a, 2007) |
| Team self-correction training | Fosters awareness of team processes and effectiveness so that team members can evaluate and correct their behaviors in the future. This is not only personal awareness but also awareness of the behaviors of other team members. Instructs teams to provide constructive feedback; may help mitigate errors due to miscommunications that naturally occur when adopting new technologies. | Blickensderfer et al. (1997), Smith-Jentsch et al. (1998), Salas et al. (2007) |
| Distributed team training | Allows physically and synchronously separated teams to develop competencies that optimize teamwork. These strategies help team members adjust their interactions to the complexities of separation, technology, and a plethora of other logistical issues. | Townsend et al. (1996), Carroll (1999) |
| Internationalization-based strategies | | |
| Individual-level strategies | Multicultural training was initially centered on either didactic (information-giving) or experiential instructional strategies. Cultural awareness and attribution training have increased in popularity as well for multicultural training. | Deshpande and Viswesvaran (1992), Kealey and Protheroe (1996) |
| Attribution training | Instructs trainees in various strategies that allow them to see the source of behavior (i.e., make attributions) similarly as individuals from a given culture. This also provides a better understanding and appreciation for other cultures. | Befus (1988), Bhawuk (2001), Littrell and Salas (2005) |
| Cultural awareness training | Emphasizes understanding the personal culture of the trainee, along with the various biases, norms, and thought patterns involved in that culture. Assumes that an awareness of one's own culture leads to a greater appreciation for and understanding of other cultures. | Bennett (1986), Befus (1988), Collins and Pieterse (2007), Thomas and Cohn (2006) |
| Didactic training | Provides trainees with various culturally relevant facts such as living and working conditions, cultural dimensions, values, logistics of travel, shopping, and attire and even food. Geographic, political, economic, and social structure information is also conveyed. | Kealey and Protheroe (1996), Morris and Robie (2001), Sanchez-Burks et al. (2007) |
| Experiential training | Emphasizes experiencing the aspects and consequences of cultural differences in various scenarios. Often occurs in the form of simulations but can even include face-to-face exposure to the targeted culture or role-playing exercises. Not only improves trainees' intercultural competence but also can serve a purpose similar to attribution training. | Kealey and Protheroe (1996), Morris and Robie (2001), Fowler and Pusch (2010) |
| Team-level strategies | Multicultural team training strategies train teams of culturally diverse individuals to optimize their performance. Five domains typically characterize multicultural team training programs: (1) enhancing specific aspects of performance through improvement of general team characteristics, (2) using a traditional team performance framework, (3) applying specific training tools and feedback, (4) utilizing a multimethod training method, and (5) training in a limited time frame. | Salas and Cannon-Bowers (2001) |
| Team leader training | Trains leaders on two domains of multicultural team leadership: (1) leadership and (2) culture. Often these are taught asynchronously, but some training programs have attempted to train leaders on both of these simultaneously. | Kozlowski et al. (1996), Thomas (1999), Burke et al. (2005) |

**Table 3** (*continued*)

| Strategy | Definition | References |
|---|---|---|
| Team building | Incorporates team members into organizational- or team-level change development and implementation. Forces team members to interact strategically. Emphasizes one of four team-building goals: goal setting, interpersonal relationships, problem solving, and role clarification. | Dyer (1977), Beer (1980), Buller (1986), Salas et al. (1999), Shay and Tracey (2009) |
| Role playing | Requires trainees to interact with other team members through the use of scripted scenarios. Progression through the roles can make team members aware of (1) their own culture (i.e., enculturation) or (2) other cultures (i.e., acculturation). Role playing can also support both of these goals when trainees experience multiple roles. | Bennett (1986), Roosa et al. (2002) |

*Source:* Adapted from Salas et al. (2006b).

government, manufacturing, and the military (e.g., Salas et al., 2008; Tannenbaum and Yukl, 1992; Guzzo and Dickson, 1996). Increasing workplace complexity has driven the need for teams (Salas et al., 2008), and advancements in technology, geopolitical stability, and research have enabled organizations to meet this need on an international level. This trending toward teams and organizational adaptability has leveled many traditional hierarchical structures (Kozlowski and Bell, 2002; Zaccaro and Baker, 2003). The vast majority of organizations utilize some form of team structure (Banker et al., 1996); similarly, in a 2001 survey by Fiore and colleagues, 80% of surveyed workers reported being members of at least one team. The generally held philosophy among organizations is that teams are the answer to many workplace problems; thus, as organizational life becomes increasingly complex, teams are implemented more readily.

This practical trending toward teams has led to a significant research emphasis on teams and team training (Salas et al., 2008; Salas and Cannon-Bowers, 2001). One of the major drivers of team training research has been in the commercial and military aviation communities (e.g., Salas et al., 1995). Crew resource management (CRM; Shuffler et al., 2010; Wiener et al., 1993) is specifically aviation relevant; other strategies such as cross-training and team self-correction training have also developed, validated in other organizational settings, and been widely discussed in the literature. While an in-depth discussion of these strategies would be essentially redundant, a summary of many team training strategies and additional resources has already been provided in Table 3. For a more exhaustive summary of team training strategies, we encourage the reader to explore other influential pieces in the literature (see Salas and Cannon-Bowers, 2001; Wilson et al., 2005; Salas et al., 2007, 2008). In addition to some of the more common strategies, team training strategies will increasingly be impacted by technology as well as globalization. These two concepts are explored at length in the following two sections.

### 5.6.2 Technology

As the strength and availability of technology increase at a breakneck pace, it is no surprise that training designers

have increasingly incorporated technology into their programs. However, technology-based instructional strategies currently only comprise little more than 30% of total practices (Paradise, 2008). With the recent and ongoing explosion of technological advancement in recent years, organizations must address not only incorporating technology into their training programs but also training employees to effectively utilize the new technologies. These issues, however, must not prevent organizations from adopting new and beneficial technologies. At one time, using "cutting-edge" technologies such as e-mail, text messaging, and videoconferencing seemed cumbersome and unnecessary to some. Technologies such as these are now considered the norm; we accept and expect them.

Recent surges in technology have allowed ever-increasing degrees of integration between technology (e.g., computer and Web based) and training (Goldstein and Ford, 2002; Rivera and Paradise, 2006). One of the alluring potentialities of intelligent (i.e., technologically based) tutoring systems is that it may reduce or eliminate the need for human instructors, especially given certain learning tasks. As early as 1995, Anderson and colleagues proposed that intelligent software could be programmed to monitor, assess, diagnose, and remediate performance for several tasks. More recently, it was shown that technology-driven tutoring software was superior to traditional instructional strategies when trainees were allowed time for practice because it provided opportunities for discovery (Brunstein et al., 2009). Learner control is a major component, and advantage, of many technology-based training programs. Allowing trainees to have control over certain elements such as training method, time for practice, and timing of feedback (e.g., Milheim and Martin, 1991) has been shown to have positive effects on trainee attitudes and motivation (e.g., Morrison et al., 1992) as well as learning and performance (Schmidt and Ford, 2003). In fact, in a recent meta-analysis, Sitzmann and colleagues showed that it was this element of control, and not the training media itself, that caused significant increases in learning (Sitzmann et al., 2006). Following is a brief summary of some of these technological instructional strategies as well as techniques to encourage and facilitate adoption of these new methods. The

technology of training, while still under much research, is an increasingly salient phenomenon. Organizations and training designers alike would do well to embrace it, given its considerable ability to improve upon the effectiveness and cost-efficiency of more traditional training methods.

***Simulation-Based Training and Games*** By far, simulation and game-based training have revolutionized the field of training, especially aviation and military training. In the late 1960s, theories regarding simulation-based training was new and unique (Ruben, 1999). At that point, training and learning were thought to occur primarily in a classroom setting, driven by books, articles, and lectures. However, technology-assisted training strategies are now readily accepted by nearly all and are often preferred by some trainees. Much of this has to do with the ability to actually practice targeted KSAOs, something which is often much more difficult in a classroom setting. Furthermore, the flexibility of many simulation programs means KSAOs can be practiced in a wide variety of settings and situations; this increases the likelihood of training transfer and generalization (Richman et al., 1995; Salas et al., 2006; Sitzmann et al., 2006). Accordingly, we are seeing, with increasing frequency, the use of simulations and computer games for instruction, practice, and feedback.

Training through simulation continues to be a popular method in business, education, and the military (Jacobs and Dempsey, 1993). The military is the main investor in simulation studies, having contributed $150 million to $250 million each year since the 1960s towards the field (Fletcher, 2009). Research has shown that simulation-based training strategies are effective, but the literature is unclear as to *why* it is effective. Several studies have offered preliminary data (e.g., Ortiz, 1994; Bell and Waag, 1998; Jentsch and Bowers, 1998), but explanatory models are yet to be widely accepted. Due to this paucity in the literature, it may be premature to say that simulation (in and of itself) leads to learning. Furthermore, a major confound exists in current research in that much of the training research evaluates trainee reactions while ignoring actual performance (see Salas et al., 1998). Some research, however, has suggested that the major advantage simulation-based training and other technology-based training methods have over traditional training strategies is the control they offer trainees, which has been shown to have positive effects on reactions, motivation, and performance (Orvis et al., 2009; Sitzmann et al., 2006). Other advantages are decreased learning time, complex but controllable practice environment, and the ability to prepare trainees for critical but rare events (Salas et al., 2009). However, further research is still required to provide a more thorough explanatory framework as to the effectiveness of simulations, even as the method becomes more and more ubiquitous.

As previously suggested, unique and relevant practice opportunities are a major advantage of simulation-based training. Structuring training around relevant scenarios is a vital and effective instructional strategy for individuals and teams operating in complex, time-oriented environments. Simulations can train a wide variety of domain-relevant KSAOs. Flight simulation has consistently been the most common application of simulation-based training and has driven much of the research in training technology. Adaptive decision making (e.g., Gillan, 2003), discrimination, performance (Aiba et al., 2002), response time (Harris and Khan, 2003), performance under workload (Wickens et al., 2002), and team issues (Prince and Jentsch, 2001) have all benefitted from research grounded in flight simulators. Similarly, driving simulators have begun to leave the domain of simple evaluation and have become popular methods of all forms of driver training (e.g., Fisher et al., 2002; Roenker et al., 2003). Medicine has also adopted simulation-based training (e.g., the METI doll; METI, 2010) to train both individual- and team-based skills. Additionally, multicultural training has embraced simulation-based training to improve the effectiveness of cross-cultural interactions (see "Individual Cultural training Strategies" in Section 5.6.3). Second Life™, a free, user-driven computer game, has been adapted to train for cultural awareness and competence, for example (Fowler and Pusch, 2010). Other arenas where simulation-based training has shown great promise have been health care (e.g., Rosen et al., 2008; Steadman et al., 2006), transportation (e.g., Tichon, 2007a), and the military (Salas et al., 2006). For example, presenting train conductors with complex, life-threatening scenarios (e.g., workers on the track), and training them accordingly, will foster in trainees an ability to make complex decisions quickly when necessary (Tichon, 2007a).

Simulation-based training varies greatly with regard to cost, fidelity, and functionality. Early simulators were typically extremely low in physical fidelity (i.e., realism), but with recent advances in technology, simulators and virtual environments now have the ability to emulate environmental characteristics such as terrain, equipment failures, motion, vibration, and much more. While simulations high in physical fidelity may better engage trainees in training and may also help in preparing for the details of complex situations (Bell et al., 2008), systems low in physical fidelity can often just as effectively represent targeted KSAOs (e.g., Jentsch and Bowers, 1998). Furthermore, these low-fidelity systems can still have positive effects on immersion, engagement, motivation, and learning over lecture and/or text-based instructional strategies, especially when the system has some sort of game element (i.e., interactivity, competition) (Bell et al., 2008, Ricci et al., 1995).

Despite the increasing availability of high-fidelity simulations, the trend recently has been toward lower fidelity, computer-based simulation systems, even when training complex skills. Beyond cost effectiveness, another practical rationale for this is that simulators low in physical fidelity can actually excel in psychological fidelity (i.e., representation of constructs and processes implicit in performance). These training systems then should actually result in more skills transfer and behavioral generalization after training (e.g., Bell et al., 2008; Gopher et al., 1994). Driven by such thinking, researchers have examined simple off-the-shelf computer games for training complex skills;

studies have shown that these low-fidelity simulations can indeed improve performance in highly complex situations (e.g., Fong, 2006; Gopher et al., 1994; Goettl et al., 1996; Jentsch and Bowers, 1998). It is important to note that these simulations were programmed around a skill-oriented task analysis; task analysis is vital to ensure the psychological fidelity of simulation training systems.

While simulation-based training has increased in prevalence consistently throughout recent years, the unfortunate truth is that many simulation systems are designed with an eye first to technology and only secondarily to more important learning factors such as cognition, training design, and effectiveness (Cannon-Bowers and Bowers, 2008; Salas et al., 1998). Researchers have yet to consistently integrate the science of training practically with simulation training (Salas and Kozlowski, 2010; Bell et al., 2008, Cannon-Bowers and Bowers, 2008). Incorporating what we know about training and learning (e.g., event- and competency-based training approaches) with simulations has been suggested as a way to remedy this situation (Cannon-Bowers et al., 1998; Fowlkes et al., 1998; Oser et al., 1999bb). That is, simulations and computer games should utilize training objectives, diagnostic measures of processes and outcomes, feedback, and guided practice. Though more research needs to be done in this area, preliminary studies (e.g., Tichon, 2007b; Colegrove and Bennet, 2006) have shown that, by designing simulations around specific events and measuring specific task performance, both learning and training analysis benefit greatly. See the review by Salas and colleagues (2008) for a list of principles for developing scientifically grounded simulation-based training. Finally, while it has been stated that an empirically verified framework regarding the processes involved in simulation-based training is lacking, Cannon-Bowers and colleagues (1998) developed a general model of simulation- and scenario-based training that includes the major components necessary for any good training program (see Figure 5). However, there has yet to be research done toward developing a comprehensive model regarding the antecedents

and processes of simulation-based training effectiveness. Promising steps have been made toward this end by Wilson and colleagues (2009) in their review of the literature on the processes of learning; their work theoretically links aspects of learning to common components of simulations and gaming. This work, then, is the closest there currently is to a comprehensive theory of simulation effectiveness and should serve as a jumping-off point for future research.

***Behavior Modeling Training***  Behavioral role modeling, or behavior modeling training, is another type of simulation that has been researched heavily in recent years. This technique has been shown to be effective in training for desirable outcomes such as organizational citizenship behaviors (e.g., Skarlicki and Latham, 1997), assertiveness (Smith-Jentsch et al., 1996bb), and various health care behaviors (Schwarzer, 2008). Taylor et al. (2005) conducted a meta-analysis of 117 studies on behavior modeling training, which reveals several characteristics of effective training regimens. Behavior modeling training led to the most transfer when trainees (1) were exposed to positive and negative models, (2) participated in scenario development, (3) set goals in training, (4) were trained alongside their supervisors, and (5) were extrinsically motivated to transfer training. Additionally, when trainees were cued as to which behavioral skills to attend to, skill learning was greatest; similar effects were found for length of training. Finally, it should be noted that the effects of behavior modeling training declined with time for declarative and procedural knowledge but remained stable or increased for results-level outcomes (e.g., productivity).

***Distance Learning***  Advances in technology have clearly impacted the way training systems are designed, and this trend is projected to continue indefinitely. As of 2006, approximately 36% of the nearly $110 billion per year training industry was devoted to technology-oriented training systems, and of this massive industry, 60% was self-paced online learning (Rivera and
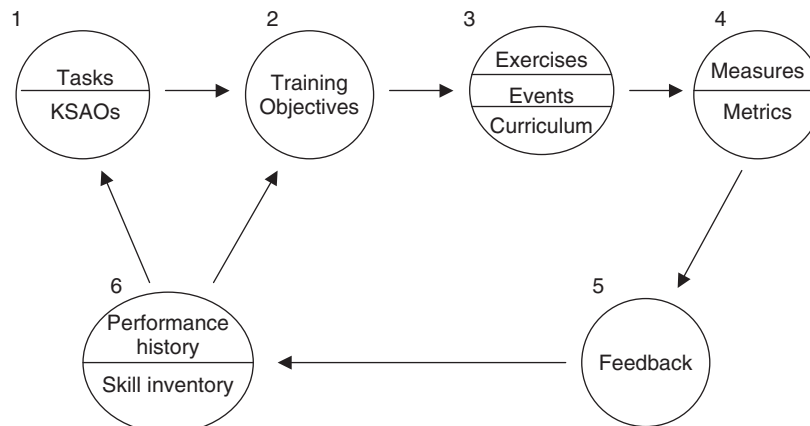


**Figure 5**  Components of scenario-based training. (From Cannon-Bowers et al., 1998.)

Paradise, 2006). While traditional (i.e., classroom-based) training methods continue to be the norm, organizations are with increasing frequency turning to technologies such as videoconferencing, electronic performance support systems, and online Internet/intranet training programs to supplement their development efforts. And while often used interchangeably in the literature, *distance learning* is the result of *distance training*. Distance training is a broad conglomeration of training strategies that are all characterized by a trainer/trainee separation of time and/or space. This broad training strategy encompasses several other methods discussed subsequently.

As with simulation-based training, there is a temptation for organizations to develop and/or adopt distance training methods without paying mind to the science behind the training. The increasing globalization of many organizations has encouraged the development and utilization of distance training strategies, but convenience is not an excuse for use. For training to be beneficial, it must be designed in accordance with the research on training and training effectiveness. Research is continually shedding more light on the nuances of distance learning. One of the major differences between distance learning and classroom learning is the issue of motivation. Without an instructor on hand to enforce (or encourage) behavior, much of the impetus of distance learning falls on the trainee. Accordingly, elements of training such as learner orientation (e.g., Klein, et al., 2006), learner control (e.g., Sitzmann et al., 2006; Orvis et al., 2009), and trainee reactions and self-efficacy (e.g., Sitzmann et al., 2008; Long et al., 2008) have been emphasized as key components of trainee motivation and performance in distance settings.

Other theoretical and logistical issues must be considered when designing and implementing distance training systems. Some of these include the importance of trainer/trainee face-to-face interactions, how to answer questions and feedback when physically and synchronously separated, where the training should fall on the content richness continuum, and the degree of learner control that should be provided. To date, these issues have been dealt with in piecemeal fashion (Bell et al., 2008). Kozlowski and Bell (2007) have presented a framework for distributed (distance) learning system design that incorporates what we know about training design and cognition as well as many available technologies. However, further research is required to determine if this framework is sufficient to address all these issues or if more theory is needed.

***E-Learning*** Any training that is conducted electronically refers to e-learning, though the vast majority of e-learning is computer based (Kaplan-Leiserson, 2002). E-learning has been defined as "a wide set of applications and processes, such as Web-based learning, computer-based learning, virtual classrooms, and digital collaboration" (Kaplan-Leiserson, 2002, p. 85). E-learning, as with many technologically oriented training designs, offers many advantages but brings with it a few new issues to consider as well. Some of the advantages of e-learning are (1) flexibility to make training mobile or local, (2) ability to incorporate multiple instructional media, (3) cost efficiency of continuing and reissuing training, (4) provision of learner control, and (5) allowing for synchronous or asynchronous training. A few issues to consider when implementing e-learning are (1) trainee motivation, (2) trainee self-efficacy and experience with computers, (3) higher start-up costs than classroom instruction, and (4) the fact that trainees ideal for e-learning are typically the employees with the least time to devote to training (Brown and Ford, 2002; DeRouin et al., 2004; Kozlowski and Bell, 2007; Klein et al., 2006; Shee and Wang, 2008).

Technological advancements have enabled e-learning to meet today's increasingly time-crunched training demands. However, recent surveys on the state and prevalence of e-learning have revealed some interesting trends. For example, some of the more advanced aspects of e-learning, such as mobile delivery or simulation, were ranked as being "rarely or never" used (Rossett and Marshall, 2010); typically these were avoided not only for economic reasons but also because of the difficulty involved with technology adoption. Rather, the most prevalent trend was that training and education typically occur in the classroom, with e-learning used as a supplementary tool (Rossett and Marshall, 2010). One arena in which the technological advances associated with e-learning have been adopted to a greater extent is distributed team training. Globalization and the increasing reliance on teams have ushered the modern workplace into an era categorized not only by teams as previously discussed but also by distributed teams. That is, team members are frequently no longer colocated but are "mediated by time, space or technology" (Driskell et al., 2003, p.3). This disconnect requires team members to rely on various communication technologies such as e-mail, videoconferencing, telephone, or faxing (Townsend et al., 1996). This dispersion, however, creates obvious problems for traditional, classroom-based training methods. The military has been most saliently affected by these team trends and has therefore exerted the most influence in the development of distributed team training methods.

Military efforts have developed the distributed mission training (DMT) system, which consists of multiple elements that combine to provide trainees with real-time, scenario-based training. These elements (real, virtual, and constructive) require coordination and communication with both real and virtual teammates (Carroll, 1999); performance on these tasks is analyzed and recorded, and trainees also receive feedback. Another key element of DMT systems is trainees' ability to access prior performance statistics from an online file and continue training from a save point. The development of DMT has resulted in two main training platforms that incorporate simulated training scenarios, virtual teammates, cognitive agents, and actual military personnel: Synthetic Cognition for Operational Team Training (SCOTT) (Zachary et al., 2001) and Synthetic Teammates for Realtime Anywhere Training and Assessment (STRATA) (Bell, 2003). While much of DMT research is military specific, the underlying principles regarding its development are relevant to any variety of distributed team training (Bell, 1999).

*Learner Control* As has already been intimated, learner control is a unique and important part of e-learning. Because learner control, which refers to "a mode of instruction in which one or more key instructional decisions are delegated to the learner" (Wydra, 1980, p.3), is a fairly general concept, it has been explored in the literature in a typically general fashion (Steinberg, 1977; Goforth, 1994; Brown and Ford, 2002; Orvis et al., 2009). The structure of e-learning lends itself perfectly to learner control. As a general rule, learner control is a major advantage e-learning has over more traditional training forms. For example, in their 2006 meta-analysis, Sitzmann and colleagues showed that Web-based instruction was not more effective than classroom instruction; however, when it allowed for learner control, it led to significantly more learning than classroom instruction. It has also been shown that the ability to control one's learning progress may be the most important aspect of the e-learning process (Shee and Wang, 2008). On the other hand, several studies have shown a negative relationship between learner control and training effectiveness (e.g., Tennyson, 1980; Murphy and Davidson, 1991; Lai, 2001). Content relevance has been suggested as a reason for these negative effects (DeRouin et al., 2004); trainee motivation also often has strong effects on the relationship between e-learning and training effectiveness (Klein et al., 2006).

*Collaborative Learning* Training can be augmented through collaborative learning strategies. Collaborative learning refers to any training situation where trainees receive training in groups. This refers not necessarily to team training (where teams are trained for team-based competencies) but simply to training situations where groups of trainees collaborate to learn. Certain aspects of the group experience facilitate learning more than individual learning efforts, especially when training efforts are targeted at increasing existing knowledge (Toomela, 2007). Dyadic tools that relied heavily on technology benefited military and nonmilitary pilots and navigators; however, when trainees were aversive to interaction, these benefits became nonsignificant (Arthur et al., 1996, 1997). Collaborative learning can provide opportunities for social learning (Shebilske et al., 1998) and can significantly reduce resource demands on the instructor (Shebilske et al., 1992). However, in collaborative distance and e-learning settings, the instructor–trainee interaction may be even more important than in colocated training settings (Marks et al., 2005).

*Error Training* As technology becomes more and more prevalent and workers shift from active players to passive observers in the work cycle, error training has become more and more prevalent in practice and research. Error training differs from proceduralized or error-avoidant training in that it encourages exploration and occasionally induces errors in the training process but also provides feedback that assists the trainee in improving performance (Frese and Altman, 1989; Heimbeck et al., 2003; Lorenzet, Salas, and

Tannenbaum., 2005). This better simulates the complexity of the real world, where employees are required to respond to variant situations, including the consequences of their own errors. Furthermore, the feedback provided in error management includes strategies for avoiding future mistakes. The advantages of error management over error avoidance are well-document in individual studies (e.g., Dormann and Frese, 1994; Ivancic and Hesketh, 1995) as well as in a recent meta-analysis of 24 empirical studies (Keith and Frese, 2008).

Error occurrence and error correction are the two main elements of error management training (Lorenzet et al., 2005). There are four ways the training designer can approach error occurrence in a program: (1) avoid, (2) allow, (3) induce, or (4) guide. Error-avoidant training guides the trainee through the program so as to either minimize or totally avoid errors; generally this is not beneficial for transfer of training, but in situations where the knowledge is procedural and static, this is actually the most effective approach (Keith and Frese, 2008). Allowing (but not inducing) errors, or exploratory learning, is helpful for transferring learning but can be difficult for trainers to anticipate where errors will occur, making the feedback process difficult (Gully et al., 2002; Keith and Frese, 2008). Inducing errors integrates a level of dynamism into the program such that changes are likely to induce errors; however, there is still no guarantee that errors will occur (Dormann and Frese, 1994). Guided error occurrence involves intentionally guiding trainees into a particular error, then providing strategies for avoiding that error. A recent meta-analysis has shown any kind of error training (i.e., *allow, induce,* and *guide*) to be more beneficial than error-avoidant training; furthermore, it was shown that error training strategies that involved inducing or guiding errors were more effective than simply allowing errors to occur (Keith and Frese, 2008).

Traditionally, two approaches to error correction have been utilized in designing error management training programs. Exploratory training design generally involves allowing trainees to struggle and work through errors until they solve any issues (Frese et al., 1991). This allows trainees to develop individualized strategies for addressing errors, but if trainees attribute errors incorrectly or come up with inappropriate strategies for error management, this may negatively impact adaptive transfer (Keith and Frese, 2008). The other approach is targeting feedback to correct errors (i.e., supported correction); feedback can be delivered either electronically or through a human instructor (Lorenzet et al., 2005). In a recent meta-analysis, Keith and Frese (2008) showed that providing error management instructions is significantly more effective than simple exploratory training.

A few general guidelines have been supported regarding error management training. Guided error training along with supported correction is generally the most effective error training (Lorenzet et al., 2005, Keith and Frese, 2008). While this tends to decrease within-training performance, trainees are better able to apply what is learned in training to different situations (Keith and Frese, 2008). This type of training also

leads to higher posttraining self-efficacy levels (Lorenzet et al., 2005). One caveat: Error-free training is equally effective (and more efficient) when training individuals for proceduralized, static tasks (Keith and Frese, 2008).

***Stress Exposure Training*** High-stress environments often add another layer of complexity onto training; the demands inherent in such situations can negatively impact trainees' cognitive resources and other abilities (Driskell et al., 2008). Stress exposure training (SET) incorporates increased situational demands into the training program so as to provide the KSAOs necessary to deal with real-world stressors (Driskell and Johnston, 1998). This form of training is especially important in situations where errors are likely to have major (and even life-threatening consequences). SET involves three stages of training: (1) identify common stressors and their effect on the job/task, (2) train trainees to deal with these potential stressors, and (3) gradually expose trainees to more of these stressors in a simulated training environment and provide feedback on their performance (Driskell et al., 2008). Importantly, SET has been shown to generalize across stressors (e.g., time pressure and high workload) and tasks (e.g., handling a hostage situation and pursuing a suspect) (Driskell et al., 2008).

***On-the-Job Training*** Job training can occur within or outside the actual job. Despite the increasing saliency of globalization and the trending toward various forms of distance learning, on-the-job training (OJT) remains one of the most prevalent forms of instruction (Paradise and Patel, 2009). OJT involves supervisors and/or trainers facilitating a training program in the exact physical and social contexts relevant to the specific job tasks (Wehrenber, 1987; Sacks, 1994; De Jong and Versloot, 1999). This training is distinguished from training that occurs on the job site but is separate from task performance (e.g., training workshops). OJT provides two main benefits to trainees and organizations: (1) training transfer is more likely to occur because it occurs in a totally relevant context and (2) tighter supervision within the training will prevent incorrect job strategies from being adopted by trainees.

Apprenticeship training is one form of OJT; it typically involves classroom instruction as well as supervision of job tasks and has traditionally been relegated to vocational trades (e.g., electricians). However, other types of organizations have begun to adopt an apprenticeship model (Goldstein, 1993). While the specifics of apprenticeship programs vary, typically they require trainees to be trained for a specified amount of time and to learn various essential skills before the trainee becomes a "journeyman" and is able to perform the job without supervision (Goldstein, 1993; Lewis, 1998; Hendricks, 2001). The European model of apprenticeship is markedly different from the U.S. method: Various European countries employ apprenticeship programs on a much larger scale and often incorporate higher learning into the apprenticeship experience (Steedman, 2005).

Mentoring is the other primary form of OJT and consists of a working relationship between job novice and expert (Wilson and Johnson, 2001). Industry reports indicate that formal and informal mentoring occurs in a large percentage of workplace settings (Paradise and Patel, 2009; Munro, 2009). The mentoring relationship has been shown to be related to many positive outcomes: increased communication and job satisfaction (Mobley et al., 1994; Forret et al., 1996), better leadership skills, more effective learning strategies, exhibiting more organizational citizenship behaviors (OCBs), and less stress and burnout (Munro, 2009). Mentoring is similar to apprenticeship in that the level of supervision allows the expert to model appropriate behaviors and monitor and correct the trainee's inappropriate behaviors (Scandura et al., 1996). Experts in the mentoring relationship also benefit; they practice vital skills while instructing trainees (Forret el al., 1996), receive greater satisfaction within the relationship, become aware of how others work, and form alliances and friendships within the workplace (Munro, 2009).

Informal learning is indirectly related to on-the-job training, as it refers to any time employees engage in learning that is not comprised of specific efforts at training. Jacobs and Park (2009) suggest a conceptual framework for delineating basic aspects of workplace learning on three domains: (1) on/off the job, (2) structured/unstructured, and (3) passive/active. These three domains interact to form eight kinds of workplace learning, with each form of learning requiring organizations, training designers, and learners to attend to different learning variables. For example, informal learning would primarily be categorized as unstructured and could either be active (i.e., conscious) or passive (i.e., during the course of regular work) and on or off (i.e., continuing education classes, personal study) the job. While not an official form of training, informal learning can serve the overall training function of improving employee performance, and organizations, leaders, and managers should be aware of a few things to facilitate informal learning in the workplace. The most impactful thing organizations can do to encourage informal learning is to create an environment conducive to continuous learning. This means (1) being supportive of employee efforts to informally learn, (2) providing opportunities to informally learn, and (3) making the tools and processes for learning available to employees that are interested in it (Tannenbaum et al., 2010). Other aspects of informal learning are relevant (Jacobs and Park, 2009; Tannenbaum et al., 2010) but are not the primary focus of this chapter on official training

### 5.6.3 Internationalization

It is a well known truth that the business world has become an increasingly global community (Tsui et al., 2007). Over 63,000 companies are involved on some level in the global marketplace (Chao and Moon, 2005), with many major companies (e.g., BP, Siemens, Honda) operating in over 100 countries (Adler and Gundersen, 2008). Constant, rapid advancements in technology coupled with an ever-increasing reliance on teams in organizations have enabled and necessitated (respectively) teams which are separated by time, space, and culture (Salas et al., 2008; Bell and Kozlowski,

2002). Unfortunately, the breakneck pace of such phenomena has yielded practice and research that are essentially out of sync (Shen, 2005). For example, while as far back as 2000 over 60% of companies offered some form of multicultural training, of those that did, a majority of this "training" consisted of one-day cultural debriefing seminars (Littrell and Salas, 2005). Insufficient expatriate training has resulted in millions of dollars lost for many organizations, due to early assignment termination, damaged relationships, burnout, and organizational departure (Littrell and Salas, 2005). And while expatriate studies continue to dominate the literature, strategies must be developed to address the issues inherent in all multicultural settings. The following sections, therefore, discuss research regarding the impact culture has on the individual (e.g., managers; see Shen, 2005) as well as the team.

Culturally oriented training has been given many names by practitioners and researchers: intercultural, diversity, multicultural, cross-cultural training, and so on. Some use these titles interchangeably, while others argue for differentiation (see Gudykunst et al., 1996). For the purposes of this chapter, we refer to culturally oriented training as *multicultural training*. The definition of multicultural training is then the development of behavioral, cognitive, and affective patterns in trainees that will make successful cross-cultural interactions more likely (Landis and Brislin, 1996; Morris and Robie, 2001). Like other forms of training, multicultural training should be intended not only to increase general cultural knowledge but also to improve skills, abilities, attitudes, and other characteristics relevant to cross-cultural success (Bhagat and Prien, 1996), which is typically attributed to personal adjustment, interpersonal adjustment, and task-relevant effectiveness (Littrell et al., 2006).

**Individual Cultural Training Strategies**   Multicultural training has typically focused either on directly providing cultural information to trainees or providing them with cultural experiences (Deshpande and Viswesvaran, 1992; Kealey and Protheroe, 1996). In addition to these methods, attribution, cultural awareness, cognitive behavior, interaction, and language training have all arisen as methods for multicultural training (Litrell and Salas, 2005). Following is a brief discussion of the most commonly used techniques:

1.  *Attribution Training*. Individuals tend to assign a reason to others' behaviors—this is known as attribution. The goal of attribution training is to help individuals better understand the attributions that individuals from other cultures may tend to make and to be able to take others' perspectives more effectively (Littrell and Salas, 2005; Befus, 1988). This not only provides a valuable set of skills but also serves to deepen the understanding and appreciation for other cultures of the trainee (Bhawuk, 2001).

2.  *Cultural Awareness Training*. The theory behind cultural awareness training is that in acknowledging one's own cultural background (and the biases inherent therein) cultural differences become more apparent as well as more understandable (Befus, 1988). This method may utilize a glut of training tools, from simulations, to role play, to group discussions (Thomas and Cohn, 2006; Bennett, 1986), with the ultimate goal being improved attitudes toward and strategies for interacting with culturally different others. An important characteristic of cultural awareness training is that the main goal inherent is the development of a culturally aware mindset. This mindset can manifest in either a static fashion (i.e., "I know how to be culturally aware") or a dynamic fashion (i.e., "I am always seeking to become more culturally aware"), but it is this mindset development that distinguishes cultural awareness training from other forms of multicultural training (Collins and Pieterse, 2007). Again, it is thought that by understanding what contributes to his or her own culture the trainee is better able to understand what should underlie other cultures.

3.  *Didactic Training*. Any form of cultural instruction that involves information giving can be considered didactic multicultural training. These may include information regarding travel, food, living conditions, and a host of other things. The format of didactic training is very similar to traditional classroom instruction in that it emphasizes cognitive goals (Bennett, 1986) such as understanding differences in political and economic structures between countries (Kealey and Protheroe, 1996; Morris and Robie, 2001). The ultimate goal of didactic multicultural training programs is to develop a working understanding of cultural factors to assist in future cross-cultural interactions (Littrell and Salas, 2005; Morris and Robie, 2001).

    Multicultural training can be delivered through various didactic media (e.g., informal briefings, cultural assimilators) (Littrell et al., 2006; Brewster, 1995) and are quite variable within. Informal briefings, for example, may be casual conversations and/or interviews with culturally experienced (and successful) individuals, such as past trainees or re/expatriates (Shen, 2005; Brewster, 1995; Kealey and Protheroe, 1996). "Briefings," however, are a fairly general term and can include lectures, videos, or information booklets on the targeted country/culture (Littrell and Salas, 2005; Grove and Torbiörn, 1993). Formal training strategies that are more in-depth may strategically cover topics such as religion, geography, history, and more (Littrell et al., 2006). Cultural assimilators consist of a variety of didactic methods that aim at teaching trainees to think like a target culture (Bhawuk, 1998, 2001). Trainees are presented with critical incident scenarios in which a response is required; in these tests, the goal is to select the response that an individual from the targeted culture would likely choose

(Kealey and Protheroe, 1996; Morris and Robie, 2001). Relational ideology training is a type of cultural assimilator that trains for emotion regulation and display in various cross-cultural settings (Sanchez-Burks et al., 2007).

4. *Experiential Training*. As the name suggests, experiential training involves having trainees learn by doing or *experiencing* a certain thing; this allows them to practice strategies and responses to various scenarios. In the case of multicultural training, it provides trainees with the skills and knowledge for interacting successfully with culturally different others (Kealey and Protheroe, 1996; Morris and Robie, 2001). Experiential training can not only impart intercultural skills and knowledge, but also serve the same purpose as attribution training, enabling trainees to see some things with the same mindset as culturally different others (Morris and Robie, 2001). Role playing, workshops, and simulations are among the most popular experiential training strategies for multicultural training (Littrell et al., 2006).

Simulations are excellent forms of experiential multicultural training, because they allow for safe practice opportunities (see the above section on simulations). BAFA BAFA is one of the most popular simulation games; in this game, trainees are assigned to one of two fictional, diametrically opposed cultures and then are tasked with interacting with individuals from the other culture (Bhawuk and Brislin, 2000). Since the development of BAFA BAFA, many more cultural simulation games have been developed. BARNGA (Steinwachs, 1995) is a card-based simulation in which trainees play card games with secretly different rules and are then required to play with other trainees with their own set of rules; this essentially emulates the process individuals must work through when dealing with others' cultural "rules." DIVERSOPHY is a cultural simulation that can be played on Second Life (a free, user-driven computer game). Intercultural simulation games have been applied in military, educational, business, and medical settings. For a deeper review of these simulations, see Fowler and Pusch (2010).

Initial reactions and anecdotal evidence tend to support the notion that individuals can be trained to be more cross-culturally savvy (Black and Mendenhall, 1990; Deshpande and Viswesvaran, 1992). However, there is a relative dearth of empirical evidence regarding the effectiveness of multicultural training, especially given the glut of work done on the theoretical and practical levels (Selmer, 2001; Selmer et al., 1998). This is especially vital, for as these training programs become more and more prevalent, it is becoming increasingly clear that effectiveness in these situations is not a guarantee. In a recent qualitative review of the literature, Kulik and Roberson (2008) showed that often multicultural awareness training programs were either marginally or not

at all effective. It therefore seems possible that certain moderators to the training–effectiveness relationship might exist, such as organization-, job-, and individual-level attributes (Bird and Dunbar, 1991; Bhagat and Prien, 1996). Recent studies have shown that the type of multicultural training, or its emphasis, may moderate the training relationship. For example, Waxin and Panaccio (2005) found that training type significantly moderated the training–expatriate adjustment relationship such that experiential training programs were significantly more effective than didactic programs. Similarly, a study comparing three emphases for diversity training (i.e., cultural awareness, attribution, and cultural skill-training) showed that only skill-based training was significantly related to intent to transfer training (Roberson et al., 2009). Further studies are needed to quantitatively explore the effectiveness of cultural training programs.

***Multicultural Team Training Strategies***  As the prevalence of teams has increased along with the saliency of globalization, multicultural teams have naturally become more and more common. Accordingly, multicultural team training strategies have become a necessity. To address this need, traditional team training programs have been adapted so that trainees are prepared for the unique aspects inherent in multicultural team interactions. Five domains typically characterize multicultural team training programs: (1) enhancing specific aspects of performance through improvement of general team characteristics, (2) using a traditional team performance framework, (3) applying specific training tools and feedback, (4) utilizing a multimethod training method, and (5) training in a limited time frame (Salas and Cannon-Bowers, 2001). Following is a brief review of three common multicultural team training strategies that have been shown to foster greater multicultural team performance and effectiveness.

1. *Team Leader Training*. Cultural differences add a unique layer of complexity to teams, and if team leaders are unable to effectively manage this complexity, teams will fail (Moore, 1999; Salas et al. 2004). Multicultural team leader training typically does not consist of specific training; rather, training for leadership and training for cultural awareness occur separately. For example, team coaching, which facilitates better communication (Kozlowski et al., 1996), can help mitigate poor communication in multicultural teams (Thomas, 1999). Martin and Lumsden (1987) provide a good overview of coaching strategies (e.g., praise or reward desired behaviors, foster encouraging environment).

However, few training programs exist to date that specifically provide leaders with the skills to manage multicultural teams (Burke et al., 2005). "Functional Learning Levers—The Team Leader Toolkit" (FuLL TiLT, Burke et al., 2005) is a program that trains good leadership skills *while taking culture into account*. This training

program incorporates experiential learning and critical incident studies to expose trainees to positive and negative models of multicultural leadership behaviors, with the end goal being trainees' understanding and development of culturally appropriate leadership skills. This is an important step to multicultural team performance, as research has shown variation in leadership style effectiveness across cultures (Ochieng and Price, 2009).

2. *Team Building*. While team building is certainly not unique to multicultural training, it has been shown to be particularly useful in this context. The essence of team building is including a group of individuals in a change development/implementation process (Salas et al., 1999). Team building is frequently used to develop interpersonal relationship skills in multicultural settings (Butler and Zander, 2008), foster trust in culturally diverse teams (Ochieng and Price, 2009), and aid the adjustment process for expatriates (Shay and Tracey, 2009). Team building models can be categorized by one (or more) of four approaches: *goal setting, interpersonal relations, problem solving*, and *role clarification* (Lintunen, 2006). Team *goal setting* allows members to experience the breadth of other team members' perspectives and skills (Watson et al., 1993). An *interpersonal relations* approach encourages the development of healthy relationships within the team. The increased trust resulting from this approach is especially salient in multicultural teams where mistrust often hinders performance (Distefano and Maznevski, 2000; Triandis, 2000). *Problem solving* in the multicultural team is designed to identify and address problems within the team (Adler, 1997). Developing strategies for proactive problem solving enables more efficient future problem solving (Daily et al., 1996). *Role clarification* involves helping team members identify the purpose of each member in the team; this allows for more effective distribution of work. This approach is especially necessary in multicultural teams, where language and social differences can lead to communication and work distribution issues (Steiner, 1972; Thomas, 1999; Hofstede, 1980).

3. *Role Playing*. A third team training strategy effective in multicultural contexts is role playing. Role playing involves trainees interacting with other trainees through scripted scenarios aimed at either (1) enculturation, or learning about one's own culture (Roosa et al., 2002), or (2) acculturation, or learning about another culture. *Enculturation* is a goal of role playing shared with cultural awareness training in that it refers to becoming more aware of the biases, prejudices, and norms inherent in one's own culture. *Acculturation* is similar to didactic training in that its focus is on creating awareness of *other* cultures. When utilizing both these approaches, trainees can become more culturally aware and proficient. Role playing is a form of behavior modeling training (BMT); see the previous section on BMT for guidelines on developing effective role playing training programs. One advantage team-based role playing has over individual (or computer-assisted) approaches is that it allows for experiencing multiple perspectives of cultural interactions. Therefore, multicultural team training programs using role playing should incorporate multiple perspective-taking opportunities into the program.

## 5.7  Program Content

After conducting all the appropriate analyses, developing training objectives, selecting the instructional strategy, and designing the training, the actual content of the program needs to be decided upon (Clark, 2000). This entails structuring the delivery of content in a logical flow and addressing all the training needs and objectives. Accordingly, all tasks trainees are provided with should serve a direct purpose toward achieving training goals. This will allow trainees to focus on the training objectives and avoid distractions that might negatively impact the knowledge structures the trainee is forming during training. Finally, delivering the program content in a logical structure will allow the training to be easily implemented in other contexts and will standardize the entire process.

## 5.8  Summary

There exists an abundance of research on the design of training systems. While it may be tempting to design training based on common sense or personal knowledge, it is vital for the training designer to leverage this extensive knowledge base. Training designers must consider not only the content and the instructional strategy but also external factors such as organizational and individual characteristics when designing a training program for maximal effectiveness.

## 6  TRAINING DEVELOPMENT

The next phase identified in the ISD model involves developing and refining the components of the training program. This includes preparing course materials, creating practice activities, and developing a system for testing and measuring trainee performance.

## 6.1  Practice Scenario Development

Consistent with the previous review (Salas et al., 2006b), research continues to emphasize the critical role of practice in successful training programs. Practice enhances learning by refining knowledge structures, or mental models, within meaningful contexts (Murthy et al., 2008). In addition, practice provides an opportunity to assess performance, enabling trainees to obtain feedback and make adjustments in behavior when weaknesses are identified. Furthermore, practice can promote the transfer of training by allowing trainees to gain experience

applying learned competencies in various contexts. In order to reap these benefits, however, practice scenarios should be carefully designed prior to the training period. Doing so allows trainers and researchers to maintain control over the practice period by standardizing the selection, presentation, and timing of the competencies to be practiced. Practice scenarios can vary greatly in their degree of realism, ranging from very high fidelity (e.g., full motion simulators) to very low fidelity (e.g., role-play activities). As both types have proven to be effective (e.g., Vechhi et al., 2005; Seddon, 2008), the level of realism should be related to the content and goals of the training program. Practice scenarios should also incorporate a range of difficulty levels and should allow trainees to respond in different ways rather than requiring clear-cut answers. Additionally, learning and transfer can be facilitated by enabling trainees to practice their skills on multiple occasions and in various contexts (Prince et al., 1993).

## 6.2 Performance Measures

Performance measures also remain a key contributor to the success of training initiatives. Measuring performance provides an opportunity to assess or diagnose trainee competence and provide feedback, both of which are central to the learning process. Arguably, training cannot effectively lead to changes in knowledge and behavior without incorporating measures of performance. Strong performance measures can ultimately feed back on the success or failure of training and highlight any deficiencies to guide ongoing improvements (Tichon, 2007b). Performance measurement can be facilitated by following certain preparatory guidelines. First, steps can be taken to simplify the measurement process for those responsible for making assessments. Opportunities for measurement should be incorporated into carefully designed practice scenarios. Doing so ensures that target competencies are practiced and measured appropriately. Instructors can thus easily identify trigger points at which performance should be observed and recorded.

Second, an overall system for measuring performance and providing feedback should be established prior to training. This can be a complicated process, as multiple factors should be considered when determining the appropriate strategy. For example, objective measures of performance, such as timing or number of errors, can be obtained automatically through the use of high-technology simulations. While this is a convenient way to collect performance measures, it is not ideal for all practice scenarios because it cannot easily capture data related to the processes (e.g., communication, decision making) used to reach performance outcomes. Team performance in particular cannot easily be captured due to the dynamic natures of teams and team processes. The limitations associated with automatic performance measures are especially apparent during periods of high workload (e.g., those experienced by trauma teams). Under such conditions, teams often communicate and coordinate at the implicit level, and as a result, significant processes become impossible for simulation-based systems to detect. In contrast,

human observers can draw inferences by observing and assessing behaviors using preestablished criteria such as checklists or observation forms. Utilizing human observers, however, could introduce errors and bias into your performance measures. To reduce such issues, at least two observers should be used and steps should be taken to establish reliability (i.e., consistency between evaluators' ratings) and validity (accuracy of evaluators' ratings) (e.g., Brannick et al., 2002; Holt et al., 2002). Choosing between objective and subjective performance measures essentially requires a trade-off that should be guided by the content and goals of the training program.

Lastly, training and practice should be designed to incorporate multiple opportunities for performance measurement. Assessments should be taken on various occasions throughout the simulation in order to gain an accurate representation of trainees' performance.

## 6.3 Summary

Developing practice scenarios and performance measurement strategies are critical steps toward the success of any training program. Learning and transfer are facilitated when trainees are given opportunities to apply target competencies and receive feedback that can guide future performance.

## 7 TRAINING IMPLEMENTATION

At this point, the training program should be fully developed and the organization should prepare to implement it. During this phase, a training location with the appropriate resources (e.g., computers for computer-based training) should be selected. Instructors should be trained, training should be pilot tested, and any final adjustments should be made (Clark, 2000). Once this has been completed, the training program will be fully prepared for delivery. During the implementation phase, steps should be taken to foster a learning climate and support transfer and maintenance (Salas and Stagl, 2009). For example, training objectives should be clearly communicated, and trainees should be prompted to set proximal and distal goals.

## 8 TRAINING EVALUATION

Just as important as developing and implementing training is the process of evaluating the effectiveness of the training program. It is critical not only to determine whether or not training was effective but also to evaluate how or why it was effective or ineffective. Without this phase, it is impossible to make improvements to the training program or re-create it for use in different situations. Training evaluation is thus essential to carrying out the overall goals of a training program. Evaluation concerns and methodologies will be discussed below.

### 8.1 Evaluation Design Concerns

Training evaluation is essentially a system for measuring the intended outcomes or goals of the training program.

The evaluation process includes the examination of such things as measurement, design, learning objectives, and acquisition of the target knowledge, skills, and abilities. Training evaluation is vital to the overall training process because it allows organizations and researchers to determine whether or not training led to meaningful changes in performance and other organizational outcomes of value.

Unfortunately, many organizations do not carry out the evaluation phase after implementing their training program due to the high costs and intensive labor often associated with doing so (Salas and Stagl, 2009). Evaluation can be a difficult process because it might require specialized expertise, a team of people to collect and analyze performance data, and it might need to be conducted in the field or on the job. Moreover, organizations sometimes avoid evaluating their training due to the politics involved or the possibility of uncovering bad news. Fortunately, researchers have taken strides to develop innovative, practical systems for facilitating the evaluation process. For example, evaluation can be simplified by basing the method of evaluation on the specific evaluation questions of interest (Sackett and Mullen, 1993) and by assessing performance differences between training-relevant content and training-irrelevant content following the training period (Haccoun and Hamtiaux 1994). Training evaluation research has spanned multiple training domains such as team training (e.g., Salas et al., 1999), stress training (e.g., Friedland and Keinan, 1992), and computer training (e.g., Simon and Werner, 1996), to name a few.

Since the previous review (Salas et al., 2006b), researchers have continued to develop and refine training evaluation techniques. Many have emphasized the value of assessing multiple training outcomes separately in order to gain a complete picture of the training's overall effectiveness. For example, taking a "voice-centered" approach in which trainees' perceptions of the training are analyzed (Fairtlough, 2007) and assessing trainees' level of satisfaction with the training program (Fullard, 2007) continue to garner support in the literature as effective evaluation strategies. Researchers have also argued for the use of six evaluation levels, namely reactions, learning, job behavior, job performance, organizational team performance, and wider, societal effects (Galanour and Priporas, 2009). Generally, research continues to provide evidence suggesting that carefully implemented training programs are in fact effective. What is less clear, however, is how best to evaluate them. High-caliber, comprehensive training evaluations are unfortunately rare (e.g., Ralphs and Stephan, 1986). In the following sections, we will discuss some of the costs and procedures involved in the evaluation process.

### 8.2  Costs of Training Evaluations

Various practical concerns are often major deterrents to the evaluation of training programs. Implementing the evaluation can put a strain on both temporal and monetary resources. Researchers have explored ways to reduce the costs of training evaluation. For example, trainers can assign different numbers of participants to training and control groups (Yang et al., 1996). Having unequal group sizes with a larger overall sample size may allow for the same level of statistical power at a lower cost. Training designers can also substitute a less expensive proxy criterion measure in place of the target criterion when evaluating training effectiveness. This technique increases the sample size needed to achieve a given statistical power while expending fewer resources. Training designers may need to negotiate a trade-off between reducing costs through proxy criterion measures and potentially increasing costs by utilizing a larger sample size (Arvey et al., 1992).

### 8.3  Kirkpatrick's Typology and Beyond

The most widely used evaluation methods are those based on Kirkpatrick's (1976) four-level model (Anguinis and Kraiger, 2009). In this approach, training is evaluated based on four criteria: (1) trainee reactions (i.e., what trainees think of the training), (2) learning (i.e., what trainees learned), (3) behavior (i.e., how trainees' behavior changes), and (4) organizational results (i.e., impact on organization). Reactions can be assessed by asking trainees if they liked the training and if they perceived it as useful. Tests and exercises can be used to measure the degree to which trainees have acquired the trained competencies. Behavior changes can be measured by observing trainees' performance in the workplace. Finally, organizational results can be evaluated by examining such things as turnover, costs, efficiency, and quality. Each of these levels is further described in Table 4.

Since its inception, Kirkpatrick's (1976) multilevel approach has been widely utilized and supported. Both individual studies (e.g., Noe and Schmitt, 1986; Wexley, 1986) and meta-analytic reviews (Burke and Day, 1986) have garnered support for its use as an effective method for evaluating training. Despite its widespread use, support for the overall framework is limited, as most studies have evaluated only one of the four levels. Specifically, trainee reactions (e.g., Bunker and Cohen, 1977) and trainee learning (e.g., Alliger and Horowitz, 1989) are most commonly evaluated, while the other levels are largely ignored. Salas and colleagues (2001), for example, examined the literature on CRM training and found that 41% of the studies they reviewed used some levels of the Kirkpatrick model to evaluate their training, but the vast majority of them limited their evaluation to only one or two of the levels, typically reactions and learning or reactions and behavior. The results of this review were thus ambiguous and incomplete, highlighting the importance of conducting comprehensive training evaluations.

The recognition of such issues has led to various revisions of the original framework and changes to the way it is used. In regards to measurement, it is important to note that Kirkpatrick's model was designed for a relatively inexperienced audience. As such, researchers have been quick to point to conceptual flaws and other limitations of the model (Snyder et al., 1980; Clement, 1982; Alliger and Janak, 1989). For example, the framework fails to incorporate relevant trainee outcomes such as motivation and self-efficacy

**Table 4  Kirkpatrick's (1976) Multilevel Training Evaluation Typology**

| Level | What Is Being Measured/Evaluated | Measurement | Sample Questions |
|---|---|---|---|
| 1. Reactions | • Learner and/or instructor reactions after training<br>• Satisfaction with training<br>• Ratings of course materials<br>• Effectiveness of content delivery | • Self-report survey<br>• Evaluation or critique | • Did you like the training?<br>• Did you think the trainer was helpful?<br>• How helpful were the training objectives? |
| 2. Learning | • Attainment of trained competencies (i.e., knowledge, skills, and attitudes)<br>• Mastery of learning objectives | • Final examination<br>• Performance exercise<br>• Knowledge pre- and posttests | • True or false: Large training departments are essential for effective training.<br>• Supervisors are closer to employees than is upper management. |
| 3. Behavior | • Application of learned competencies on the job<br>• Transfer of training<br>• Improvement in individual and/or team performance | • Observation of job performance | • Do the trainees perform learned behaviors?<br>• Are the trainees paying attention and being observant?<br>• Have the trainees shown patience? |
| 4. Results | • Operational outcomes<br>• Return on training investment<br>• Benefits to organization | • Longitudinal data<br>• Cost–benefits analysis<br>• Organizational outcomes | • Have there been observable changes in employee turnover, employee attitudes, and safety since the training? |

*Source*: Adapted from Childs and Bell (2002) and Wilson et al. (2005).

(Gist et al., 1988; Gist, 1989; Tannenbaum et al., 1991). Further, measures relating to the value of training, such as content validity (Ford and Wroten, 1984), cost-effectiveness, and utility (Schmidt and Bjork, 1982; Cascio, 1989), are not considered. Finally, because learning is conceptualized as increases in declarative knowledge, Kirkpatrick's typology is difficult to merge with more recent developments in learning theory, such as cognitive skill acquisition (Anderson, 1982; Ackerman, 1987).

In response to these concerns, efforts have been made to improve individual levels of Kirkpatrick's framework and to rebuild or add to the model as a whole. Several researchers, for example, have criticized the first level of the model, reactions, for its use of self-report measures of trainee reactions. Self-report measures have been very popular and arguably overused due to their ease of use, but they do not necessarily provide an accurate or complete representation of training effectiveness. Specifically, measuring the degree to which trainees liked the training does not provide information regarding their learning or performance and thus does not reflect the efficacy of the training program (Alliger et al., 1997). Kraiger and colleagues (1993) addressed this issue in their discussion of three outcomes they deemed essential to training evaluation: affective (i.e., reactions), cognitive (i.e., learning), and skill-based (i.e., behavior) outcomes. The first level, reactions, includes more traditional measures of trainee affect, or how much they liked the training, but also includes a measure of perceived utility, or the degree to which trainees considered the training useful. Utility judgments indicate

the extent to which trainees believe the training will help them on their jobs and can provide information about their learning and potential application of trained competencies. In fact, perceived utility has consistently demonstrated a positive relationship with the transfer of training (Burke and Hutchins, 2007). The authors did not eliminate affective reactions from their model, however, as such measures can still offer information of value. Evaluators can gather information about organizational factors, such as organizational support, through trainees' reports of their feelings toward the training. Thus both forms of trainee reactions were incorporated in the revised framework, and both are often included in more modern models of training evaluation.

The second level, learning, is focused on determining whether or not trainees successfully learned the competencies targeted in the training program. As mentioned above, learning has traditionally been assessed through declarative knowledge tests but can also be measured through procedural tests immediately following training. Trainees' application of learned competencies on the actual job is measured in the behavioral level of evaluation, in which changes in behavior are examined after the training period, in the work environment. Finally, organizational outcomes such as reduced costs and improved performance are assessed at the highest level of evaluation, results. Evaluation at this level is also concerned with the validity of the training program. Intraorganizational validity (i.e., is performance consistent across multiple groups of employees?) and interorganizational validity (i.e., will the training program in one department be effective in another?), for

Causality model of training
evaluation (Kirkpatrick, 1976)

Alternative causality model
(Alliger and Janak, 1989)
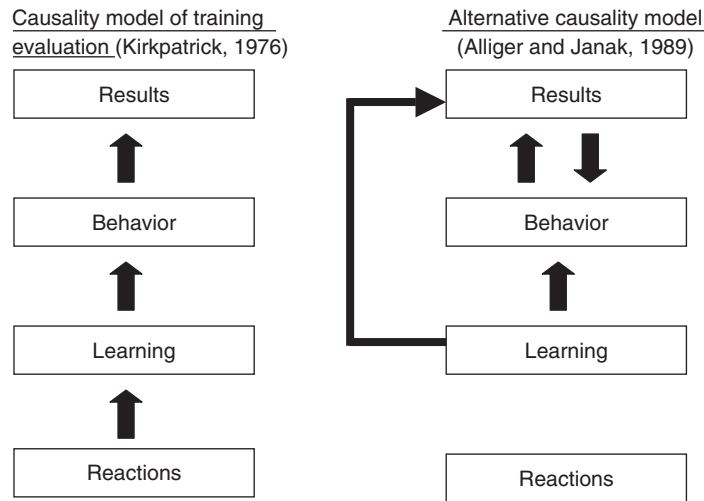
Figure 6    Comparison of two training evaluation models.

example, are important results to examine. Unfortunately, this type of data is rarely collected due to the high costs and labor required to do so.

Alliger and colleagues (1997) also refined and expanded the Kirkpatrick typology based on their meta-analytic review of the literature (Alliger and Janak, 1989). As depicted in Figure 6 and Table 5, their framework extended and clarified the methods involved in the original typology. Specifically, the authors created different classifications for trainee reactions and learning and emphasized the evaluation of the transfer of training rather than the broader behavior category.

Like Kraiger, et al.'s (1993) model, the reaction phase is divided into two categories: affective reactions and utility judgments. Importantly, the learning phase is also comprised of multiple categories, including immediate knowledge, knowledge retention, and behavior/skill demonstration. Including multiple aspects of these levels enables evaluators to gain a more complete understanding of training effectiveness. The authors also argued for the use of multiple evaluation methods such as multiple-choice tests, open-ended questions, and the recall of facts. The model is further strengthened by the behavior/skill component of learning evaluation in which trainees' performance is assessed within the training context. Simulations, performance ratings, and behavioral role-plays can indicate the extent to which trainees have acquired the target competencies.

The last two phases of the revised model include transfer and results. Transfer is evaluated after the training has been completed and trainees have returned to their jobs. Evaluations of transfer are different than those of knowledge retention because they emphasize on-the-job performance, which requires the maintenance and generalization of trained competencies. The results phase is concerned with broad training outcomes such as productivity gains, improved customer satisfaction, reduced production costs, and increased profits. While

each of these evaluation phases can provide valuable information about training effectiveness, it is important to note that evaluation results may or may not be related to training. Such information is difficult to capture and is subject to the influence of several organizational factors aside from training. Evaluation results should thus be interpreted and acted on with caution.

Although revisions to the original Kirkpatrick typology have addressed several of its limitations, other concerns have arisen in the literature. Kraiger and Jung (1997), for example, argued for the use of instructional training objectives in the evaluation of learning outcomes. Other scholars have developed methods for assessing learners' knowledge and in skills in specific domains (Goldsmith and Kraiger, 1997). More recently, researchers have continued to refine the original model. As mentioned previously, Galanou and Priporas (2009) proposed the use of six evaluation levels: reactions, learning, job behavior, job performance, organizational team performance, and wider, societal effects. Other studies have reevaluated specific phases of the model such as trainee reactions. A recent study suggested the adoption of a "voice-centered" approach to training evaluation in which the trainees' in-depth description of the training program is the primary unit of analysis (Fairtlough, 2007).

Despite its limitations and criticisms, the Kirkpatrick (1976) typology remains one of the most widely used frameworks for guiding training evaluation. The original model has served as a valuable foundation for researchers and practitioners alike. Additionally, research shows that much can be gained from supplementing and revising the individual phases of the model and its overall structure.

## 8.4  Summary

It is critical for organizations to evaluate the effectiveness of their training following the completion of a

**Table 5  Alliger et al.'s (1997) Augmented Kirkpatrick Training Taxonomy**

| Step | | Definition |
|---|---|---|
| 1. Reactions | | |
| | a. Affective reactions | Measures emotional self-report of trainees given immediately, with little if any thought; impressions. |
| | b. Utility judgments | Evaluates trainee opinions or judgments about the transferability and utility of the training; behaviorally based opinions. |
| 2. Learning | | |
| | a. Immediate knowledge | Assesses how much trainees learned from training (i.e., how much they know about what they were trained). Uses multiple choice, open-ended questions, lists, etc. |
| | b. Knowledge retention | Assesses what trainees know about training, much like immediate knowledge tests, but administered after some time has passed, to test retention. Used in combination with or instead of immediate knowledge tests. |
| | c. Behavior/skill demonstration | Measures behaviors/skills indicators of performance exhibited during training as opposed to on the job. Uses simulations, behavioral reproduction, ratings of training performance, and performance-centered scorings in classes. |
| 3. Transfer | | Measures output, outcomes, and work samples to assess on-the-job performance. Measured some time after training to assess some measurable aspect of job performance. Assesses transfer of training to job setting. |
| 4. Results | | Assesses what organizational impact training had after the fact. Uses measurement of productivity gains, customer satisfaction, any change in cost, an improvement in employee morale, and profit margin, among others. Measurement is often difficult, due to organizational limitations and because results are the most distal from training. Caution should be used, however, because results are often regarded as the basis for judging training success, but judgments are often based on false expectations. |

*Source:* Originally published in Salas et al. (2006b).

training program. Kirkpatrick's typology has served as a valuable springboard for various evaluation models and tools. Although practical concerns and high costs prevent many organizations from conducting evaluations, doing so can inform future training programs and contribute to organizations' future success.

## 9  TRANSFER OF TRAINING

The transfer of training refers to the extent to which trained competencies are applied, generalized, and maintained in the work environment (Baldwin and Ford, 1988). Transfer leads to meaningful changes in job performance and thus is essentially the primary goal of any training initative. As such, the transfer of training remains a prominent area of interest for both researchers and organizations alike. The previous review highlighted research indicating that transfer is influenced by factors such as organizational context, delay between training and use on the job, and situational cues and consequences. Since then, researchers have provided ample evidence that transfer is also influenced by the three main components of Baldwin and Ford's (1988) model of transfer: trainee characteristics, training design, and the work environment. Several trainee characteristics, for example, have exhibited consistent, positive relationships with the transfer of training. Meta-analytic findings show a strong correlation between cognitive ability and positive transfer outcomes (Blume et al., 2010). Trainees higher in cognitive ability are more likely to successfully acquire, utilize, and maintain trained competencies in the appropriate contexts. Self-efficacy, or one's belief in their ability to accomplish a task (Bandura, 1982), has also been linked to training transfer through meta-analysis (Blume et al., 2010). Research suggests that self-efficacy partially contributes to transfer through its influence on motivation (e.g., Chiaburu and Linsday, 2008), another trainee characteristic that positively predicts the transfer of training (Baldwin et al., 2009). Specifically, pretraining motivation, motivation to learn, and motivation to transfer have all demonstrated significant relationships with the transfer of trained competencies. More recently, perceived utility, or the value associated with participating in training, has also emerged as a predictor of training outcomes (Burke and Hutchins, 2007). Transfer is enhanced when trainees perceive a clear link between trained competencies and valued job outcomes (Chiaburu and Lindsay, 2008) and when training instructions match job requirements (Velada et al., 2007).

Transfer can also be facilitated through the use of certain training design and delivery methods. Since the last review (Salas et al., 2006b), no major research has negated the reported effectiveness of behavioral modeling techniques. Behavioral modeling is a learning approach which incorporates clearly defined explanations of behaviors to be learned, models displaying the effective use of these behaviors, opportunities for trainees to practice learned skills, and the provision of feedback and social reinforcement (Taylor et al., 2005).

## 9.1  Posttraining Environment

In addition to the characteristics of trainees and training design, those of the posttraining environment also play a significant role in the transfer of training. By facilitating or hindering the use of trained competencies, environmental factors largely determine whether or not learned behaviors are exhibited once trainees return to the work setting. If the posttraining environment does not encourage the transfer of target competencies, even well-designed, properly implemented training programs will fail to yield long-term behavioral change. Several components of the posttraining environment have been shown to contribute to transfer outcomes. A positive transfer climate, for example, encourages trainees to apply target knowledge and skills to the job (e.g., Colquitt et al., 2000). Transfer climate has been described as observable or perceived situations that inhibit or facilitate the use of learned skills (Rouiller and Goldstein, 1993). Such situations might include cues that prompt trainees to use new skills, consequences for the correct, incorrect, or lack of use, and social support in the form of feedback and incentives. Research continues to show a significant relationship between transfer climate and training outcomes (e.g., Blume et al., 2010; Gilpin-Jackson and Busche, 2007; Burke et al., 2008).

Supervisor and peer support have also exhibited strong relationships with the transfer of training (e.g., Blume et al., 2010). Supervisor support can be manifested in various ways. Goal setting, for example, can be used to enhance transfer outcomes. Supervisors should prompt employees to set goals for utilizing new competencies in the workplace (Burke and Hutchins, 2007). Supervisors can also provide support by providing recognition and rewards (Salas and Stagl, 2009), communicating expectations (Burke and Hutchins, 2007), and maintaining a high level of involvement (Gilpin-Jackson and Busche, 2007; Saks and Belcourt, 2006). Trainees can support each other by observing one another using trained skills, coaching each other, and sharing ideas about course content (Gilpin-Jackson and Busche, 2007; Hawley and Barnard, 2005).

Not surprisingly, new competencies will not transfer to the workplace unless employees are given ample opportunities to apply them (Burke and Hutchins, 2007). Research shows that deficient time, resources, and opportunities to perform can seriously hinder the use of trained knowledge and skills on the job (e.g., Clarke, 2002; Gilpin-Jackson and Busche, 2007). Transfer is enhanced when trainees are provided with sufficient opportunities and resources for the application of their new skills. Additionally, the delay between training and opportunity to perform should be minimized. Transfer can be further facilitated through the use of follow-up activities (Salas and Stagl, 2009). After, action reviews, for example, debrief trainees, provide further education and enable trainees to reflect on their experiences through practice and discussion. Posttraining interventions such as relapse prevention, self-management, and goal setting can all serve to promote the transfer of training (Baldwin et al., 2009).

Finally, trained competencies are transferred more readily when trainees perceive a continuous learning culture. A continuous learning culture encourages the acquisiton of new knowledge, skills, and attitudes by reinforcing achievement and encouraging innovation and competetion (Tracey et al., 1995). When this climate is ingrained in an organization, learning will be part of the daily work environment, and employees will be more likely to utilize new competencies on the job.

## 9.2  Job Aids

The transfer of training can also be facilitated through the use of job aids. Job aids are tools that are used to assist in the performance of a job or task (Swezey, 1987). Job aids are beneficial because they reduce the amount of time employees need to spend in training, requiring them to spend less time away from their jobs. They can also improve performance by minimizing the cognitive load required to memorize job information, thus freeing up cognitive resources that can be directed toward other aspects of performance. Job aids can be particularly useful in stressful environments in which critical components of a task are more likely to be forgotten or unintentionally omitted. Several types of job aids exist, including informational, procedural, and decision making and coaching. These are listed in Table 6 and further discussed below.

### 9.2.1  Informational Aids

Informational job aids contain material similar to that of on-the-job manuals and reference books. These materials are critical when job information is impossible to memorize (e.g., an aircraft maintenance manual). They are also used to reduce the cognitive demands (e.g., recall of memorized information) associated with performing the job. Informational job aids typically include facts relating to names, locations, dates, and times that are relevant to the job (Rossett and Gautier-Downes, 1991) and are available in paper or electronic formats. Such aids enhance performance by making pertinent job information easily accessible.

**Table 6  Types of Job Aids**

| Type | Description | When to Use |
|---|---|---|
| Informational | Provides access to large amounts of information, such as telephone directory or online database. | During task |
| Procedural | Provides step-by-step instructions for completing a task, such as directions for installing a faucet. | During task |
| Decision making and coaching | Provides a heuristic to guide the user through a thought process to choose the best solution. | Before, during, and after task |

*Source:* Originally published in Salas et al. (2006b).

### 9.2.2  Procedural Aids

Procedural job aids provide step-by-step instructions explaining how to complete a task (Swezey, 1987). In addition to illustrating which actions to take in sequential order, they often also include feedback regarding what the results of each step should look like. Checklists, for example, are a type of procedural aid used to assist employees in remembering and completing each component of their task. Though they have traditionally been provided in paper formats, many companies now offer them online or through other electronic mediums.

### 9.2.3  Decision-Making and Coaching Aids

Decision-making and coaching job aids provide heuristics, or references that guide employees to think a certain way in order to determine the best decision or solution to a problem (Rossett and Gautier-Downes, 1991). Unlike procedural aids, they do not provide detailed, sequential information. Instead, they provide cues, such as ideas or questions, which guide the user toward the path that will lead to the optimal solution. The specific steps used to reach the solution are free to vary.

Job aids have traditionally been implemented when employees are unsure about job information or how to complete a task. Decision-making and coaching aids, however, can be used prior to and after the specific time they are needed. As such, these types of job aids can also be considered training aids because they provide learning opportunities that can benefit future task performance.

Job aids can serve to enhance performance in various contexts. Table 7 provides guidelines describing situations in which job aids should be implemented.

### 9.2.4  Development of Job Aids

To develop a job aid, a task analysis must first be conducted in order to identify the knowledge, skills, equipment, and technical data required to perform the task (Swezey, 1987). The specific steps used to perform the task as well as the appropriate sequence of those steps should also be determined. Once this information

**Table 7  Situations When Job Aids Should Be Used**

- The performance of a task is infrequent and the information is not expected to be remembered.
- The task is complex or has several steps.
- The costs of errors are high.
- Task performance is dependent on knowing a large amount of information.
- Performance depends on dynamic information or procedures.
- Performance can be improved through self-correction.
- The task is simple and there is high turnover volume.
- There is not enough time for training or training resources are not available.

*Source*: Adapted from Rossett and Gautier-Downes (1991).

has been gathered through task analysis, the type of job aid can be determined and the tool can be fully developed. Upon completion, the job aid should be tested and modified to ensure its effectiveness. Further, job aids should be updated as information, procedures, or decision-making processes change (Rossett and Gautier-Downes, 1991).

### 9.2.5  Training Aids

Job aids can also be developed for use during training. Training aids differ from job aids in that they are not used to complete a specific task on the job, but rather, they aid in skill and knowledge acquisition during training. Specifically, training aids are documents, manuals, or devices that are designed to assist trainees in learning the appropriate competencies that are associated with a task or job (Swezey, 1987). Like other types of job aids, training aids are increasingly available in computer-based formats as well as more traditional paper formats.

### 9.2.6  Examples of Job Aids

Various types of job aids are available and implemented in organizations. We discuss two commonly used job aids, namely, manuals and decision support systems.

***Manuals***  Manuals can be used to present both informational and procedural job aids. Information that is especially long or complex can be provided in manuals that employees can utilize as valuable reference tools. Manuals can also be used to house information that is utilized on a daily basis. For example, many organizations provide a directory listing the contact information of employees and other key personnel. Employees can then easily access contact information without having to memorize it or spend time tracking it down. Manuals can also include procedures for tasks that are not performed regularly and thus might not be memorized by employees. Electronic documents and databases are increasingly being used in place of paper manuals, as they are often more convenient to access and easy to update.

Organizations have also used manuals to aid the training process. On their first day of training, employees have traditionally been provided with a training manual that incorporates all the information that is relevant to their company and performing their job. Manuals have typically served as a supplement to classroom training that enabled trainees to learn the nuances and finer details of the job independently. The use of manuals as training aids has become less common, however, due to technological advances and the development of other learning strategies such as simulation and e-learning.

***Decision Support Systems***  As computers become increasingly integral to most work environments, classroom-based training programs continue to be replaced by computer-based platforms. Not surprisingly, computer-based job and training aids, such as decision support systems (DSS) and intelligent tutoring systems (ITSs), have also been developed to complement

such training programs and are rapidly replacing more traditional, paper-based aids.

Designed to improve and support human decision making (Brody et al., 2003), DSSs can be utilized both on the job and during training. When used as job aids, DSSs facilitate and improve decision making during actual task performance. The use of DSSs developed by the Navy, for example, led to increased confidence in decisions made and more effective performance during a decision-making task (Zachary et al., 1998). More recently, DSSs have been put forth as valuable tools for aiding in medical diagnoses (Lindgaard, et al., 2009) and response effectiveness in emergency situations (Yoon et al., 2008).

As training aids, DSSs are integrated into the overall training program, typically alongside simulations, and aid in the development of critical thinking and decision-making skills. DSSs are a valuable addition to scenario-based training exercises. During each simulation, the trainee can utilize the DSS during the decision-making process and receive feedback for each decision that is made. This strategy has been successfully implemented in the military in which DSSs provide real-time strategy training and feedback in a secure training environment (Bell, 2003). DSSs have also garnered support as effective aids in the training of emergency responses in the transportation industry (Yoon et al., 2008) and have even been proposed as a means of facilitating risky business decision-making training (Borrajo et al., 2010).

The ITS, a specific type of DSS, can also be used to facilitate training (Ong and Ramachandran, 2003). As training aids, ITSs can teach a variety of knowledge domains but are difficult to implement because they require extensive knowledge of the subject and strategies for error management. Examples and connections to relevant topics are also required to effectively aid the decision-making process. While ITSs require a great deal of time and effort to design and program, their potential benefits may outweigh the costs for some organizations. Training can be conducted without the physical presence of a facilitator and trainees can learn at their own pace, which can enhance the transfer of training. When used as job aids, the development of ITSs does not require extensive resources, as users should only need small amounts of specific information to supplement their existing knowledge of how to perform a task or reach a decision.

Job and training aids continue to be improved through advances in technology and the science of training. As the associated benefits of such aids increase and the costs required to implement them decrease, we will likely see more and more organizations incorporate them into their training programs. Utilizing job and training aids assists trainees in acquiring knowledge and skills and guides them when applying them on the job. Such tools greatly increase the likelihood that trained competencies will be transferred to the work environment.

### 9.3 Summary

The completion of training evaluation does not mark the end of the training process. Rather, organizations should work to establish a positive transfer climate and provide critical resources such as support systems, opportunities to perform, and job aids. If efforts are not made to facilitate the transfer of training, learned competencies will not be applied on the job, and the goals of the training program will ultimately not be met.

## 10  CONCLUSIONS

### 10.1  Future Directions

The science of training has expanded greatly over the past few decades; this is due largely in part to increasing levels of scientific rigor (Chen and Klimoski, 2007). Furthermore, the breakneck pace of technological advancements has both enabled and necessitated constant growth in the science of training. Two recent training fields likely to see much growth in the coming years are emotions and cognitive neuroscience.

***Emotion Regulation***   Recent research suggests that trainees' ability to regulate their own emotions may play a role in the training process. Emotion regulation generally refers to deliberate, effortful processes that serve to override individuals' spontaneous emotional responses (Koole, 2009). During such processes, people maintain, increase, or decrease positive and negative emotions, changing the way they experience and express their emotions (Gross, 1999). Emotion regulation has proven to have important implications for the workplace, particularly in relation to performance. The term emotional labor, for example, refers specifically to the management of individuals' emotions in the context of their jobs (Schaubroeck and Jones 2000). Engaging in emotional labor can have a negative impact on job performance (e.g., Duke et al., 2009). Recent findings suggest emotion regulation can also have implications for performance within training contexts. Specifically, several scholars have proposed that emotions are particularly relevant when training involves active learning or when trainees actively participate in the learning process, rather than passively absorbing information (Bell and Kozlowski, 2008). Being an active learner can be difficult and anxiety provoking; thus, it is important for trainees to be able to control their emotional reactions during training. Negative emotions can consume attentional resources, hindering learning and performance (Kanfer and Ackerman, 1989). This is particularly true during the early stages of training, when cognitive demands are at their highest.

Fortunately, strategies have been developed to help trainees regulate their emotions during training and alleviate the negative effects of anxiety on learning (Bell and Kozlowski, 2008). For example, instructing trainees to increase the frequency of positive thoughts and reduce the frequency of negative thoughts, along with positive reinforcement, can effectively reduce negative affect and improve performance during training (Kanfer and Ackerman, 1990). Further, work by Bell and Kozlowki (2008) suggests that emotion regulation interventions can reduce anxiety and help sustain trainees' motivation and performance in situations where trainees

**Table 8  Steps in Designing, Delivering, and Evaluating Training Systems**

| Step | Outcome |
|---|---|
| | ***Training Analysis*** |
| 1.  Organizational analysis | Identifies:<br>• Where training is needed<br>• When training is needed<br>• Resources and constraints<br>• Support for transfer |
| 2.  Job/task analysis | Identifies:<br>• Task specifications (e.g., what tasks, under what conditions)<br>• Task characteristics (e.g., equipment needed for task)<br>• Competencies (KSAOs) needed to perform task |
| 2a.  Cognitive task analysis | Identifies:<br>• Cognitive processes and requirements for the task |
| 3.  Person analysis | Identifies:<br>• Who needs training<br>• What they need to be trained on |
| | ***Training Design*** |
| 4.  Develop training objectives. | • Desired outcomes/goals are identified.<br>• Assumptions about training are identified.<br>• Objectives are documented.<br>• Competencies are established. |
| 5.  Consider individual characteristics. | Identifies trainee characteristics that may affect training:<br>• Cognitive ability<br>• Self-efficacy<br>• Goal orientation<br>• Motivation |
| 6.  Consider organizational characteristics. | Identifies organizational characteristics that may affect training:<br>• Organizational culture<br>• Policies and procedures<br>• Situational influences<br>• Prepractice conditions |
| 7.  Establish practice opportunities. | • Practice opportunities are specified (e.g., when they will occur during training, number of opportunities provided, levels of difficulty). |
| 8.  Establish feedback opportunities. | • Identifies when feedback will be provided (e.g., immediately after training) and at what level (e.g., individual, team, both) it is specified.<br>• Trainees know how they did.<br>• Trainees know where improvements are necessary. |
| 9.  Select an instructional strategy. | • The best instructional strategy or combination of strategies will be selected to train competencies of interest based on the needs of the organization (e.g., teams, technology, internationalization). |
| 10.  Outline the program content. | • Sequence and structure of the training program is laid out. |
| | ***Training Development*** |
| 11.  Develop practices scenarios. | • Realistic practice scenarios are scripted that engage trainees.<br>• Scenarios of varying difficulty are scripted. |
| 12.  Develop performance measures. | • The measurement plan is identified.<br>• Criteria for success are developed.<br>• Performance measures are established.<br>• Tools for assisting performance measurement are developed (e.g., observation checklists). |

**Table 8** (*continued*)

| Step | Outcome |
|------|---------|
| *Training Implementation* | |
| 13. Select the instructional setting/location. | • Available training site is identified.<br>• Training environment is prepared. |
| 14. Train instructors. | • Instructors are adequately trained to conduct the instruction.<br>• Instructors are knowledgeable in terms of the program content to handle questions and/or problems that may arise. |
| 15. Conduct a pilot test. | • Issues or concerns with training are identified.<br>• Feedback is received from trainees.<br>• Necessary adjustments are made to the training program. |
| 16. Conduct the instruction. | • Developed instructional materials are put into practice.<br>• Training program is live and functional.<br>• Training program is completed. |
| *Training Evaluation* | |
| 17. Consider evaluation design issues. | • Experimental plan is laid out (e.g., posttest only, control group vs. no control group).<br>• Where evaluations will be conducted is specified (i.e., in the field, on the job, both). |
| 18. Consider costs of training evaluations. | • Low-cost alternatives are explored (e.g., unequal sample sizes between trained and untrained groups; low-cost proxy criterion measure selected). |
| 19. Evaluate training system at multiple levels. | • Data on training's effectiveness are collected at multiple levels and analyzed.<br>• Data on job performance are collected and analyzed. |
| *Transfer of Training* | |
| 20. Establish a positive posttraining environment. | • Organization and supervisors support competencies on the job.<br>• Continuous learning climate is established.<br>• Trainees are rewarded.<br>• Behaviors that contradict those that are trained are discouraged. |
| 21. Use job aids. | • Performance on the job is enhanced. |

*Source:* Originally published in Salas et al. (2006b).

experience stress. These studies suggest that emotion regulation may be an important factor to consider both as a trainee characteristic and as an element of training design, particularly in very difficult or stressful training situations.

***Cognitive Neuroscience*** One arena of study that looks to be promising in relation to the science of training is neuroscience (Salas and Kozlowski, 2010). Neuroscience explores the circuitry of the brain; research in this area may help training designers gain a better understanding of how the brain (and by extension, learning) operates. Neuroergonomics is a subfield of neuroscience; it differs from neuroscience in that traditional neuroscience is interested purely in the functioning of the brain. To this end, cognitive neuroscientists utilize painstaking, and occasionally invasive, methods (e.g., functional magnetic resonance imaging and computerized tomography scan) in order to observe how the brain works. However, research has shown that brain activity differs by context (i.e., labs, simulations, or real world) (Parasuraman and Wilson, 2008). This presents a problem, however, if research in neuroscience is to be applied to the field; accordingly, neuroergonomics has emerged in recent years, examining the brain at work in field settings. Studying brain functioning in relation to human interaction with various

systems, be they technologies or training regimens can help product and training designers develop systems that are maximally suited to human performance (Parasuraman, 2009). At an even deeper level of analysis, molecular genetics allows researchers to link certain genes to specific cognitive functions (e.g., selective attention, working memory, vigilance). Furthermore, genetic makeup can be tracked to more complex brain activity, such as decision making (Parasuraman, 2009). Implications for training may be tailoring training methods to certain genotypes, among other things.

## 10.2 Summary

A strong, capable workforce is critical to the development and success of most organizations. In order to develop their employees' competencies and maintain a competitive advantage, organizations should place a heavy emphasis on training. The purpose of this chapter was to provide an updated training review and offer further guidance related to the design, delivery, and evaluation of training programs (see Table 8). We maintain our position encouraging training designers to take a systematic approach to the training process by carefully considering each component involved. Organizations have much to gain from utilizing the science of training that has been developed and refined by scientists and professionals spanning multiple fields.

# REFERENCES

Ackerman, P. (1987), "Individual Differences in Skill Learning: An Integration of Psychometric and Information Processing Perspectives," *Psychological Bulletin*, Vol. 102, No. 1, pp. 3–27.

Adler, N. J. (1997), *International Dimensions of Organizational Behavior*, 3rd ed., International Thompson Publishing, Cincinnati, OH.

Adler, N. J., and Gundersen, A. (2008), *International Dimensions of Organizational Behavior*, 5th ed., Thomson Higher Education, Mason, OH.

Aguinis, H., and Kraiger, K. (2009), "Benefits of Training and Development for Individuals and Teams, Organizations, and Society," *Annual Review of Psychology*, Vol. 60, pp. 451–474.

Aiba, K., Kadoo, A., and Nomiyama, T. (2002), "Pilot Performance of the Night Approach and Landing under the Limited Runway Light Conditions," *Reports of Aeromedical Laboratory*, Vol. 42, Nos. 2–3, pp. 19–29.

Alliger, G. M., and Horowitz, H. M. (1989), "IBM Takes the Guessing Out of Testing," *Training and Development Journal*, April, pp. 69–73.

Alliger, G. M., and Janak, E. A. (1989), "Kirkpatrick's Levels of Training Criteria: Thirty Years Later," *Personnel Psychology*, Vol. 42, pp. 331–342.

Alliger, G. M., Tannenbaum, S. I., and Bennett, W., Jr. (1997), "A Meta-Analysis of the Relations among Training Criteria," *Personnel Psychology*, Vol. 50, pp. 341–358.

Anderson, J. R. (1982), "Acquisition of Cognitive Skill," *Psychological Review*, Vol. 89, No. 4, pp. 369–406.

Anderson, J. R., Corbett A. T., Koedinger K. R., and Pelletier, R. (1995), "Cognitive Tutors: Lessons Learned," *Journal of Learning Science*, Vol. 4, pp. 167–207.

Arthur, W., Young, B., Jordan, J. A., and Shebilske, W. L. (1996), "Effectiveness of Individual and Dyadic Training Protocols: The Influence of Trainee Interaction Anxiety," *Human Factors*, Vol. 38, pp. 79–86.

Arthur, W., Day, E. A., Bennett, W., McNelly, T. L., and Jordan, J. A. (1997), "Dyadic versus Individual Training Protocols: Loss and Reacquisition of a Complex Skill," *Journal of Applied Psychology*, Vol. 82, pp. 783–791.

Arthur, W. A., Bennet, W., Edens, P. S., and Bell, S. T. (2003), "Effectiveness of Training in Organizations: A Meta-Analysis of Design and Evaluation Features," *Journal of Applied Psychology*, Vol. 88, No. 2, pp. 234–245.

Arvey, R. D., Salas, E., and Gialluca, K. A. (1992), "Using Task Inventories to Forecast Skills and Abilities," *Human Performance*, Vol. 5, pp. 171–190.

Baker, D., Prince, C., Shrestha, L., Oser, R., and Salas, E. (1993), "Aviation Computer Games for Crew Resource Management Training," *International Journal of Aviation Psychology*, Vol. 3, pp. 143–156.

Baldwin, T. T., and Ford, J. K. (1988), "Transfer of Training: A Review and Directions for Future Research," *Personnel Psychology*, Vol. 41, pp. 63–105.

Baldwin, T. T., and Magjuka, R. J. (1997), "Training as an Organizational Episode: Pretraining Influences on Trainee Motivation," in *Improving Training Effectiveness in Work Organizations*, J. K. Ford, S. Kozlowski, K. Kraiger, E. Salas, and M. Teachout, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 99–127.

Baldwin, T.T., Ford, K. J., and Blume, B. D. (2009), "Transfer of Training 1988–2008: An Updated Review and Agenda for Future Research," *International Review of Industrial and Organizational Psychology*, Vol. 24, pp. 41–70.

Bandura, A. (1982), "Self-Efficacy Mechanism in Human Agency," *American Psychologist*, Vol. 37, No. 2, pp. 122–147.

Banker, R. D., Field, J. M., Schroeder, R. G., and Sinha, K. K. (1996), "Impact of Work Teams on Manufacturing Performance: A Longitudinal Study," *Academy of Management Journal*, Vol. 39, No. 4, pp. 867–890.

Beer, M. (1980), *Organization Change and Development: A Systems View*, Scott, Foresman, Glenview, IL.

Befus, C. P. (1988), "A Multilevel Treatment Approach for Culture Shock Experience by Sojourners," *International Journal of Intercultural Relations*, Vol. 12, pp. 381–400.

Bell, B. (2003), "On-Demand Team Training with Simulated Teammates: Some Preliminary Thoughts on Cognitive Model Reuse," *TTG News*, pp. 6–8.

Bell, B. S., and Kozlowski, S. W. J. (2002), "A Typology of Virtual Teams: Implications for Effective Leadership," *Group and Organization Management*, Vol. 27, No. 1, pp. 14–49.

Bell, B. S., and Kozlowski, S. W. J. (2008), "Active Learning: Effects of Core Training Design Elements on Self-Regulatory Processes, Learning, and Adaptability," *Journal of Applied Psychology*, Vol. 93, No. 2, pp. 296–316.

Bell, B. S., Kanar, A. M., and Kozlowski, S. W. J. (2008), "Current Issues and Future Directions in Simulation-Based Training in North America," *International Journal of Human Resource Management*, Vol. 19, No. 8, pp. 1416–1434.

Bell, H., and Waag, W. (1998), "Evaluating the Effectiveness of Flight Simulators for Training Combat Skills: A Review," *International Journal of Aviation Psychology*, Vol. 8, pp. 223–242.

Bell, H. H. (1999), "The Effectiveness of Distributed Mission Training," *Communications of the ACM*, Vol. 42, No. 9, pp. 72–78.

Bennett, J. M. (1986), "Modes of Cross-Cultural Training: Conceptualizing Cross-Cultural Training as Education," *International Journal of Intercultural Relations*, Vol. 10, pp. 117–134.

Bhagat, R., and Prien, K. O. (1996), "Cross-Cultural Training in Organizational Contexts," in *Handbook of Intercultural Training*, 2nd ed., D. Landis and R. S. Bhagat, Eds., Sage Publications, Thousand Oaks, CA, pp. 216–230.

Bhawuk, D. P. S. (1998), "The Role of Culture Theory in Cross-Cultural Training: A Multimethod Study of Culture-Specific, Culture-General, and Culture Theory-Based Assimilators," *Journal of Cross-Cultural Psychology*, Vol. 29, No. 5, pp. 630–655.

Bhawuk, D. P. S. (2001), "Evolution of Culture Assimilators: Toward Theory-Based Assimilators," *International Journal of Intercultural Relations*, Vol. 25, pp. 141–163.

Bhawuk, D. P. S., and Brislin, R. W. (2000), "Cross-Cultural Training: A Review," *Applied Psychology: An International Review*, Vol. 49, pp. 162–191.

Bird, A., and Dunbar, R. (1991), "Getting the Job Done over There: Improving Expatriate Productivity," *National Productivity Review*, Vol. 10, pp. 145–156.

Black, J. S., and Mendenhall, M. E. (1990), "Cross-Cultural Training Effectiveness: A Review and a Theoretical Framework for Future Research," *Academy of Management Review*, Vol. 15, pp. 113–136.

Blickensderfer, E. L., Cannon-Bowers, J. A., and Salas, E. (1997), "Theoretical Bases for Team Self-Correction: Fostering Shared Mental Models," in *Advances in Interdisciplinary Studies in Work Teams Series*, Vol. 4, M. Beyerlein, D. Johnson, and S. Beyerlein, Eds., JAI Press, Greenwich, CT, pp. 249–279.

Blume, B. D., Ford, J. K., Baldwin, T. T., and Huang, J. L. (2010), "Transfer of Training: A Meta-Analytic Review," *Journal of Management*, Vol. 36, No. 4, pp. 1065–1105.

Borrajo, F., Bueno, Y., de Pablo, I., Santos, B., Fernandez, F., Garcia, J., and Segredo, I. (2010), "SIMBA: A Simulator for Business Education and Research," *Decision Support Systems*, Vol. 48, No. 3, pp. 498–506.

Bowers, C. A., Blickensderfer, E. L., and Morgan, B. B. (1998), "Air Traffic Control Specialist Team Coordination," in *Human Factors in Air Traffic Control*, M. W. Smolensky and E. S. Stein, Eds., Academic, San Diego, CA, pp. 215–236.

Brannick, M. T., Prince, C., and Salas, E. (2002), "The Reliability of Instructor Evaluations of Crew Performance: Good News and Not So Good News," *International Journal of Aviation Psychology*, Vol. 12, pp. 241–261.

Branson, R. K., Rayner, G. T., Cox, J. L., Furman, J. P., King, F. J., and Hannum, W. H. (1975), *Interservice Procedures for Instructional Systems Development*, 5 vols., TRADOC Pam 350–30 NAVEDTRA 106A, U.S. Army Training and Doctrine Command, Ft. Monroe, VA.

Bretz, R. D., Jr., and Thompsett, R. E. (1992), "Comparing Traditional and Integrative Learning Methods in Organizational Training Programs," *Journal of Applied Psychology*, Vol. 77, pp. 941–951.

Brewster, C. (1995), "Effective Expatriate Training," in *Expatriate Management: New Ideas for International Business*, J. Selmer, Ed., Quorem Books, Westport, CT, pp. 57–71.

Brody, R. G., Kowalczyk, T. K., and Coulter, J. M. (2003), "The Effect of a Computerized Decision Aid on the Development of Knowledge," *Journal of Business and Psychology*, Vol. 18, No. 2, pp. 157–174.

Brown, K. G., and Ford, J. K. (2002), "Using Computer Technology in Training: Building an Infrastructure for Active Learning," in *Creating, Implementing, and Managing Effective Training and Development*, K. Kraiger, Ed., Jossey-Bass, San Francisco, pp. 192–233.

Brunstein, A., Betts, S., and Anderson, J. R. (2009), "Practice Enables Successful Learning under Minimal Guidance," *Journal of Educational Psychology*, Vol. 101, No. 4, pp. 790–802.

Buckley, R., and Caple, J. (2008), "Training Objectives," in *The Theory & Practice of Training*, 5th rev. ed., Kogan Page, London, pp. 116–131.

Buller, P. F. (1986), "The Team Building–Task Performance Relation: Some Conceptual and Methodological Refinements," *Group and Organizational Studies*, Vol. 11, pp. 147–168.

Bunch, K. J. (2007), "Training Failure as a Consequence of Organizational Culture," *Human Resource Development Review*, Vol. 6, No. 2, pp. 142–163.

Bunker, K. A., and Cohen, S. L. (1977), "The Rigors of Training Evaluation: A Discussion and Field Demonstration," *Personnel Psychology*, Vol. 30, No. 4, pp. 525–541.

Burke, C. S., Hess, K. P., Priest, H. A., Rosen, M., Salas, E., Paley, M., and Riedel, S. (2005), "Facilitating Leadership in a Global Community: A Training Tool for Multicultural Team Leaders," Paper No. 2294, presented at Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL, pp. 1–12.

Burke, L. A., and Hutchins, H.M. (2007), "Training Transfer: An Integrative Literature Review," *Human Resource Development Review*, Vol. 6, No. 3, pp. 263–296.

Burke, L. A., and Hutchins, H. M. (2008), "Study of Best Practices in Training Transfer and Proposed Model of Transfer," *Human Resource Development Quarterly*, Vol. 19, No. 2, pp. 107–128.

Burke, M. J., and Day, R. R. (1986), "A Cumulative Study of the Effectiveness of Managerial Training," *Journal of Applied Psychology*, Vol. 71, No. 2, pp. 232–245.

Burke, M. J., Chan-Serafin, S., Salvador R., Smith, A., and Sarpy, S. A. (2008), "The Role of National Culture and Organizational Climate in Safety Training Effectiveness," *European Journal of Work and Organizational Psychology*, Vol. 17, No. 1, pp. 133–152.

Burke, R. J. (1997), "Organizational Hierarchy and Cultural Values," *Psychological Reports*, Vol. 81, pp. 832–834.

Butler, C., and Zander, L. (2008), "The Business of Teaching and Learning through Multicultural Teams," *Journal of Teaching in International Business*, Vol. 19, No. 2, pp. 192–218.

Buttom, S. B., Mathieu, J. E., and Zajac, D. M. (1996), "Goal Orientation in Organizational Research: A Conceptual and Empirical Foundation," *Organizational Behavior and Human Decision Processes*, Vol. 67, pp. 26–48.

Campbell, J. P. (1971), "Personnel Training and Development," *Annual Review of Psychology*, Vol. 22, pp. 565–602.

Cannon-Bowers, J. A., and Bowers, C. A. (2008), "Synthetic Learning Environments," in *Handbook of Research on Educational Communications and Technology*, 3rd ed., J. M. Spector, M. D. Merrill, J. van Merriënboer, and M. P. Driscoll, Eds., Taylor and Francis Group, New York, pp. 317–327.

Cannon-Bowers, J. A., and Salas, E. (1997), "A Framework for Measuring Team Performance Measures in Training," in *Team Performance Assessment and Measurement: Theory, Methods, and Applications*, M. T. Brannick, E. Salas, and C. Prince, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 45–62.

Cannon-Bowers, J. A., Salas, E., Tannenbaum, S. I., and Mathieu, J. E. (1995), "Toward Theoretically-Based Principles of Trainee Effectiveness: A Model and Initial Empirical Investigation," *Military Psychology*, Vol. 7, pp. 141–164.

Cannon-Bowers, J. A., Burns, J. J., Salas, E., and Pruitt, J. S. (1998), "Advanced Technology in Scenario-Based Training," in *Making Decisions under Stress: Implications for Individual and Team Training*, J. A. Cannon-Bowers and E. Salas, Eds., American Psychological Association, Washington, DC, pp. 365–374.

Carroll, L. A. (1999), "Multimodal Integrated Team Training," *Communications of the ACM*, Vol. 42, No. 9, pp. 68–71.

Cascio, W. F. (1989), "Using Utility Analysis to Assess Training Outcomes," in *Training and Development in Organizations*, I. L. Goldstein, Ed., Jossey-Bass, San Francisco, pp. 63–88.

Chao, G. T., and Moon, H. (2005), "The Cultural Mosaic: A Metatheory for Understanding the Complexity of Culture," *Journal of Applied Psychology*, Vol. 90, No. 6, pp. 1128–1140.

Chen, G., and Klimoski, R. J. (2007), "Training and Development of Human Resources at Work: Is the State of Our Science Strong?" *Human Resource Management Review*, Vol. 17, pp. 180–190.

Chen, I. (2008), "Instructional Design Methodologies," in *Handbook of Research on Instructional Systems and Technology*, T. T. Kidd and H. Song, Eds., Information Science Reference, Hershey, PA, pp. 1–14.

Cheramie, R. A., and Simmering, M. J. (2010), "Improving Individual Learning for Trainees with Low Conscientiousness," *Journal of Managerial Psychology*, Vol. 25, No. 1, pp. 44–57.

Chiaburu, D. S., and Linsday, D. R. (2008), "Can Do or Will Do? The Importance of Self-Efficacy and Instrumentality for Training Transfer," *Human Resource Development International*, Vol. 11, No. 2, pp. 199–206.

Childs, J. M., and Bell, H. H. (2002), "Training Systems Evaluation," in *Handbook of Human Factors Testing and Evaluation*, S. G. Charlton and T. G. O'Brien, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 473–509.

Christoph, R. T., Schoenfeld, G. A., Jr., and Tansky, J. W. (1998), "Overcoming Barriers to Training Utilizing Technology: The Influence of Self-Efficacy Factors on Multimedia-Based Training Receptiveness," *Human Resources Development Quarterly*, Vol. 9, pp. 25–38.

Clark, D. (2000), "Introduction to Instructional System design," available: http://www.nwlink.com/~donclark/hrd/sat1.html#model.

Clarke, N. (2002), "Job/Work Environment Factors Influencing Training Transfer Within a Human Service Agency: Some Indicative Support for Baldwin and Ford's Transfer Climate Construct," *International Journal of Training and Development*, Vol. 6, pp. 146–162.

Clement, R. W. (1982), "Testing the Hierarchy Theory of Training Evaluation: An Expanded Role for Trainee Reactions," *Public Personnel Management*, Vol. 11, No. 2, pp. 176–184.

Colegrove, C. M., and Bennett, W. (2006), "Competency-Based Training: Adapting to Warfighter Needs," Warfighter Readiness Research Division, AFRL-HE-AZ-TR-2006-0014.

Collins, N. M, and Pieterse, A. L. (2007), "Critical Incident Analysis Based Training: An Approach for Developing Active Racial/Cultural Awareness," *Journal of Counseling & Development*, Vol. 85, pp. 14–23.

Colquitt, J. A., LePine, J. A., and Noe, R. A. (2000), "Toward an Integrative Theory of Training Motivation: A Meta-Analytic Path Analysis of 20 Years of Research," *Journal of Applied Psychology*, Vol. 85, No. 5, pp. 678–707.

Cooke, N. J. (1999), "Knowledge Elicitation," in *Handbook of Applied Cognition*, F. T. Durso, R. S. Nickerson, R. W. Schvaneveldt, S. T. Dumais, D. S. Lindsay, and M. T. H. Chi, Eds., Wiley, New York, pp. 479–509.

Daily, B., Whatley, A., Ash, S. R., and Steiner, R. L. (1996), "The Effects of a Group Decision Support System on Culturally Diverse and Culturally Homogeneous Group Decision Making," *Information and Management*, Vol. 30, pp. 281–289.

Degani, A., and Wiener, E. L. (1997), "Philosophy, Policies, Procedures and Practices: The Four 'P's of Flight Deck Operations," in *Aviation Psychology in Practice*, N. Johnston, N. McDonald, and R. Fuller, Eds., Avebury, Aldershot, Hampshire, England, pp. 44–67.

De Jong, J. A., and Versloot, B. (1999), "Structuring On-the-Job Training: Report of a Multiple Case Study," *International Journal of Training and Development*, Vol. 3, No. 3, pp. 186–199.

DeRouin, R. E., Fritzsche, B. A., and Salas E. (2004), "Optimizing E-Learning: Research Based Guidelines for Learner-Controlled Training," *Human Resource Management Journal*, Vol. 43, No. 2/3, pp. 147–162.

Deshpande, S. P., and Viswesvaran, C. (1992), "Is Cross-Cultural Training of Expatriate Managers Effective: A Meta Analysis," *International Journal of Intercultural Relations*, Vol. 16, pp. 295–310.

Dierdorff, E. C., and Surface, E. A. (2008), "Assessing Training Needs: Do Work Experience and Capability Matter?" *Human Performance*, Vol. 21, pp. 28–48.

Distefano, J. J., and Maznevski, M. L. (2000), "Creating Value with Diverse Teams in Global Management," *Organizational Dynamics*, Vol. 29, No. 1, pp. 45–63.

Dormann, T., and Frese, M. (1994), "Error Training: Replication and the Function of Exploratory Behavior," *International Journal of Human–Computer Interaction*, Vol. 6, No. 4, pp. 365–372.

Driskell, J. E., Johnston, J. H., and Salas, E. (2001), "Does Stress Training Generalize to Novel Settings?" *Human Factors*, Vol. 43, No. 1, pp. 99–110.

Driskell, J. E., Radtke, P. H., and Salas, E. (2003), "Virtual Teams: Effects of Technological Mediation on Team Processes," *Group Dynamics*, Vol. 7, No. 4, pp. 297–323.

Driskell, J. E., and Johnston, J. H. (1998), "Stress Exposure Training," in *Making Decisions Under Stress: Implications for Individual and Team Training*, J. A. Cannon-Bowers and E. Salas, Eds., American Psychological Association, Washington, DC, pp. 191–217.

Driskell, J. E., Salas, E., Johnston, J. H., and Wollert, T. N. (2008), "Stress Exposure Training: An Event-Based Approach," in *Performance under Stress*, P. A. Hancock and J. L. Szalma (Eds.), Ashgate Publishing Limited, Hampshire, England, pp. 271–286.

Duke, A. B., Goodman, J. M., Treadway, D. C., and Breland, J. W. (2009), "Perceived Emotional Support as a Moderator of the Emotional Labor/Outcomes Relationships," *Journal of Applied Social Psychology*, Vol. 39., No. 5, pp. 1013–1034.

Dweck, C. S., and Leggett, E. L. (1988), "A Social-Cognitive Approach to Motivation and Personality," *Psychological Review*, Vol. 95, pp. 256–273.

Dyer, W. G. (1977), *Team Building: Issues and Alternatives*, Addison-Wesley, Reading, MA.

Ehrenstein, A., Walker, B., Czerwinski, M. and Feldman, E. (1997), "Some Fundamentals of Training and Transfer: Practice Benefits Are Not Automatic," in *Training for a Rapidly Changing Workplace: Applications of Psychological Research, American Psychological Association*, Washington, DC, pp. 31–60.

Ellermann, H. H., and Free, E. L. (1990), "A Subject-Controlled Environment for Paired Associate Learning," *Journal of Computer-Based Instruction*, Vol. 17, pp. 97–102.

Elliot, A. J., and Church, M. A. (1997), "A Hierarchical Model of Approach and Avoidance Achievement Motivation," *Journal of Personality and Social Psychology*, Vol. 72, pp. 218–232.

Elliot, A. J., and McGregor, H. A. (2001), "A 2 x 2 Achievement Goal Framework," *Journal of Personality and Social Psychology*, Vol. 80, No. 3, pp. 501–519.

Entin, E. E., and Serfaty, D. (1999), "Adaptive Team Coordination," *Human Factors*, Vol. 41, No. 2, pp. 312–325.

Fairtlough, A. C. (2007), "Adapting the Voice-Centred Relational Method of Data Analysis: Reading the Trainees' Accounts of Their Learning on a Pilot Programme for Practitioners Working with Parents," *Learning in Health and Social Care*, Vol. 6, No. 1, pp. 2–13.

Feldman, D. (1988), *Managing Careers in Organizations*, Scott, Foresman, Glenview, IL.

Festner, D., and Gruber, H. (2008), "Conditions of Work Environments in Fostering Transfer of Training," in *Emerging Perspectives of Workplace Learning*, S. Billett, C. Harteis, and A. Eteläpelto (Eds.), Sense Publishers, Rotterdam, The Netherlands, pp. 215–228.

Fiore, S. M., Salas, E., and Cannon-Bowers, J. A. (2001), "Group Dynamics and Shared Mental Model Development," in *How People Evaluate Others in Organizations*, M. London, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 309–336.

Fisher, D. L., Laurie, N. E., and Glaser, R. (2002), "Use of a Fixed-Based Driving Simulator to Evaluate the Effects of Experience and PC-Based Risk Assessment Training on Drivers' Decision," *Human Factors*, Vol. 44, No. 2, pp. 287–302.

Fletcher, J. D. (2009), "Education and Training Technology in the Military," *Science*, Vol. 323, No. 72, pp. 72–75.

Fong, G. (2006), "Adapting COTS games for military experimentation," *Simulation & Gaming*, Vol. 37, No. 4, pp. 452–465.

Ford, J. K., and Noe, R. A. (1987), "Self-Assessed Training Needs: The Effects of Attitudes toward Training, Managerial Level, and Function," *Personnel Psychology*, Vol. 40, pp. 39–53.

Ford, J. K., and Wroten, S. P. (1984), "Introducing New Methods for Conducting Training Evaluation and for Linking Training Evaluation to Program Redesign," *Personnel Psychology*, Vol. 37, pp. 651–665.

Ford, J. K., Quinones, M. A., Sego, D. J., and Sorra, J. S. (1992), "Factors Affecting the Opportunity to Perform Trained Tasks on the Job," *Personnel Psychology*, Vol. 45, pp. 511–527.

Ford, J. K., Kozlowski, S., Kraiger, K., Salas, E., and Teachout, M., Eds. (1997), *Improving Training Effectiveness in Work Organizations*, Lawrence Erlbaum Associates, Mahwah, NJ.

Ford, J. K., Smith, E. M., Weissbein, D. A., Gully, S. M., and Salas, E. (1998), "Relationships of Goal Orientation, Metacognitive Activity, and Practice Strategies with Learning Outcomes and Transfer," *Journal of Applied Psychology*, Vol. 83, pp. 218–233.

Forret, M. L., Turban, D. B., and Dougherty, T. W. (1996), "Issues Facing Organizations When Implementing Formal Mentoring Programmes," *Leadership and Organization Development Journal*, Vol. 17, No. 3, pp. 27–30.

Fowler, S. M., and Pusch, M. D. (2010), "Intercultural Simulation Games: A Review (of the United States and Beyond)," *Simulation Gaming*, Vol. 41, No. 1, pp. 94–115.

Fowlkes, J., Dwyer, D. J., Oser, R. L., and Salas, E. (1998), "Event-Based Approach to Training (EBAT)," *International Journal of Aviation Psychology*, Vol. 8, No. 3, pp. 209–221.

Frese, M., and Altman, A. (1989), "The Treatment of Errors in Learning and Training," in *Developing Skills with Information Technology*, L. Bainbridge and S. A. Quintanilla, Eds., Wiley, Chichester, West Sussex, England, pp. 65–86.

Frese, M., Brodbeck, F., Heinnbokel, T., Mooser, C., Schleiffenbaum, E., and Thiemann, P. (1991), "Errors in Training Computer Skills: On the Positive Functions of Errors," *Human–Computer Interaction*, Vol. 6, pp. 77–93.

Friedland, N., and Keinan, G. (1992), "Training Effective Performance in Stressful Situations: Three Approaches and Implications for Combat Training," *Military Psychology*, Vol. 4. pp. 157–174.

Fullard, F. (2007), "A Model to Evaluate the Effectiveness of Enterprise Training Programmes," *International Entrepreneurship and Management Journal*, Vol. 3, No. 3, pp. 263–276.

Galanour, E., and Priporas C. V. (2009), "A Model for Evaluating the Effectiveness of Middle Managers' Training Courses: Evidence from a Major Banking Organization in Greece," *International Journal of Training and Development*, Vol. 13, No. 4, pp. 221–246.

Ghiselli, E. E., (1966), *The Validity of Occupational Aptitude Tests*, Wiley, New York.

Ghodsian, D., Bjork, R., and Benjamin, A. (1997), "Evaluating Training during Training: Obstacles and Opportunities," in *Training for a Rapidly Changing Workplace: Applications of Psychological Research*, M. A. Quinones and A. Ehrenstein, Eds., American Psychological Association, Washington, DC, pp. 63–88.

Gillan, C. A. (2003), "Aircrew Adaptive Decision Making: A Cross-Case Analysis," *Dissertation Abstracts International, Section B: The Sciences and Engineering*, Vol. 64, No. 1-B, p. 438.

Gilpin-Jackson, Y., and Busche, G. R. (2007), "Leadership Development Training Transfer: A Case Study of Post-Training Determinants," *Journal of Management Development*, Vol. 26, No. 10, pp. 980–1004.

Gist, M., Rosen, B., and Schwoerer, C. (1988), "The Influence of Training Method and Trainee Age on the Acquisition of Computer Skills," *Personnel Psychology*, Vol. 41, No. 2, pp. 255–265.

Gist, M. E. (1989), "Effects of Alternative Training Methods on Self-Efficacy and Performance in Computer Software Training," *Journal of Applied Psychology*, Vol. 74, No. 6, pp. 884–891.

Glendon, A. I., and Stanton, N. A. (2000), "Perspectives on Safety Culture," *Safety Science*, Vol. 34, pp. 193–214.

Goettl, B. P., Yadrick, R. M., Connolly-Gomez, C., Regian, W. J., and Shebilske, W. L. (1996), "Alternating Task Modules in Isochronal Distributed Training of Complex Tasks," *Human Factors*, Vol. 38, pp. 330–346.

Goforth, D. (1994), "Learner Control = Decision Making + Information: A Model and Meta Analysis," *Journal of Educational Computing Research*, Vol. 11, pp. 1–26.

Goldsmith, T., and Kraiger, K. (1997), "Structural Knowledge Assessment and Training Evaluation," in *Improving Training Effectiveness in Work Organizations*, J. K. Ford, S. Kozlowski, K. Kraiger, E. Salas, and M. Teachout, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 19–46.

Goldstein, I. L. (1993), *Training in Organizations*, 3rd ed., Brooks/Cole, Pacific Grove, CA.

Goldstein, I. L., and Ford, J. K. (2002), *Training in Organizations: Needs Assessment, Development, and Evaluation*, 4th ed., Wadsworth, Belmont, CA.

Gopher, D., Weil, M., and Bareket, T. (1994), "Transfer of Skill from a Computer Game Trainer to Flight," *Human Factors*, Vol. 36, pp. 387–405.

Gordon, S. E., and Gill, R. T. (1997), "Cognitive Task Analysis," in *Naturalistic Decision Making*, C. E. Zsambok and G. Klein, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 131–140.

Gross, J. J. (1999), "Emotion Regulation: Past, Present, Future," *Cognition and Emotion*, Vol. 13, No. 5, pp. 551–573.

Grove, C., and Torbiörn, I. (1993), "A New Conceptualization of Intercultural Adjustment and the Goals of Training," in *Education for the Intercultural Experience*, R. M. Paige, Ed., Intercultural, Yarmouth, ME, pp. 205–233.

Gudykunst, W. B., Guzley, R. M., and Hammer, M. R. (1996), "Designing Intercultural Training," in *Handbook of Intercultural Training*, 2nd ed., D. Landis and R. W. Brislin, Eds., Sage Publications, Thousand Oaks, CA, pp. 61–80.

Guldenmund, F. W. (2000), "The Nature of Safety Culture: A Review of Theory and Research," *Safety Science*, Vol. 34, pp. 215–257.

Gully, S. M., Payne, S. C., Kiechel Koles, K. L., and Whiteman, J. K. (2002), "The Impact of Error Training and Individual Differences on Training Outcomes: An Attribute–Treatment Interaction Perspective," *Journal of Applied Psychology*, Vol. 87, No. 1, pp. 143–155.

Guzzo, R. A., and Dickson, M. W. (1996), "Teams in Organizations: Recent Research on Performance and Effectiveness," *Annual Review of Psychology*, Vol. 47, pp. 307–338.

Haccoun, R. R., and Hamtiaux, T. (1994), "Optimizing Knowledge Tests for Inferring Learning Acquisition Levels in Single Group Training Evaluation Designs: The Internal Referencing Strategy," *Personnel Psychology*, Vol. 47, pp. 593–604.

Harris, D., and Khan, H. (2003), "Response Time to Reject a Takeoff," *Human Factors and Aerospace Safety*, Vol. 3, No. 2, pp 165–175.

Hawley, J. D., and Barbard, J. K. (2005), "Work Environment Characteristics and Implications for Training Transfer: A Case Study of the Nuclear Power Industry," *Human Resource Development International*, Vol. 8, No. 1, pp. 65–80.

Heimbeck, D., Frese, M., Sonnentag, S., and Keith, N. (2003), "Integrating Errors into the Training Process: The Function of Error Management Instructions and the Role of Goal Orientation," *Personnel Psychology*, Vol. 56, pp. 333–361.

Hendricks, C. C. (2001), "Teaching Causal Reasoning through Cognitive Apprenticeship: What Are the Results from Situated Learning?" *Journal of Educational Research*, Vol. 94, No. 5, pp. 302–311.

Hoffman, R. R., and Lintern, G. (2006), "Eliciting and Representing the Knowledge of Experts," in *Cambridge Handbook of Expertise and Expert Performance*, K. A. Ericsson, N. Charness, and P. Feltovich, Eds., Cambridge University Press, New York, pp. 203–222.

Hofmann, D. A., and Stetzer, A. (1996), "A Cross-Level Investigation of Factors Influencing Unsafe Behaviors and Accidents," *Personnel Psychology*, Vol. 49, pp. 307–339.

Hofmann, D. A., Jacobs, R., and Landy, F. (1995), "High Reliability Process Industries: Individual, Micro, and Macro Organizational Influences on Safety Performance," *Journal of Safety Research*, Vol. 26, pp. 131–149.

Hofstede, G. (1980), *Culture's Consequences: International Differences in Work Related Values*, Sage Publications, Beverly Hills, CA.

Holt, R. W., Hansberger, J. T., and Boehm-Davis, D. A. (2002), "Improving Rater Calibration in Aviation: A Case Study," *International Journal of Aviation Psychology*, Vol. 12, pp. 305–330.

Ivancic, K., and Hesketh, B. (1995), "Making the Best of Errors during Training," *Training Research Journal*, Vol. 1, pp. 103–125.

Jacobs, J. W., and Dempsey, J. V. (1993), "Simulation and Gaming: Fidelity, Feedback, and Motivation," in *Interactive Instruction and Feedback*, J. V. Dempsey and G. C. Sales, Eds., Educational Technology, Englewood Cliffs, NJ, pp. 197–229.

Jacobs, R. L., and Park, Y. (2009), "A Proposed Conceptual Framework of Workplace Learning: Implications for Theory Development and Research in Human Resource Development," *Human Resource Development Review*, Vol. 8, pp. 133–150.

Jentsch, F., and Bowers, C. (1998), "Evidence for the Validity of PC-Based Simulations in Studying Aircrew Coordination," *International Journal of Aviation Psychology*, Vol. 8, pp. 243–260.

Johnston, J. H., and Cannon-Bowers, J. A. (1996), "Training for Stress Exposure," in *Stress and Human Performance*, J. E. Driskell and E. Salas, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 223–256.

Kanfer, R., and Ackerman, P. L. (1989), "Motivation and Cognitive Abilities: An Integrative/Aptitude-Treatment Interaction Approach to Skill Acquisition," *Journal of Applied Psychology*, Vol. 74, No. 4, pp. 657–690.

Kanfer, R., and Ackerman, P. L. (1990), "Ability and Metacognitive Determinants of Skill Acquisition and Transfer," Air Force Office of Scientific Research, Minneapolis, MN.

Kaplan-Leiserson, E. (2002), "E-Learning Glossary," available: http://www.learning-circuits.org/glossary.html, accessed July 1, 2002.

Kealey, D. J., and Protheroe, D. R. (1996), "The Effectiveness of Cross-Cultural Training for Expatriates: An Assessment of the Literature on the Issue," *International Journal of Intercultural Relations*, Vol. 20, pp. 141–165.

Keith, N., and Frese, M. (2008), "Effectiveness of Error Management Training: A Meta-Analysis," *Journal of Applied Psychology*, Vol. 93, No. 1, pp. 59–69.

Kirkpatrick, D. L. (1976), "Evaluation of Training," in *Training and Development Handbook: A Guide to Human Resource Development*, 2nd ed., R. L. Craig, Ed., McGraw-Hill, New York, pp. 1–26.

Klein, G., and Militello, L. (2001), "Some Guidelines for Conducting a Cognitive Task Analysis," in *Advances in Human Performance and Cognitive Engineering Research*, Vol. 1, E. Salas, Ed., Elsevier Science, Amsterdam, pp. 161–199.

Klein, H. J., Noe, R. A., and Wang, C. (2006), "Motivation to Learn and Course Outcomes: The Impact of Delivery Mode, Learning Goal Orientation, and Perceived Barriers and Enablers," *Personnel Psychology*, Vol. 59, pp. 665–702.

Knapp, D. E., and Kustis, G. A. (1996), "The Real 'Disclosure': Sexual Harassment and the Bottom Line," in *Sexual Harassment in the Workplace: Perspectives, Frontiers, and Response Strategies*, M. S. Stockdale, Ed., Sage Publications, Thousand Oaks, CA, pp. 199–213.

Koole, S. L. (2009), "The Psychology of Emotion Regulation: An Integrative Review," *Cognition and Emotion*, Vol. 23, No. 1, pp. 4–41.

Kozlowski, S. J., and Bell, B. S. (2002), "A Typology of Virtual Teams: Implications for Effective Leadership," *Group and Organization Management*, Vol. 27, No. 1, pp. 14–49.

Kozlowski, S. W. J., and Bell, B. S. (2006), "Disentangling Achievement Orientation and Goal Setting: Effects on Self-Regulatory Processes," *Journal of Applied Psychology*, Vol. 91, No. 4, pp. 900–916.

Kozlowski, S. W. J., and Bell, B. S. (2007), "A Theory-Based Approach for Designing Distributed Learning Systems," in *Toward a Science of Distributed Learning*, S. M. Fiore and E. Salas, Eds., American Psychological Association, Washington, DC, pp. 15–39.

Kozlowski, S. W. J., Gully, S. M., Salas, E., and Cannon-Bowers, J. A. (1996), "Team Leadership and Development: Theory, Principles, and Guidelines for Training Leaders and Teams," in *Advances in Interdisciplinary Studies of Work Teams: Team Leadership*, M. Beyerlein, S. Beyerlein, and D. Johnson, Eds., JAI Press, Greenwich, CT, pp. 3:253–292.

Kozlowski, S. W. J., and Salas, E. (Eds.). (2010), *Learning, Training, and Development in Organizations*. Taylor & Francis Group, New York, NY.

Kraiger, K., and Jung, K. (1997), "Linking Training Objectives to Evaluation Criteria," in *Training for a Rapidly Changing Workplace: Applications of Psychological Research*, M. A. Quinones and A. Ehrenstein, Eds., American Psychological Association, Washington, DC, pp. 151–176.

Kraiger, K., Ford, J. K., and Salas, E. (1993), "Application of Cognitive, Skill-Based, and Affective Theories of Learning Outcomes to New Methods of Training Evaluation," *Journal of Applied Psychology*, Vol. 78, No. 2, pp. 311–328.

Kulik, C. T., and Roberson, L. (2008), "Common Goals and Golden Opportunities: Evaluations of Diversity Education in Academic and Organizational Settings," *Academy of Management Learning & Education*, Vol. 7, No. 3, pp. 309–331.

Lai, Shu-Ling (2001), "Controlling the Display of Animation for Better Understanding," *Journal of Research on Computing in Education*, Vol. 33, available: http://www.iste.org/jrte/33/5/lai.html, accessed June 13, 2001.

Landis, D., and Brislin, R. W., Eds. (1996), *Handbook of Intercultural Training*, 2nd ed., Sage Publications, Thousand Oaks, CA.

Latham, G. P. (1989), "Behavioral Approaches to the Training and Learning Process," in *Training and Development in Organizations*, L. L. Goldstein, Ed., Jossey-Bass, San Francisco, pp. 256–295.

Lewin, K. (1951), *Field Theory in Social Science: Selected Theoretical Papers*, Harper & Row, New York.

Lewis, M. (1998), "What Workers Need to Succeed," *American Machinist*, Vol. 142, No. 3, pp. 88–98.

Lim, D. H., and Morris, M. L. (2006), "Influence of Trainee Characteristics, Instructional Satisfaction, and Organizational Climate on Perceived Learning and Training Transfer," *Human Resource Development Quarterly*, Vol. 17, No. 1, pp. 85–115.

Lindgaard, G., Pyper, C., Frize, M., and Walker, R. (2009), "Does Bayes Have It? Decision Support Systems in Diagnostic Medicine," *International Journal of Industrial Ergonomics*, Vol. 39, No. 3, pp. 524–532.

Lintunen, T. (2006), "Social and Emotional Learning in the School Physical Education Context," in *Current Research Topics in Exercise and Sport Psychology in Europe*, F. Boen, B. de Cuyper, and J. Opdenacker, Eds., Lannoo Campus Publishers, Naamsesteenweg, Belgium, pp. 25–33.

Littrell, L. N., and Salas, E. (2005), "A Review of Cross-Cultural Training: Best Practices, Guidelines, and Research Needs," *Human Resource Development Review*, Vol. 4, No. 3, pp. 305–334.

Littrell, L. N., Salas, E., Hess, K. P., Paley, M., and Riedel, S. (2006), "Expatriate Preparation: A Critical Analysis of 25 Years of Cross-Cultural Training Research," *Human Resource Development Review*, Vol. 5, No. 3, pp. 355–388.

Lohman, M. C. (2009), "A Survey of Factors Influencing the Engagement of Information Technology Professionals in Informal Learning Activities," *Information Technology, Learning, and Performance Journal*, Vol. 25, No. 1, pp. 43–53.

Long, L. K., DuBois, C. Z., and Faley, R. H. (2008), "Online Training: The Value of Capturing Trainee Reactions," *Journal of Workplace Learning*, Vol. 20, No. 1, pp. 21–37.

Lorenzet, S. J., Salas, E., and Tannenbaum, S. I. (2005), "Benefitting from Mistakes: The Impact of Guided Errors on Skill Development and Self-Efficacy," *Human Resource Development Quarterly*, Vol. 16, No. 3, pp. 301–322.

Major, D. A., Turner, J. E., and Fletcher, T. D. (2006), "Linking Proactive Personality and the Big Five to Motivation to Learn and Development Activity," *Journal of Applied Psychology*, Vol. 91, No. 4, pp. 927–935.

Marks, M. A. (2000), "A Critical Analysis of Computer Simulations for Conducting Team Research," *Small Group Research*, Vol. 31, No. 6, pp. 653–675.

Marks, R. B., Sibley, S. D., and Arbaugh, J. B. (2005), "A Structural Equation Model of Predictors for Effective Online Learning," *Journal of Management Education*, Vol. 29, pp. 531–563.

Martin, G. L., and Lumsden, J. A. (1987), *Coaching: An Effective Behavioral Approach*, C.V. Mosby, St. Louis, MO.

Martocchio, J. J. (1992), "Microcomputer Usage as an Opportunity: The Influence of Context in Employee Training," *Personnel Psychology*, Vol. 45, pp. 529–551.

Martocchio, J. J., and Judge, T. A. (1997), "Relationship between Conscientiousness and Learning in Employee Training: Mediating Influences of Self-Deception and Self-Efficacy," *Journal of Applied Psychology*, Vol. 82, pp. 764–773.

Martocchio, J. J., and Webster, J. (1992), "Effects of Feedback and Cognitive Playfulness on Performance in Microcomputer Software Training," *Personality Psychology*, Vol. 45, pp. 553–578.

Mathieu, J. E., and Tesluk, P. E. (2010), "A Multilevel Perspective on Training and Development Effectiveness," in *Learning, Training, and Development in Organizations*, S. W. J. Kozlowski and E. Salas, Eds., Taylor and Francis Group, New York, pp. 405–440.

Mathieu, J. E., Tannenbaum, S. I., and Salas, E. (1992), "Influences of Individual and Situational Characteristics on Measures of Training Effectiveness," *Academy of Management Review*, Vol. 35, pp. 828–847.

METI. (2010). "Medical Education and Simulation." Retrieved from http://www.meti.com/index.html.

Milheim, W. D., and Martin, B. L. (1991), "Theoretical Bases for the Use of Learner Control: Three Different Perspectives," *Journal of Computer-Based Instruction*, Vol. 18, pp. 99–105.

Miller, J. E., Patterson, E. S., and Woods, D. D. (2006), "Elicitation by critiquing a cognitive task analysis methodology," *Cognition Technology and Work*, Vol. 8, pp. 90–102.

Mobley, G. M., Jaret, C., Marsh, K., and Lim, Y. Y. (1994), "Mentoring, Job Satisfaction, Gender, and the Legal Profession," *Sex Roles: A Journal of Research*, Vol. 31, pp. 79–98.

Moe, M. T., and Blodget, H. (2000), *The Knowledge Web*, Part 3, *Higher Web Universities Online*, Merrill Lynch & Co., New York.

Moore, S. (1999), "Understanding and Managing Diversity among Groups at Work: Key Issues for Organisational Training and Development," *Journal of European Industrial Training*, Vol. 23, Nos. 4–5, pp. 208–217.

Morris, M. A., and Robie, C. (2001), "A Meta-Analysis of the Effects of Cross-Cultural Training on Expatriate Performance and Adjustment," *International Journal of Training and Development*, Vol. 5, pp. 112–125.

Morrison, G. R., Ross, S. M., and Baldwin, W. (1992), "Learner Control of Context and Instructional Support in Learning Elementary School Mathematics," *Educational Technology Research and Development*, Vol. 40, pp. 5–13.

Munro, C. R. (2009), "Mentoring Needs and Expectations of Generation-Y Human Resources Practitioners: Preparing the Next Wave of Strategic Business Partners," *Journal of Management Research*, Vol. 1, No. 2, pp. 1–25.

Murphy, M. A., and Davidson, G. V. (1991), "Computer-Based Adaptive Instruction: Effects of Learner Control on Concept Learning," *Journal of Computer-Based Instruction*, Vol. 18, No. 2, pp. 51–56.

Murthy, N. N., Challagalla, G. N., Vincent, L. H., and Shervani, T. A. (2008), "The Impact of Simulation Training on Call Center Agent Performance: A Field-Based Investigation," *Management Science*, Vol. 54, No. 2, pp. 384–399.

Narayan, A., and Steele-Johnson, D. (2007), "Relationships between Prior Experience of Training, Gender, Goal Orientation, and Training Attitudes," *International Journal of Training and Development*, Vol. 11, No. 3, pp. 166–180.

National Safety Council (2010), *Summary from Injury Facts, 2010 Edition*, available: http://www.nsc.org/news_resources/injury_and_death_statistics/Documents/Summary_2010_Ed.pdf.

Nijman, D., Nijhof, W. J., Wognum, A. A. M., and Veldkamp, B. P. (2006), "Exploring Differential Effects of Supervisor Support on Transfer of Training," *Journal of European Industrial Training*, Vol. 30, No. 7, pp. 529–549.

Noe, R. A. (1986), "Trainee Attributes and Attitudes: Neglected Influences on Training Effectiveness," *Academy of Management Review*, Vol. 4, pp. 736–749.

Noe, R. A., and Schmitt, N. (1986), "The Influence of Trainee Attitudes on Training Effectiveness: Test of a Model," *Personnel Psychology*, Vol. 39, pp. 497–523.

Ochieng, E. G., and Price, A. D. (2009), "Framework for Managing Multicultural Project Teams," *Engineering, Construction, and Architectural Management*, Vol. 16, No. 6, pp. 527–543.

Ong, J., and Ramachandran, S. (2003), "Intelligent Tutoring Systems: Using AI to Improve Training Performance and ROI," available: http://www.shai.com/papers, accessed July 18.

Ortiz, G. A. (1994), "Effectiveness of PC-Based Flight Simulation," *International Journal of Aviation Psychology*, Vol. 4, pp. 285–291.

Orvis, K. A., Horn, D. B., and Belanich, J. (2008), "The Roles of Task Difficulty and Prior Videogame Experience on Performance and Motivation in Instructional Videogames," *Computers in Human Behavior*, Vol. 24, pp. 2415–2433.

Orvis, K. A., Fisher, S. L., and Wasserman, M. E. (2009), "Power to the People: Using Learner Control to Improve Trainee Reactions and Learning in Web-Based Instructional Environments," *Journal of Applied Psychology*, Vol. 94, No. 4, pp. 960–971.

Oser, R. L., Cannon-Bowers, J. A., Salas, E., and Dwyer, D. J. (1999a), "Enhancing Human Performance in Technology-Rich Environments: Guidelines for Scenario-Based Training," in *Human/Technology Interaction in Complex Systems*, E. Salas, Ed., JAI Press, Greenwich, CT, pp. 175–202.

Oser, R. L., Gualtieri, J. W., Cannon-Bowers, J. A., and Salas, E. (1999b), "Training Team Problem Solving Skills: An Event-Based Approach," *Computers in Human Behavior*, Vol. 15, Nos. 3–4, pp. 441–462.

Paradise, A. (2007), *State of the Industry: ASTD's Annual Review of Trends in Workplace Learning and Performance*, American Society for Training & Development, Alexandria, VA.

Paradise, A. (2008), "Investment in Learning Remains Strong," *Training and Development*, Vol. 62, No. 11, pp. 44–51.

Paradise, A., and Patel, L. (2009), "State of the Industry: ASTD's Annual Review of Trends in Workplace Learning and Performance," American Society for Training & Development, Alexandria, VA.

Pasher, H., McDaniel, M., Rohrer, D., and Bjork, R. (2008), "Learning Styles: Concepts and Evidence," *Psychological Science in the Public Interest*, Vol. 9, pp. 104–119.

Parasuraman, R. (2009). "Assaying Individual Differences in Cognition with Molecular Genetics: Theory and Application," *Theoretical Issues in Ergonomics Science*, Vol. 10, No. 5, pp. 399–416.

Parasuraman, R., and Wilson, G. F. (2008), "Putting the Brain to Work: Neuroergonomics Past, Present, and Future," *Human Factors*, Vol. 50, pp. 468–474.

Peters, L. H., O'Connor, E. J., and Eulberg, J. R. (1985), "Situational Constraints: Sources, Consequences, and Future Considerations," in *Research in Personnel and Human Resources Management*, Vol. 3, K. Rowland and G. Ferris, Eds., JAI Press, Greenwich, CT, pp. 79–113.

Phillips, J. M., and Gully, S. M. (1997), "Role of Goal Orientation, Ability, Need for Achievement, and Locus of Control in the Self-Efficacy and Goal Setting Process," *Journal of Applied Psychology*, Vol. 82, pp. 792–802.

Prince, C., and Jentsch, F. (2001), "Aviation Crew Resource Management Training with Low-Fidelity Devices," in *Improving Teamwork in Organizations*, E. Salas, C. A. Bowers, and E. Edens, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 147–164.

Prince, C., Oser, R., Salas, E., and Woodruff, W. (1993), "Increasing Hits and Reducing Misses in CRM/LOS Scenarios: Guidelines for Simulator Scenario Development," *International Journal of Aviation Psychology*, Vol. 3, No. 1, pp. 69–82.

Quinones, M. A. (1995), "Pretraining Context Effects: Training Assignment as Feedback," *Journal of Applied Psychology*, Vol. 80, pp. 226–238.

Quinones, M. A. (1997), "Contextual Influencing on Training Effectiveness," in *Training for a Rapidly Changing Workplace: Applications of Psychological Research*, M. A. Quinones and A. Ehrenstein, Eds., American Psychological Association, Washington, DC, pp. 177–200.

Ralphs, L. T., and Stephan, E. (1986), "HRD in the Fortune 500," *Training and Development Journal*, Vol. 40, pp. 69–76.

Ree, M. J., Carretta, T. R., and Teachout, M. S. (1995), "Role of Ability and Prior Job Knowledge in Complex Training Performance," *Journal of Applied Psychology*, Vol. 80, pp. 721–730.

Ricci, K. E., Salas, E., and Cannon-Bowers, J. A. (1995), "Do Computer Based Games Facilitate Knowledge Acquisition and Retention?" *Military Psychology*, Vol. 8, pp. 295–307.

Richman, H. B., Staszewski, J. J., and Simon, H. A. (1995), "Simulation of Expert Memory Using EPAM IV," *Psychological Review*, Vol. 102, pp. 305–330.

Rivera, R. J., and Paradise, A. (2006), "State of the Industry in Leading Enterprises: ASTD's Annual Review of Trends in Workplace Learning and Performance," American Society for Training & Development, Alexandria, VA.

Roberson, L., Kulik, C. T., and Pepper, M. B. (2009), "Individual and Environmental Factors Influencing the Use of Transfer Strategies after Diversity Training," *Group & Organization Management*, Vol. 34, No. 1, pp. 67–89.

Roenker, D. L., Cissell, G. M., Ball, K. K., Wadley, V. G., and Edwards, J. E. (2003), "Speed-of Processing and Driving Simulator in Training Result in Improved Driving Performance," *Human Factors*, Vol. 45, No. 2, pp. 218–233.

Roosa, M. W., Dumka, L. E., Gonzales, N. A., and Knight, G. P. (2002), "Cultural/Ethnic Issues and the Prevention Scientist in the 21st Century," *Prevention and Treatment*, Vol. 5, available: http://journals.apa.org/prevention/volume5/pre0050005a.html.

Rosen, M. A., Salas, E., Wilson, K. A., King, H. B., Salisbury, M., Augenstein, J. S., Robinson, D. W., and Birnbach, D. J. (2008), "Measuring Team Performance in Simulation-Based Training: Adopting Best Practices for Healthcare," *Journal of the Society for Simulation in Healthcare*, Vol. 3, No. 1, pp. 33–41.

Rossett, A., and Gautier-Downes, J. (1991), *A Handbook of Job Aids*, Jossey-Bass/Pfeiffer, San Francisco.

Rossett, A., and Marshall, J. (2010), "E-learning: What's Old Is New Again," *Training and Development*, Vol. 64, No. 1, pp. 34–38.

Rouiller, J. Z., and Goldstein, I. L. (1993), "The Relationship between Organizational Transfer Climate and Positive Transfer of Training," *Human Resource Development Quarterly*, Vol. 4, pp. 377–390.

Rowold. J. (2007), "The Impact of Personality on Training-Related Aspects of Motivation: Test of a Longitudinal Model," *Human Resource Development Quarterly*, Vol. 18, No. 1, pp. 9–31.

Ruben, B. D. (1999), "Simulations, Games, and Experience-Based Learning: The Quest for a New Paradigm for Teaching and Learning," *Simulation and Gaming*, Vol. 30, No. 4, pp. 498–505.

Sackett, P. R., and Mullen, E. J. (1993), "Beyond Formal Experimental Design: Towards an Expanded View of the Training Evaluation Process," *Personnel Psychology*, Vol. 46, pp. 613–627.

Sacks, M. (1994), *On-the-Job Learning in the Software Industry*, Quorum Books, Westport, CT.

Saks, A. M. (1995), "Longitudinal Field Investigation of the Moderating and Mediating Effects of Self-Efficacy on the Relationship Between Training and Newcomer Adjustment," *Journal of Applied Psychology*, Vol. 80, pp. 221–225.

Saks, A. M., and Belcourt, M. (2006), "An Investigation of Training Activities and Transfer of Training in Organizations," *Human Resource Management*, Vol. 45, No. 4, pp. 629–648.

Salas, E., and Cannon-Bowers, J. A. (1997), "Methods, Tools, and Strategies for Team Training," in *Training for a Rapidly Changing Workplace: Applications of Psychological Research*, M. A. Quinones and A. Ehrenstein, Eds., American Psychological Association, Washington, DC, pp. 249–279.

Salas, E., and Cannon-Bowers, J. A. (2000a), "Designing Training Systematically," in *The Blackwell Handbook of Principles of Organizational Behavior*, E. A. Locke, Ed., Blackwell, Malden, MA, pp. 43–59.

Salas, E., and Cannon-Bowers, J. A. (2000b), "The Anatomy of Team Training," in *Training and Retraining: A Handbook for Business, Industry, Government, and the Military*, S. Tobias and J. D. Fletcher, Eds., Macmillan, New York, pp. 312–335.

Salas, E., and Cannon-Bowers, J. A. (2001), "The Science of Training: A Decade of Progress," *Annual Review of Psychology*, Vol. 52, pp. 471–499.

Salas, E., and Klein, G., Eds. (2000), *Linking Expertise and Naturalistic Decision Making*, Lawrence Erlbaum Associates, Mahwah, NJ.

Salas, E., and Kozlowski, S. W. J. (2010), "Learning, Training and Development in Organizations: Much Progress and a Peek Over the Horizon," in *Learning, Training, and Development in Organizations*, S. W. J. Kozlowski and E. Salas, Eds., Taylor and Francis Group, New York, pp. 461–476.

Salas, E., and Stagl, K.C. (2009), "Design Training Systematically and Follow the Science of Training," in *Handbook of Principles of Organizational Behavior: Indispensible knowledge for Evidence-Based Management*, 2nd ed., E. Locke, Ed., John Wiley & Sons, Hoboken, NJ, pp. 59–84.

Salas, E., Burgess, K. A., and Cannon-Bowers, J. A. (1995), "Training Effectiveness Techniques," in *Research Techniques in Human Engineering*, J. Weiner, Ed., Prentice-Hall, Englewood Cliffs, NJ, pp. 439–471.

Salas, E., Bowers, C. A., and Blickensderfer, E. (1997a), "Enhancing Reciprocity between Training Theory and Training Practice: Principles, Guidelines, and Specifications," in *Improving Training Effectiveness in Work Organizations*, J. K. Ford, S. Kozlowski, K. Kraiger, E. Salas, and M. Teachout, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 19–46.

Salas, E., Cannon-Bowers, J. A., and Johnston, J. H. (1997b), "How Can You Turn a Team of Experts into an Expert Team?: Emerging Training Strategies," in *Naturalistic Decision Making*, C. E. Zsambok and G. Klein, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 359–370.

Salas, E., Bowers, C. A., and Rhodenizer, L. (1998), "It Is Not How Much You Have but How You Use It: Toward a Rational Use of Simulation to Support Aviation Training," *International Journal of Aviation Psychology*, Vol. 8, pp. 197–208.

Salas, E., Rozell, D., Mullen, B., and Driskell, J. E. (1999), "The Effect of Team Building on Performance: An Integration," *Small Group Research*, Vol. 30, No. 3, pp. 309–329.

Salas, E., Burke, C. S. Bowers, C. A., and Wilson, K. A. (2001), "Team Training in the Skies: Does Crew Resource Management (CRM) Training Work?" *Human Factors*, Vol. 43, No. 4, pp. 641–674.

Salas, E., Burke, C. S., Fowlkes, J. E., and Wilson, K. A. (2004), "Challenges and Approaches to Understanding Leadership Efficacy in Multi-Cultural Teams," in *Advances in Human Performance and Cognitive Engineering Research*, Vol. 4, M. Kaplan, Ed., Elsevier, Oxford, pp. 341–384.

Salas, E., Priest, H., Wilson, K., and Burke, C. S. (2006a), "Scenario-Based Training: Improving Military Mission Performance and Adaptability," in *Military Life: The Psychology of Serving in Peace and Combat*, Vol. 2, A. B. Adler, C. A. Castro, and T. W. Britt, Eds., Praeger Security International, Westport, CT, pp. 32–53.

Salas, E., Wilson, K. A., Priest, H. A., and Guthrie, J. (2006b), "Training in Organizations: The Design, Delivery, and Evaluation of Training Systems," in *Handbook of Human Factors and Ergonomics* (3$^{rd}$ ed), G. Salvendy Ed., John Wiley and Sons, Hoboken, NJ, pp. 472–512.

Salas, E., Stagl, K. C., Burke, C. S., and Goodwin, G. F. (2007), "Fostering Team Effectiveness in Organizations: Toward an Integrative Theoretical Framework of Team Performance," in *Modeling Complex Systems: Motivation, Cognition and Social Processes*, J. W. Shuart, W. Spaulding, and J. Poland, Eds., Nebraska Symposium on Motivation, Vol. 52, University of Nebraska Press, Lincoln, NE.

Salas, E., Cooke, N. J., and Rosen, M. A. (2008), "On Teams, Teamwork, and Team Performance: Discoveries and Developments," *Human Factors*, Vol. 50, No. 3, pp. 540–547.

Salas, E., and Kozlowski. S. W. J. (2010). "Learning, Training, and Development in Organizations: Much Progress and Peak over the Horizon. Learning, Training, and Development in Organizations," in *Learning, Training, and Development in Organizations*, S. W. J. Kozlowski and E. Salas, Eds., Taylor and Francis Group, New York, NY, pp 461–476.

Salas, E., Wildman, J., and Piccolo, R. (2009), "Using Simulation-Based Training to Enhance Management Education," *Academy of Management Learning & Education*, Vol. 8, No. 4, pp. 559–573.

Sanchez-Burks, J., Lee, F., Nisbett, R., and Ybarra, O. (2007), "Cultural Training Based on a Theory of Relational Ideology," *Basic and Applied Social Psychology*, Vol. 29, No. 3, pp. 257–268.

Satish, U., and Streufert, S. (2002), "Value of a Cognitive Simulation in Medicine: Towards Optimizing Decision Making Performance of Healthcare Personnel," *Quality and Safety in Health Care*, Vol. 11, pp. 163–167.

Scaduto, A., Lindsay, D., and Chiaburu, D. S. (2008), "Leader Influences on Training Effectiveness: Motivation and Outcome Expectation Processes," *International Journal of Training and Development*, Vol. 12, No. 3, pp. 158–170.

Scandura, T. A., Tejeda, M. J., Werther, W. B., and Lankau, M. J. (1996), "Perspectives on Mentoring," *Leadership and Organizational Development Journal*, Vol. 17, pp. 50–56.

Schaubroeck, J., and Jones, J. (2000), "Antecedents of Workplace Emotional Labor Dimensions and Moderators of Their Effects on Physical Symptoms," *Journal of Organizational Behavior*, Vol. 21, Special Issue, pp. 163–183.

Schein, E. H. (1992), *Organizational Culture and Leadership*, 2nd ed., Jossey-Bass, San Francisco.

Schmidt, A. M., and Ford, J. K. (2003), "Learning within a Learner Control Training Environment: The Interactive Effects of Goal Orientation and Metacognitive Instruction on Learning Outcomes," *Personnel Psychology*, Vol. 56, pp. 405–429.

Schmidt, R. A., and Bjork, R. A. (1992), "New Conceptualizations of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training," *Psychological Science*, Vol. 3, pp. 207–217.

Schraagen, J. M., Chipman, S. F., and Shalin, V. L., Eds. (2000), *Cognitive Task Analysis*, Lawrence Erlbaum Associates, Mahwah, NJ.

Schwarzer, R. (2008), "Modeling Health Behavior Change: How to Predict and Modify the Adoption and Maintenance of Health Behaviors," *Applied Psychology: An International Review*, Vol. 57, No. 1, pp. 1–29.

Seddon, J. (2008), "Vets and Videos: Student Learning from Context-Based Assessment in a Pre-Clinical Science Course," *Assessment & Evaluation in Higher Education*, Vol. 33, No. 5, pp. 559–566.

Selmer, J. (2001), "The Preference for Predeparture or Postarrival Cross-Cultural Training: An Exploratory Approach," *Journal of Managerial Psychology*, Vol. 16, pp. 50–58.

Selmer, J., Torbiörn, I., and de Leon, C. T. (1998), "Sequential Cross-Cultural Training for Expatriate Business Managers: Pre-Departure and Post-Arrival," *International Journal of Human Resource Management*, Vol. 9, pp. 831–840.

Serfaty, D., Entin, E. E., Johnston, J. H., and Cannon-Bowers, J. A. (1998), "Team Coordination Training," in *Making Decisions under Stress: Implications for Individual and Team Training*, J. A. Cannon-Bowers and E. Salas, Eds., American Psychological Association, Washington, DC, pp. 221–245.

Shay, J. P., and Tracey, J. B. (2009), "Expatriate Adjustment and Effectiveness: The Mediating Role of Managerial Practices," *Journal of International Management*, Vol. 15, pp. 401–412.

Shebilske, W. L., Regian, J. W., Arthur, W., Jr., and Jordan, J. A. (1992), "A Dyadic Protocol for Training Complex Skills," *Human Factors* Vol. 34, pp. 369–374.

Shebilske, W. L., Jordan, J. A., Goettl, B. P., and Paulus, L. E. (1998), "Observation versus Hands-on Practice of Complex Skills in Dyadic, Triadic, and Tetradic Training-Teams," *Human Factors* Vol. 40, pp. 525–540.

Shee, D. Y., and Wang, Y. (2008), "Multi-Criteria Evaluation of the Web-based E-learning System: A Methodology Based on Learner Satisfaction and Its Applications," *Computers & Education*, Vol. 50, pp. 894–905.

Shen, J. (2005), "International Training and Mangement Development: Theory and Reality," *Journal of Management Development*, Vol. 24, No. 7/8, pp. 656–666.

Shuffler, M. L., Salas, E., and Xavier, L. F. (2010), "The Design, Delivery, and Evaluation of Crew Resource Management Training," in *Crew Resource Management*, 2nd ed., B. Kanki, R. Helmreich, and J. Anca, Eds., Academic Press, San Diego, CA, pp. 205–232.

Shute, V. J., and Gawlick, L. A. (1995), "Practice Effects on Skill Acquisition, Learning Outcome, Retention, and Sensitivity to Relearning," *Human Factors*, Vol. 37, pp. 781–803.

Simon, S. J., and Werner, J. M. (1996), "Computer Training through Behavior Modeling, Self-Paced, and Instructional Approaches: A Field Experiment," *Journal of Applied Psychology*, Vol. 81, pp. 648–659.

Sitzmann, T., Kraiger, K., Stewart, D., and Wisher, R. (2006), "The Comparative Effectiveness of Web-Based and Classroom Instruction: A Meta-Analysis," *Personnel Psychology*, Vol. 59, pp. 623–664.

Sitzmann, T., Brown, K. G., Casper, W. J., Ely, K., and Zimmerman, R. D. (2008), "A Review and Meta-Analysis of the Nomological Network of Trainee Reactions," *Journal of Applied Psychology*, Vol. 93, No.2, pp. 280–295.

Skarlicki, D. P., and Latham, G. P. (1997), "Leadership Training in Organizational Justice to Increase Citizenship Behavior within a Labor Union: A Replication," *Personnel Psychology*, Vol. 50, pp. 617–633.

Smith-Jentsch, K. A., Jentsch, F. G., Payne, S. C., and Salas, E. (1996a), "Can Pretraining Experiences Explain Individual Differences in Learning?" *Journal of Applied Psychology*, Vol. 81, pp. 909–936.

Smith-Jentsch, K., Salas, E., and Baker, D. P. (1996b), "Training Team Performance-Related Assertiveness," *Personnel Psychology*, Vol. 49, pp. 909–936.

Smith-Jentsch, K. A., Zeisig, R. L., Acton, B., and McPherson, J. A. (1998), "Team Dimensional Training: A Strategy for Guided Team Self-Correction," in *Making Decisions under Stress: Implications for Individual and Team Training*, J. A. Cannon-Bowers and E. Salas, Eds., American Psychological Association, Washington, DC, pp. 271–297.

Snyder, R., Raben, C., and Farr, J. (1980), "A Model for the Systematic Evaluation of Human Resource Development Programs," *Academy of Management Review*, Vol. 5, No. 3, pp. 431–444.

Steadman, R. H., Coates, W. C., Huan, Y. M., Matevosian, R., Larmon, B. R., McCullough, L., and Ariel, D. (2006), "Simulation-Based Training Is Superior to Problem-Based Learning for the Acquisition of Critical Assessment and Management Skills," *Critical Care Medicine*, Vol. 34, No. 1, pp. 151–157.

Steedman, H. (2005), "Apprenticeship in Europe: 'Fading' or Flourishing?" available: http://cep.lse.ac.uk/pubs/download/dp0710.pdf, accessed July 9, 2010.

Steele-Johnson, D., and Hyde, B. G. (1997), "Advanced Technologies in Training: Intelligent Tutoring Systems and Virtual Reality," in *Training for a Rapidly Changing Workplace: Applications of Psychological Research*, M. A. Quinones and A. Ehrenstein, Eds., American Psychological Association, Washington, DC, pp. 225–248.

Steinberg, E. R. (1977), "Review of Student Control in Computer-Assisted Instruction," *Journal of Computer-Based Instruction*, Vol. 3, pp. 84–90.

Steiner, I. D. (1972), *Group Process and Productivity*, Academic, New York.

Steinwachs, B. (1995), "Barnga: A Game for All Seasons," in S.M. Fowler and M. G. Munford Eds., *Intercultural Sourcebook: Cross-Cultural Training Methods*, Intercultural Press, Boston, MA, pp. 101–108.

Stevens, C. K., and Gist, M. E. (1997), "Effects of Self-Efficacy and Goal-Orientation Training on Negotiation Skill Maintenance: What Are the Mechanisms?" *Personnel Psychology*, Vol. 50, pp. 955–978.

Sugrue, B., and Rivera, R. J. (2005), "State of the Industry: ASTD's Annual Review of Trends in Workplace Learning and Performance," American Society for Training & Development, Alexandria, VA.

Swain, R.W. (2005), "Changes in Instructional System Design (ISD): Improving Training Product Delivery to United States Army Soldiers." Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA431718\&Location =U2\&doc=GetTRDoc.pdf.

Swezey, R. W. (1987), "Design of Job Aids and Procedure Writing," in *Handbook of Human Factors and Ergonomics*, G. Salvendy, Ed., Wiley, New York, pp. 1039–1057.

Swezey, R. W., and Llaneras, R. E. (1997), "Models in Training and Instruction," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York, pp. 514–577.

Tannenbaum, S. I., and Yukl, G. (1992), "Training and Development in Work Organizations," *Annual Review of Psychology*, Vol. 43, pp. 399–441.

Tannenbaum, S. I., Mathieu, J. E., Salas, E., and Cannon-Bowers, J. A. (1991), "Meeting Trainees' Expectations: The Influence of Training Fulfillment on the Development of Commitment," *Journal of Applied Psychology*, Vol. 76, pp. 759–769.

Tannenbaum, S. I., Cannon-Bowers, J. A., and Mathieu, J. E. (1993), "Factors That Influence Training Effectiveness: A Conceptual Model and Longitudinal Analysis," Report 93-011, Naval Training Systems Center, Orlando, FL.

Tannenbaum, S. I., Beard, R. L., McNall, L. A., and Salas, E. (2010), "Informal Learning and Development in Organizations," in *Learning, Training, and Development in Organizations*, S. W. J. Kozlowski and E. Salas, Eds., Taylor and Francis Group, New York, pp. 303–332.

Taylor, P. J., Russ-Eft, D. F., and Chan, D. W. L. (2005), "A Meta-Analytic Review of Behavior Modeling Training," *Journal of Applied Psychology*, Vol. 90, No. 4, pp. 692–709.

Tennyson, R. D. (1980), "Instructional Control Strategies and Content Structure as Design Variables in Concept Acquisition Using Computer-Based Instruction," *Journal of Educational Psychology*, Vol. 72, No. 4, pp. 525–532.

Thayer, P. W., and Teachout, M. S. (1995), "A Climate for Transfer Model," Report AL/HR-TP-1995-0035, Air Force Material Command, Brooks Air Force Base, TX.

Thomas, D. C. (1999), "Cultural Diversity and Work Group Effectiveness," *Journal of Cross-Cultural Psychology*, Vol. 30, No. 2, pp. 242–263.

Thomas, V. J., and Cohn, T. (2006), "Communication Skills and Cultural Awareness Courses for Healthcare Professionals Who Care for Patients with Sickle Cell Disease," *Issues and Innovations in Nursing Education*, Vol. 1, No. 1, pp. 480–488.

Tichon, J. (2007a), "The Use of Expert Knowledge in the Development of Simulations for Train Driver Training," *Cogition, Technology, and Work*, Vol. 9, No. 4, pp. 177–187.

Tichon, J. (2007b), "Training Cognitive Skills in Virtual Reality: Measuring Performance," *CyberPsychology & Behavior*, Vol. 10, No. 2, pp. 286–289.

Toomela, A. (2007), "Sometimes One is More Than Two: When Collaboration Inhibits Knowledge Construction," *Integrative Psychological & Behavioral Science*, Vol. 41, No. 2, pp. 198–207.

Townsend, A. M., DeMarie, S. M., and Hendrickson, A. R. (1996), "A Multi-Disciplinary Approach to the Study of Virtual Workgroups," in *Proceedings of the 1996 27th Annual Meeting of the Decision Sciences Institute, Orlando, FL, Decision Sciences Institute, Atlanta, GA*.

Tracey, B. J., Tannenbaum, S. I., and Kavanagh, M. J. (1995), "Applying Trained Skills on the Job: The Importance of the Work Environment," *Journal of Applied Psychology*, Vol. 80, pp. 239–252.

Tracey, J. B., Hinkin, T. R., Tannenbaum, S. I., and Mathieu, J. E. (2001), "The Influence of Individual Characteristics and the Work Environment on Varying Levels of Training Outcomes," *Human Resource Development Quarterly*, Vol. 12, pp. 5–24.

Triandis, H. C. (2000), "Culture and Conflict," *International Journal of Psychology*, Vol. 35, No. 2, pp. 145–152.

Tsui, A. S., Nifadkar, S. S., and Ou, A. Y. (2007), "Cross-National, Cross-Cultural Organizational Behavior Research: Advances, Gaps, and Recommendation," *Journal of Management*, Vol. 33, No. 3, pp. 426–478.

Tziner, A., Fisher, M., Senior, T., and Wesiberg, J. (2007), "Effects of Trainee Characteristics on Training Effectiveness," *International Journal of Selection and Assessment*, Vol. 15, No. 2, pp. 167–174.

Vandewalle, D. (1997), "Development and Validation of a Work Domain Goal Orientation Instrument," *Educational Psychology Measurement*, Vol. 57, pp. 995–1015.

van Merriënboer, J., Kester, L., and Paas, F. (2006), "Teaching Complex Rather Than Simple Tasks: Balancing Intrinsic and Germane Load to Enhance Transfer of Learning," *Applied Cognitive Psychology*, Vol. 20, pp. 343–352.

Vecchi, G. M., Van Hasselt, V. B., and Romano, S. J. (2005), "Crisis (Hostage) Negotiation: Current Strategies and Issues in High-Risk Conflict Resolution," *Aggression and Violent Behavior*, Vol. 10, No. 5, pp. 533–551.

Velada, R., Caetano, A., Michel, J. W., Lyons, B. D., and Kavanagh, M. J. (2007), "The Effects of Training Design, Individual Characteristics and Work Environment on Transfer of Training," *International Journal of Training and Development*, Vol. 11, No. 4, pp. 282–294.

Villachica, S. W., and Stone, D. L. (2010), "Cognitive Task Analysis: Research and Experience," in *Handbook of Improving Performance in the Workplace: Instructional Design and Training Delivery*, K. H. Silber and W. R. Foshay, Eds., Pfeiffer, San Francisco, pp. 227–258.

Volpe, C. E., Cannon-Bowers, J. A., Salas, E., and Spector, P. E. (1996), "The Impact of Cross-Training on Team Functioning: An Empirical Investigation," *Human Factors*, Vol. 38, pp. 87–100.

Watson, W. E., Kumar, K., and Michaelson, K. K. (1993), "Cultural Diversity's Impact on Interaction Process and Performance: Comparing Homogeneous and Diverse Task Groups," *Academy of Management Journal*, Vol. 36, pp. 590–602.

Waxin, M., and Panaccio, A. (2005), "Cross-Cultural Training to Facilitate Expatriate Adjustment: It Works!" *Personnel Review*, Vol. 34, pp. 51–67.

Wehrenberg, S. B. (1987), "Supervisors as Trainers: The Long-Term Gains of OJT," *Personnel Journal*, Vol. 66, No. 4, pp. 48–51

Wexley, K. N. (1986), "Posttraining Strategies for Facilitating Positive Transfer: An Empirical Exploration," *The Academy of Management Journal*, Vol. 29, No. 3, pp. 503–520.

Wickens, C. D., Helleberg, J., and Xu, X. (2002), "Pilot Maneuver Choice and Workload in Free Flight," *Human Factors*, Vol. 44, No. 2, pp. 171–188.

Wiener, E. L., Kanki, B. G., and Helmreich, R. L., Eds. (1993). *Cockpit Resource Management*, Academic Press, New York.

Williams, T. C., Thayer, P. W., and Pond, S. B. (1991), "Test of a Model of Motivational Influences on Reactions to Training and Learning," paper presented at the 6th Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Wilson, P. F., and Johnson, W. B. (2001), "Core Virtues for the Practice of Mentoring," *Journal of Psychology and Theology*, Vol. 29, No. 2, pp. 121–130.

Wilson, K. A., Priest, H. A., Salas, E., and Burke, C. S. (2005), "Can Training for Safe Practices Reduce the Risk of Organizational Liability?" in *Handbook of Human Factors in Litigation*, I. Noy and W. Karwowski, Eds., Taylor & Francis, London, pp. 6-1–6-32.

Wydra, F. T. (1980), *Learner Controlled Instruction*, Educational Technology Publications, Englewood Cliffs, NJ.

Yang, H., Sackett, P. R., and Arvey, R. D. (1996), "Statistical Power and Cost in Training Evaluation: Some New Considerations," *Personnel Psychology*, Vol. 49, pp. 651–668.

Yoon, S., Velasquez, J., Partridge, B., and Nof, S. (2008), "Transportation Security Decision Support System for Emergency Response: A Training Prototype," *Decision Support Systems*, Vol. 46, No. 1, pp. 139–148.

Zaccaro, S. J., and Bader, P. (2003), "E-Leadership and the Challenges of Leading E-Teams: Minimizing the Bad and Maximizing the Good," *Organizational Dynamics*, Vol. 31, No. 4, pp. 377–387.

Zachary, W., Santarelli, T., Lyons, D., Bergondy, M., and Johnston, J. (2001), "Using a Community of Intelligent Synthetic Entities to Support Operational Team Training." Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA525861&Location=U2&doc=GetTRDoc.pdf

Zachary, W. W., Ryder, J. M., and Hicinbothom, J. H. (1998), "Cognitive Task Analysis and Modeling of Decision Making in Complex Environments," in *Making Decisions Under Stress: Implications for Individual and Team Training*, J. A. Cannon-Bowers and E. Salas, Eds., American Psychological Association, Washington, DC, pp. 313–344.

Zsambok, C., and Klein, G., Eds. (1997), *Naturalistic Decision Making*, Lawrence Erlbaum Associates, Mahwah, NJ.

# CHAPTER 18

# HUMAN FACTORS IN ORGANIZATIONAL DESIGN AND MANAGEMENT

**Pascale Carayon, Peter Hoonakker, and Michael J. Smith**
**University of Wisconsin–Madison**
**Madison, Wisconsin**

## 1 INTRODUCTION

As we write this chapter the world is starting a very slow recovery from the greatest economic recession since the 1930s. There is high unemployment, a lack of new jobs, low investment in businesses, and slow growth in developing countries. The demand for products and services is down. Companies are looking for ways to cut costs, reduce overhead, and right size. This puts pressure on management to cut budgets and staffing and to squeeze higher productivity and quality from all company resources, including capital, technology, and people. In such a demanding and difficult economic environment there is a tendency for companies to develop organizational design and management policies and practices that create negative corporate cultures for the people resources. At the same time there is increased emphasis on using new "technology" to bring substantial benefits for the economic prosperity of a company. This leads to devaluing the people resources and increases job stress, low motivation, and employee apathy.

In this chapter we will focus on how to build an organizational design and management culture and process that permits people to be highly motivated, productive, and effective and lets them have a high quality of working life and be safe. A central assumption of this approach is that meeting the expectations and needs of the people in an organization is necessary for both the short- and long-term economic success of the organization.

Consistent with many organizational design and management authors (Carayon and Smith, 2000; Deming, 1986; Hackman, 2002; Hendrick, 1986, 1987; Lawler, 1986, 2003; McGregor, 1960; Smith, 1965; Smith and Carayon-Sainfort, 1989; Smith and Carayon, 1995), we view employees as the foundation on which to build a successful sustainable and healthy organization.

As we proceed, we present our perspectives on organizational design, operations, and management. Our beliefs are grounded in theory, research, and practice, and that gives us confidence that these perspectives can be helpful to many organizations.

This chapter begins by providing some background context about the history of organizations and organizational design and management. We then discuss how this context provides a basis for a greater focus on the importance of finding meaning in work and life. We describe the key attributes of healthy and sustainable work organizations, organizations that are able to respond to the difficult circumstances of the current economic reality

and future growth. We then present the work system model and design principles and methods that provide more specific guidance on actions that can be taken to create a healthy and sustainable work organization.

## 2 BACKGROUND AND HISTORICAL PERSPECTIVE ON ODAM

The background in organizational design and management discussed here comes from the traditions in North American and Western European conceptualizations of organizations and how they operate. We recognize that there are many other traditions and concepts with rich histories and important knowledge. We believe that many of these other traditions provide reasonable alternatives to our approach. We have focused our approach to the societal, cultural, and management processes based on the behaviors of organizations and individuals in the regions of the world where we have our greatest knowledge and experience.

Organizational structure has always been a primary consideration in organizational design and management. Since the dawn of human history there have been various organizational structures for groups of people who have come together around common interests. Over time, the structures of these groups have increased in their complexity as the size of the organizations grew and as varied interests had to be accommodated. As organizations became more complex, they formed multiple sectors or departments of specialization. This mirrored the growing complexity of society with multiple sectors, such as government, military, judicial, religious, agricultural, marketplace, trading, financial, construction, artisan/professional, and labor. Work organizations developed a similar makeup, with management, finance, security, manufacturing, marketing, sales, customer services, transportation, and human resources. This complexity led to a need for structure, rules, and procedures to provide effective and efficient operations.

As Smith (1965) reported, very early in human history a dominant organizational structure, the hierarchy, was developed. One leader at the top had absolute power and then delegated authority and responsibility downward in a pyramid-type progression. The greatest power and authority were held by those few near the top, while little was present near the base. This is the well-known top-down power structure of an organization. Independent of the nature of the leadership, the structure of the organization typically followed a military structure with a top leader (CEO), generals (vice presidents), colonels (division managers), lieutenants (department managers), sergeants (supervisors), corporals (lead workers), and privates (employees), with orders flowing downward. To this day, this type of hierarchy and military style of command structure remains the dominant structure of small and large organizations. One addition of the few that have commonly been implemented is the "board of directors" at the top of the pyramid representing stockholders.

Characteristic of this structure is the chain of command, with the orders for action flowing from the top down through the organization to the bottom. Functions, knowledge, and skills tend to be specialized within specific units of the organization, sometimes referred to as divisions or departments. This structure requires the coordination and integration of expertise within and across the various units into a unified operation, and this is accomplished through the management process. At the higher levels in the hierarchy, the management functions are more similar than they are different, but specialized knowledge may still differentiate one unit from another. However, at the lowest level in the hierarchy, the activities of the people are quite different across departments, and front-line supervisors are well versed in the day-to-day details of how specific activities in their department are carried out with little knowledge of the specifics of other departments.

In the early twentieth century much emphasis was put on the specialization of function and knowledge at the department level, the supervisor level, and the individual employee level. The purpose was to develop greater expertise by focusing the attention and skills of the workforce to a limited number that could be mastered. To build competence and skill, researchers and practitioners used scientific measurement methods and motivational theory to improve employee performance and productivity substantially. This led to highly structured work activities that required focused knowledge and intelligence and highly developed perceptual–motor skills. The workforce responded positively to the specialization of function. People learned new skills and took pride in the quality of what they produced, and their wages and standard of living increased substantially.

Over the next several decades the workforce became more educated and less satisfied with the narrowly focused and specialized nature of their work, which led to routine, boring work and the realization that opportunities for growth were limited. This led to problems in productivity and serious concerns for employee physical and mental health. New organizational structures and approaches to job design were developed in the early part of the twentieth century but became more widespread in the middle of the twentieth century and progressed through the rest of the century (Black and Porter, 2000; Lawler, 1986, 1996, 2003; Parker and Wall, 1998; Porter et al., 2002). However, even with the growth of new management approaches and job designs, the most dominant organizational approach in the United States and Western Europe remained a hierarchical militarylike structure with top-down power, authority, and decision making. This structure was difficult to discard because it was effective in getting the orders followed, departments coordinated, and products and services produced and delivered in a consistent manner.

Some of the important lessons that we believe emerged from studying the organization and management of work over the past 100 years are that (1) the hierarchy structure of management and control produces predictable results; (2) an effective hierarchy structure does best when there is a top-down power structure with a strong leader; (3) other forms of organizational structure can be effective, but primarily in

small organizations or in large organizations that operate as an integrated network of small businesses or in environments with high uncertainty; (4) people at all levels of the hierarchy are critical resources to the success of the organization; and (5) the best organizational designs incorporate the needs and the knowledge of the workforce in managing the organization.

Starting around the beginning of the twentieth century and up to today, much has been learned about how organizational structure and management affect employee satisfaction and performance. Early on we learned that employees responded well to taking orders from supervisors if they trusted the orders (McGregor, 1960; Smith, 1965; Taylor, 1911). They followed explicit instructions in what tasks to do and the directions and specifications about how to do the tasks. This created clear roles and responsibilities for managers and employees, removed role ambiguity, and let employees explicitly know their performance requirements. Employees responded positively to efforts to improve their skills through training and education. They appreciated new tools and technology that reduced the effort needed to do their tasks and increased the rewards that came with high achievement. The application of detailed and careful evaluation of work management, operations, and tasks led to a "scientific" basis for establishing guidelines for employee selection, supervision, the design of work tasks, and performance requirements (Drucker, 2001; McGregor, 1960; Smith, 1965; Taylor, 1911).

The consistent use of scientific work evaluation and design methods was well received by employees when the methods were perceived as unbiased and fair. This established some important human factors considerations in organizational management that led to increased employee satisfaction with their work and less stress. Human factors considerations included developing reasonable and fair work standards and creating an environment in which employees would trust the organization's decisions and feel they were treated fairly.

For the fair treatment of employees and their perceptions of fairness and trust, "scientific" analytical methods and design criteria were used as the basis for work design and management. These scientific methods and criteria were based on sound evidence of validity and reliability and were simple and clear enough to be understood and accepted by the employees. When such "fair," "trustworthy," and "scientific" requirements were applied to managing and designing work, the employees were more satisfied and less stressed and performed better than when arbitrary requirements were applied (Carayon and Smith, 2000; Lawler, 1996, 2003; Smith, 1965, 1987).

Although the formal structure of the organization and the management process are important for organizational success, we have learned that the informal and social aspects also need to be considered since they can influence the effectiveness of the formal elements (Lawler, 1986; Roethlisberger and Dickson, 1942). The informal hierarchy of leadership and management that exists in work organizations can substantially influence the attitudes and behaviors of employees. Social

processes at work can influence cohesiveness within the work unit and how close employees feel toward the organization. Both of these factors can influence employee satisfaction, stress, and productivity. Informal leaders can facilitate management of the organization by their conformity with formal directives or can inhibit organizational management by presenting contrary perspectives and directives to employees. The informal influence can be so subtle that it does not appear to confront management directives directly. Informal influences can also provide assistance to employees in obstructing the goals of management. In organizations that have labor unions, the informal processes tend to be organized and are easier to identify, and it is easier to understand their perspectives. The work group can buffer stressful aspects of work by providing social support and technical assistance to co-workers (Caplan et al., 1975; French, 1963; House, 1981).

To take advantage of the social aspects of work, many companies use special programs or processes to get employees more involved in the improvement of products and the company. For example, they use quality circles and other techniques of total quality improvement to get employees involved in the design of their own work (Carayon et al., 1999; Deming, 1986; Lawler, 1986; Smith and Carayon-Sainfort, 1989). Some companies use climate questionnaires to assess the status of employee satisfaction and stress and to define specific areas of employee concern (Lawler, 1986). These approaches provide data to help management align formal and informal structures. Successful organizations recognize the importance of aligning the formal management process with informal social processes to get employees positively involved in organizational success. An important human factors consideration is for companies to recognize the importance of the informal social aspects of work groups and to provide formal structures and processes that harness the informal group process to benefit the management of the organization and the satisfaction and success of the employees.

The power of the social process in organizations has been recognized and is one of the drivers in the shift from individually based work to using groups, or teams, of employees working together to achieve a goal (Hackman, 2002; Sainfort et al., 2001; Salas et al., 2004). Although it is commonly recognized that many complex issues require a cross-functional team-based approach, teams have also been found to be beneficial for other reasons in jobs that were previously done in isolation or on assembly lines. By creating work teams, an environment that provides social support can be fostered. Social support has been shown to reduce stress for employees (House, 1981). Team-based work processes are observed in a wide variety of organizations, from manufacturing production and assembly processes to service activities and in new product or service design processes (Sundstrom et al., 1990). Companies teach employees about the importance of teamwork, how to interact in a team, and how to coordinate with other teams.

Teams consisting of managers, marketing, sales, engineering, production, and labor are used to solve

critical product design problems and develop new products. When team operations can be achieved for making products, assembling products, providing services, and selling, they provide social and motivational benefits that lead to greater employee satisfaction and performance (Sainfort et al., 2001). However, social pressures from the group can sometimes increase employee stress, and this requires careful monitoring by the organization. Many management approaches call for the inclusion of front-line employees as part of a team formed to resolve product and service quality problems. Teams capitalize on the mixing of perspectives and knowledge to define problems and to develop preferred solutions. Teams capitalize on multiple domains of expertise, and they also capitalize on the social aspects of the process that provide positive feelings of recognition and respect to the individual participants.

An important consideration in managing work is that employees like opportunities to participate in the decision-making process (Haims and Carayon, 1998; Lawler, 1986; McGregor, 1960; Noro and Imada, 1991; Sainfort et al., 2001; Smith and Carayon-Sainfort, 1989; Wilson and Haines, 1997). Employees like to feel important, respected, and appreciated. Positive participation experiences address an employee's social and ego needs. Organizational management processes that incorporate employee participation into production, problem-solving, designing, and/or opinion-sharing activities provide benefits to both the organization and the individual employee (Wilson and Haines, 1997).

## 3 HEALTHY AND PRODUCTIVE ORGANIZATIONS

### 3.1 Historical Perspective

Since the time of Frederick Taylor's principles of scientific management of work and Henry Ford's application of the assembly line to increase production there have been competing approaches to managing the design of work and the management of the workplace. Taylor and Ford emphasized the simplification of tasks, the use of better tools and technology, and close supervision of workers to achieve substantial gains in worker productivity. They were very successful and established a template for work design and management that is still used by many companies around the world a century later. Workers benefitted through higher wages and/or pay incentives for high performance. But there was dissatisfaction with the simplified nature of the tasks and the high-performance demands placed on workers. As unionization of industries spread in the United States in the early to mid-1900s, worker resistance to these approaches increased and social reformers decried the "dehumanization" of work. Psychologists, sociologists, and management theorists began to appreciate work as an important aspect of a person's life that provided more than economic maintenance. Work was associated with self-esteem, peer esteem, social status, and aspects of life accomplishment and satisfaction.

The counter approach to scientific management came from theorists and applied management experts in Europe and the United States. They saw work as a social and psychological process that could be managed to benefit the enterprise as well as the individual worker. The Hawthorne Studies of the 1920s and 1930s brought attention to the importance of group processes, peer pressure, and peer support on worker performance independent of the quality of physical working conditions. The importance of group affiliation, informal group influences, group leadership, and supervision style on worker behavior and performance became apparent. The idea that people worked for more than economic reasons came forward. The need for "affiliation," "meaningful" work, and "recognition" was embraced. The "intrinsic" aspects of work became as important as the "extrinsic" aspects (Herzberg, 1974). There was a belief that workers would excel when given greater autonomy over their work and embrace the opportunity to perform at high levels. The human side of enterprise and human relations at the workplace became a counterbalance to scientific management.

The theories and approaches of Taylor and Ford were opposed by the theories and approaches that proposed "humanizing" work. Over the many intervening decades since the 1920s the underlying beliefs and principles of each of these varying approaches have been debated, studied, competed, and revived in many forms under different names and methods. The basics of the "scientific management" approach have retained the use of measurements and quantification, improved technology, and financial incentives to achieve improvement in worker performance. The basics of the "humanistic" approach have relied on social and psychological processes to achieve improvements in the quality of working life and worker performance.

Concurrent with the development of the humanistic approaches to work design there was a close examination of worker safety and health. Governments implemented laws and regulations to protect the safety and health of workers. As part of this movement there was an interest in how the design of work affected the mental and physical health of workers. Hans Selye's (1950) landmark work on a generalized "stress" syndrome led to interest in how the physical, psychological, and social aspects of work interacted to affect workers' physical and mental health. In addition, an emphasis was put on the "ergonomic" characteristics of work in terms of how job demands affected workers' injuries and illnesses. Aspects of Taylor's approach to work measurement and task analysis and the humanistic approaches to work design started to come together. Integrated theories of the design and management of work were developed, including participatory ergonomics and macroergonomics. These integrated theories encompassed aspects of work measurement, efficiency, productivity, quality, quality of working life, job demands, physical and psychosocial stress management, and worker safety and health into a complete package for designing and managing the workplace. At the heart of the integrated organizational design and management (ODAM) approaches was the concept that happy, satisfied, unstressed workers will be healthy and productive workers. Healthy and productive workers lead to a

healthy and productive enterprise. However, the belief that happy workers are productive workers has been debated for decades and continues to be debated.

## 3.2  Happy, Healthy, and Productive Workers

Since the 1920s many theorists and applied management experts have proposed approaches to achieve happy, satisfied, and productive employees. While there have been differences in content and emphasis in these approaches the basics have remained consistent. Jobs that provide workers with autonomy, control over aspects of work tasks, reasonable physical and psychological demands, and employment security will lead to higher employee satisfaction with work and lower job stress. Such employees will be more productive in terms of the quantity and quality of products, services, customer satisfaction, and developing new ideas and products. These employees will have greater commitment to the enterprise and lower intention to turn over.

In 1975 the U.S. National Academy of Sciences asked a group of scholars to examine the issues of improved quality of working life and productivity at the workplace. Two important articles on improving productivity and job satisfaction/quality of working life were produced. The first, by Katzell et al. (1975), concluded that job satisfaction and productivity do not follow parallel paths. Therefore it did not follow that simply increasing job satisfaction would necessarily lead to greater productivity. They believed that the objective of increasing both job satisfaction and productivity were not incompatible and could be met. But it was not sufficient to increase worker satisfaction and expect greater productivity because the two constructs were only loosely coupled. An array of methods for improving job satisfaction such as job enrichment, management by objectives, autonomous work groups, and participative management, when implemented by themselves, were more likely than not to leave productivity unchanged or at best to improve it marginally.

These approaches could even lead to reductions in productivity by the disruption of ongoing work processes. These scholars concluded that no one approach was sufficient to simultaneously affect both productivity and employee satisfaction significantly. They stated that two barriers limited the potential for increasing job satisfaction to benefit productivity. The first was "resistance to change" and the second was the insistence on focusing on just "one" approach versus using a variety of methods.

Katzell et al. (1975) proposed five concepts that they felt were supported by their examination of the literature to support both increased job satisfaction and productivity improvements:

1. *Critical-Mass Principle.* Organizational changes required to achieve changes in job satisfaction and productivity have to be sufficiently deep and far reaching to make substantial effects and not just transitory effects. The change does not have to be made all at once and can be staged. Changes in job satisfaction and productivity should not be expected until at least several stages have been undertaken.

2. *Motivation Principle.* The elements of the stages need to be integrated around the concept of developing satisfied *and* highly motivated workers. High job satisfaction is not sufficient for high performance. High motivation to perform needs to be tied to high satisfaction. Effective worker performance should be rewarded in whatever terms are meaningful to the individual, be it financial, psychological, or both. The idea is to develop workers who are "committed" to high performance because of a reward structure that leads to high job satisfaction and high productivity.

3. *Shared Benefits.* An extension of the motivational principle is that workers at *all* levels of the organization must see that the organizational changes will benefit them in terms that are important to them.

4. *Job Design.* Changes in job content need to be substantial enough to be perceptible to workers and typically include greater self-regulation, diversity in tasks, meaningfulness, challenge, and social responsibility. In addition, changes in job content need to be part of a larger program of improved policies and practices that has aspects of adequate pay, job security, proper resources, working conditions, increased mutual influence at all levels, and constructive labor–management relations.

5. *Pattern of Control.* There are three levels at which a redistribution of influence and control in organizations are important: the individual job, the work group, and the organization. Increased autonomy or self-regulation is key to increased job satisfaction and productivity for some workers but not all. Providing workers with a voice over what goes on in their work group has been shown to have favorable effects on satisfaction, work motivation, and turnover. Organizations in which members at all levels exercise greater control over what goes on in the organization typically are more productive and have more highly motivated and satisfied workers.

6. *Patterns of Compensation.* Workers who are more highly paid generally like their pay and their jobs better and are less likely to quit or be absent. Workers in a given job who are paid more than other workers with a comparable job are also likely to have higher motivation and productivity, but only if their pay level is linked to their performance.

7. *Systemwide Changes.* Studies have shown that large-scale systemwide organizational changes had greater benefits for productivity and job satisfaction increases. These extensive changes create a "new" work system.

Katzell and colleagues (1975) concluded their evaluation of the organizational change literature findings and

conclusions with six principles for potential success in increasing worker job satisfaction and productivity:

1. The pay of workers must be linked to their performance and productivity gains.
2. Workers and jobs must be matched to meet their capabilities, needs, and expectations and provide the resources for them to succeed.
3. Jobs should provide the opportunities for workers to use their abilities, make meaningful contributions, challenge, diversity of duties, and being responsible for others, but only for those workers who desire these factors.
4. Workers at all levels should have input to plans and decisions affecting their jobs and working lives.
5. Necessary resources must be provided for achievement of high performance.
6. Adequate "hygiene" conditions such as competent supervision, fair pay and fringe benefits, job security, good working conditions, and sound employee relations must exist.

The second paper, by Cummings et al. (1975), examined how to improve the quality of working life and worker productivity. It also examined an extensive literature dealing with job satisfaction, productivity, and organizational design. Their paper looked at three main types of knowledge from the literature for developing effective strategies for improving productivity and worker satisfaction: "action levers," contingencies, and change strategies.

Action levers were seen as the first step in improving working conditions that could be manipulated to create desired changes. The action levers were:

1. Pay/reward systems for performance
2. Autonomy/discretion for workers
3. Support services from technical groups
4. Training of all workers for all jobs in a department
5. Flat organizational structure
6. Rearrangement of the physical plant
7. Task variety
8. Information and feedback from users
9. Increased interpersonal interaction in groups/departments.

Cummings and colleagues (1975) concluded that the literature supported the belief that manipulation of these action levers produced positive outcomes in cost reduction, increased productivity, lower rates of rejected products, decreased employee turnover and absenteeism, and increased job satisfaction. They classified the general approaches to organizational redesign into four categories: (1) sociotechnical/autonomous work groups, (2) job restructuring such as job enrichment, job enlargement and job rotation, (3) participative management, and

(4) structural changes at the organizational level rather than the job or departmental levels.

Cummings et al. (1975) indicated that contingencies were factors that affected the success of the action levers. Action levers were manipulated in situations with human and organizational peculiarities, and these contextual differences needed to be taken into account in any change effort. Their review of the research literature led to the conclusion that the manipulation of the action levers was more likely to produce positive outcomes when the contingencies were used. The contingencies included:

1. Support for the changes from the highest involved level of the organization
2. Addressing worker "higher order" needs
3. Pay and reward systems based on a group rather than an individual basis
4. A technological process that allows for relatively self-contained task groupings
5. An adequate training program for establishing technical skills
6. Recognition that workers are capable of assuming responsibility for a "whole" task
7. First-level supervisors supportive of the changes
8. Changes that deal with abilities valued by workers
9. Changes that do not adversely impact interpersonal relationships
10. Increased participation that is seen as a legitimate part of work
11. No strong resistance to the methods of introducing participation
12. Participation that involves decisions that workers perceive as important
13. Workers possessing a need for independence
14. Participative decisions that motivate a worker to enter and remain in the organization
15. Organizational "climate" that is supportive of innovative behavior
16. High level of trust between workers and managers

Cummings et al. (1975) proposed key aspects of organizational redesign strategies taken from their evaluation of the literature:

1. Pick one of the four strategies (identified above) that fits your organization, but be flexible and incorporate elements of the other three when appropriate.
2. Department- and group-level changes have shorter time scales and smaller impact than systemwide strategies. There will be ripple effects when change occurs, and these must be accommodated for successful implementation. Unanticipated consequences are the rule, not the exception.

3. Decide whether to involve the target group actively in the change process decision making. Even though participation is a central aspect of the change process, there are situations when worker involvement in decisions will be disrupting.

4. Make the first steps in the change process small and adopt an evolutionary approach. Be flexible as the process evolves using data on successes and failures to direct future directions. Start with pilot projects and then widen the scope of the change process.

5. Build information-gathering processes into the change process.

6. Be opportunistic. Take advantage of situations that arise which provide new insights, opportunities, and chances to make gains.

7. Avoid the trap of having to prove the success of the innovations. Think long term and look to the "big picture." Do not focus on small successes or failures, and do not pressure managers over short-term failures. Managers need flexibility to adapt to the unanticipated consequences of changes.

In 1970 the U.S. National Institute for Occupational Safety and Health (NIOSH) undertook a program to define workplace factors that related to employee job stress. It is interesting that this effort had some parallels to the work of the U.S. National Academy of Sciences to examine work organization and productivity. The NIOSH program was looking at characteristics of work organization and work design with an emphasis on worker mental and physical health rather than job satisfaction and quality of working life. One of the first major studies from this program was completed in 1975 at the same time the National Academy of Sciences work design and productivity findings were coming out (Caplan et al., 1975). This study collected questionnaire responses about working conditions, job satisfaction, and health from 2010 employees in 23 occupations. The study established relationships among workers' perceptions of job stressors and behavioral, psychological, and somatic outcomes. An important finding was that the occupations that had more task complexity, higher levels of participation, and lower levels of underutilization of abilities were the most satisfied with their jobs.

Cooper and Marshall (1976) categorized sources of job stress into five categories: (1) factors intrinsic to the job such as demands and dangers, (2) factors that affect a person's role such as role ambiguity and role conflicts, (3) career factors such as promotion and job security, (4) relationships with others at work, and (5) organizational climate issues such as participation. They conducted a comprehensive literature review and concluded that job stressors led to worker stress and recurrent worker stress can lead to chronic disease.

Frankenhaeuser and Gardell (1976) and Gardell (1981a, 1981b) declared that there was an epidemic of stress due to "Taylorization" of work. Further they claimed that "psychosocial" aspects of work defined the "happiness" with work and the level of job stress. In Scandinavia and Western Europe there was a strong movement to promote worker involvement and participation in the design of work to humanize working life. Karasek (1979) found that workers in jobs with high work demands and low decision latitude had an increased risk for coronary heart disease.

What is intriguing about the happy, satisfied, productive worker literature and the job stress literature of the 1970s is that these different approaches to examining work both found that similar aspects of work design and organizational design, management, and culture were associated with better worker productivity and lesser job stress and strain. These aspects of work included matching job content to workers' expectations, skills, and capacity; providing variety in tasks; involving workers in decision making and giving workers control over tasks; providing job growth and security; promoting social support; rewarding good performance; and supportive supervision and management. These job and organizational characteristics led to a "culture" in which both workers and companies benefited.

### 3.3 Recent Perspectives on Happy, Productive Workers and Workplaces

More recent theoretical discussions and research promote and support the idea of happy, healthy, and productive work, and they bring these ideas into a contemporary context. Harter et al. (2002) proposed that worker well-being was in the best interests of communities and organizations. In particular, organizations spend substantial resources in developing human resources in efforts to generate profits and customer loyalty. Harter et al. (2002) carried out an extensive review of the literature examining work, life satisfaction, job satisfaction, health outcomes, and business outcomes. Generally, they found that working conditions affected worker health, stress, and satisfaction and individual employee satisfaction was related to job performance. The happy–productive worker concept was linked to emotional well-being and work performance. In conclusion, they stated that work was a pervasive and influential part of the individual's and community's well-being. It affected the individual's quality of life and mental health and thereby could affect the productivity of entire communities. The organization's ability to promote worker well-being rather than strains and poor emotional health was a benefit to the worker but also a benefit to an organization's bottom line.

Harter et al. (2002) proposed six characteristics of jobs to promote employee well-being and productivity:

1. Basic materials and equipment to carry out the job tasks

2. Role clarity and expectations

3. Feelings of contributing to the organization

4. Sense of belonging

5. Opportunities for growth

6. Chances to discuss personal progress

Wright and Cropanzano (2004) took a "fresh look" at the role of psychological well-being and job performance by taking a "positivist" position on the influences of work and life experiences. Their review of past research concluded that high levels of worker psychological well-being (emotional well-being) led to increased job performance and an increased capacity to appreciate new opportunities and experiences. They indicated that employee-focused, positive psychological-based interventions at work take three general forms:

1. *Composition of Workforce: Worker Selection.* Research has shown that worker psychological well-being and happiness were stable over long periods of time as long as five years. This may provide a possibility of using selection procedures that try to determine the current "happiness" and/or psychological well-being of job applicants. However, they pointed out that selecting the happiest job applicants as workers raises the specter of potentially serious ethical issues.

2. *Training Workforce.* A second option is to train workers in methods to be happier. For example, stress management training can have positive effects on worker happiness. Another strategy is training workers in strategies to change their personal perceptions from negative to positive. Constructive self-talk and other approaches to cognitive restructuring are examples of this. Still another strategy is training workers in "learned optimism," a method emphasizing positive thought patterns. Research has shown that optimistic workers perform more effectively in a wide range of occupations, especially those that require significant interaction with others.

3. *Situational Engineering of Workplace.* This approach requires changing the work environment so that it promotes worker psychological well-being. There is evidence that working conditions strongly affect worker psychological well-being. For example, reengineering the work environment such as the physical, social, work role, task, and job demands has been shown to be related to worker emotions. Family, friendly organizational policies such as flex-time and childcare programs should increase worker psychological well-being.

Quick and Quick (2004) emphasized that it was useful to consider "public health" ideas to manage the health and well-being of the workforce to achieve happy, healthy, and productive workers. From the five principles that they used to develop a "preventive stress management" framework, two were central to managing organizations to achieve worker health and productivity:

1. *Individual health and organizational health are interdependent.* Workers who are unhappy and stressed can drain positive energy that otherwise could be put to use to achieve happiness and productivity. Poor working conditions that lead to psychosocial stress produce "emotional toxins" that undercut motivation, health, and happiness.

2. *Leaders have a responsibility for individual and organizational health.* Leaders can use the tenants of "positive psychology" to promote a healthy and productive workforce. Positive psychology aims to build on an individual's strengths and competencies to promote health, well-being, and effectiveness. When leaders apply the tenants of positive psychology to the entire workforce, then they can develop the workforce into a happy, effective, and productive engine of the organization. Leaders should focus on keeping the workforce and themselves healthy, happy, and productive in the service of the organization.

Quick and Quick (2004) indicated that an emphasis on individual and corporate health was based on having happy, healthy, and effective workers and managers. Leaders and organizations that develop the skills and competencies of workers and managers can optimize the contributions that can be made to productivity and in the process produce healthy and happy workers. As with public health approaches, this approach emphasized preventive programs to keep the workforce healthy, happy, and productive. Such preventive programs were grounded in developing the capabilities of the workforce, eliminating the conditions that diminished worker competency, and designing work that led to workers who were happy and effective.

These concepts are parallel to the older ideas about enhancing worker job satisfaction, designing jobs that reduce stress, and making work meaningful, which will lead to effective and motivated workers that are highly productive.

Grawitch et al. (2006) reviewed the literature about healthy workplaces, employee well-being, and organizational improvements. They identified five workplace practices that had direct and indirect links between healthy workplace practices and organizational improvements:

1. *Providing Work and Life Balance.* Conflicts between work and family life diminish an employee's perceptions of the quality of each. The structure of work has a strong influence on family life; for example, work schedule affects the time available to interact with the family. Work–family conflicts have been tied to increased absenteeism, while corporate work–life balance programs have been tied to worker's organizational commitment and job satisfaction. Family–life balance features such as flex-time and fringe benefits may build employee loyalty and morale.

2. *Employee Growth and Development.* The opportunity to gain additional skills, knowledge, and experiences can act as motivators which can

translate into positive gains for the organiza-
tion. Training and internal career opportunities
have been shown to be significant predictors of
organizational effectiveness and job satisfaction.
Training opportunities were related to less job
stress. The effectiveness of training programs
was enhanced when workers were able to apply
what they learned to their jobs.

3. *Health and Safety.* Implementation of healthy
   workplace and safety initiatives can be seen as
   a form of organizational support of the workers.
   Research has found that high levels of worker
   stress led to higher health care expenditures and
   greater absenteeism. Health promotion programs
   led to lower absenteeism and lower health care
   costs. Stress management programs have been
   shown to decrease absenteeism and increase
   productivity.

4. *Recognition.* Employee recognition has been
   found to be a significant predictor of organiza-
   tional effectiveness, worker job satisfaction, and
   worker stress. In particular, worker compensa-
   tion was critical to a healthy workplace. Com-
   pensation and fringe benefits attract and retain
   workers.

5. *Employee Involvement.* Employee involvement
   has been related to worker satisfaction and
   morale and to lower turnover and absentee-
   ism and higher quality of work/products. Em-
   ployee involvement programs led to benefits for
   workers and for the organization.

These five organizational practices are in line with
the earlier proposals dealing with happy, satisfied, and
productive workers.

### 3.4 Recent Research on Worker Happiness, Attitude, and Productivity

Taris and Schreurs (2009) studied 66 Dutch home health
care organizations to examine the relationship among
workers' emotional health and job satisfaction and
client satisfaction and organizational productivity. Three
separate studies were undertaken: a questionnaire survey
of the organizations' workers, a questionnaire survey of
the organizations' clients, and an accounting study of
the financial condition of the participating organizations.
In the first study responses to the survey on quality of
working life were received from 56,963 workers (48.7%
response rate) in 81 organizations. A second survey
of client satisfaction with services of the organizations
was undertaken almost simultaneously with the first
study. There were 54,987 respondents (51.5% response
rate). The health care organizations were asked to
voluntarily participate in an accounting evaluation of
their financial position. The analysis was carried out by
an international accounting firm. The data were collected
for the same time period as the employee and client
surveys. Overall there were data from all three surveys
from 66 organizations.

The employee survey examined job demands, job con-
trol, social support, emotional exhaustion, and employee
job satisfaction. The client survey examined 14 aspects
of the services received. The financial survey defined
organizational productivity as the number of service
hours delivered as a percentage of the total number of
hours produced by the employees of an organization.
This accounted for overhead differences in the orga-
nizations. In addition, personnel costs per hour were
calculated. Lastly, organizational efficiency was com-
puted based on the costs of one service hour. Then
the cost of one service hour was compared using a
method that benchmarked the cheapest cost as 100%
efficient and related all of the other organizations stan-
dardized to the benchmark. Multiple regression using
blocks of variables to define employee well-being (sat-
isfaction and emotional exhaustion), job characteristics
(demands, control, social support), organizational pro-
ductivity, personnel costs per hour, organizational effi-
ciency, and client satisfaction with services was used to
test relationships.

The findings indicated that employee well-being was
related to client satisfaction with services, organizational
productivity, and personnel costs per hour but not to
organizational efficiency. Job characteristics were not
significantly related to any of the financial indicators or
client satisfaction with services. Employee satisfaction
was positively related to client satisfaction with services
but negatively related to organizational productivity.
Employee emotional exhaustion was positively related
to personnel costs (higher costs) but was negatively
related to client satisfaction (less client satisfaction) and
productivity (less productivity). Emotional exhaustion
was positively related to job demands and control
(higher emotional exhaustion with higher job demands
and control) and negatively related to social support
(higher emotional exhaustion with lower social support).
The results indicated that, the greater the job demands,
the lower the job satisfaction, but the higher the social
support and job control, the greater the job satisfaction.
Generally the results of this large-scale study indicated
that worker emotional well-being in particular and
job satisfaction in some instances had benefits for
client satisfaction with services and financial benefits
for the organization. The happy worker was a more
productive worker.

Fisher et al. (2010) examined employee attitudes,
behaviors, and business performance in a multinational
hotel chain with operations in 50 countries. A ques-
tionnaire survey was used to collect information from
employees in the hotels in Mexico and China as part of
an annual employee survey process. The survey covered
employee demographics, attitudes, and behaviors [role
congruence, communication, leadership, commitment,
job satisfaction, and organizational citizenship behav-
ior (OCB)]. There were 3606 respondents from Mexico
(from four hotels) and 7896 respondents from China
(from four hotels) for an overall employee response rate
of 39.6%. Performance for each hotel was determined by
measures of percentage of annual house profits, revenue
per available room, and guest satisfaction ratings.

The hotels in Mexico were much more profitable
than the hotels in China and had higher scores for
guest satisfaction. There were significant differences

in employee job satisfaction, organizational commitment, and organizational citizenship behavior between employees in Mexico and China with the Mexican employees being higher on all measures. Nonparametric analyses (Spearman's rho) were used to test the relationships among job satisfaction and the financial measures and guest satisfaction. The results showed positive relationships between job satisfaction and the financial measures, but not with guest satisfaction. The same findings were found for organizational commitment. Organizational citizenship behavior was positively related to annual profit but nothing else.

## 3.5 So What?

The evidence from several decades of research indicates that there is support for the belief that happy, satisfied, unstressed workers are more productive, have fewer absences, and are less likely to leave the company; all of these positive impacts on workers benefit the bottom line of an organization in Western cultures. There has been much debate about the truth and/or strength of this relationship, and the debate will continue. With the current worldwide high levels of unemployment there appears to be little incentive for organizations to worry about whether their employees are happy, satisfied, or emotionally well. There is a deep well of job applicants that organizations can draw from to find

replacements for unsatisfied and/or stressed employees. On the other hand, the public health literature indicates that health care costs continue to go up at shocking annual percentage rates in Western countries. This is a significant cost to businesses and to countries that affects commerce, competitiveness, and the long term well-being of countries, companies, and employees.

Smart companies will understand that it is good business to have happy, healthy, and productive employees that offer high-quality goods and services to customers. This makes for a healthy and productive organization.

The next question is, "What can an organization do to be healthy and productive?" Based on a variety of theoretical perspectives and research findings over more than 80 years of research, case studies, and experience, some basic considerations can be reasonably suggested. See Table 1 for a summary of the principles for healthy and productive organizations.

## 4 HEALTHY AND SUSTAINABLE ORGANIZATIONS

In the previous section we presented a description of the attributes of a healthy organization. By a healthy organization we mean that the organization supports people's needs to find meaning, balance, authenticity,

**Table 1 Principles of Healthy and Productive Organizations**

| Principle | Description |
|---|---|
| Importance of human resources | Organizations should treat their human resources at least as well as their other resources (capital, structures, equipment, materials and supplies, products and services, inventories, and customers). |
| Relationship between employee well-being and health and organizational health | Organizations should recognize that the well-being of their workers affects the well-being of their enterprises. |
| Enhancement of work motivation | Organizations should make work (jobs, facilities, supervision, policies) an activity that workers are motivated to engage in effectively and productively. Make jobs such that workers look forward to going to work. |
| Need for continuous assessment and improvement | There are many organizational cultures and management styles that will promote worker well-being and satisfaction. Organizations should monitor the attitudes of workers regularly and make changes in their organizational culture and/or management style and/or job designs as necessary to achieve acceptable levels of worker well-being and satisfaction. |
| Importance of work–family balance | Organizations should remember that work affects almost all aspects of life. It is to an organization's benefit to provide policies and programs to help workers achieve a good work–life balance. |
| Relationship between well-being and productivity | Workers with high emotional well-being will be more productive and less likely to turn over. <br><br> 1. Workers who have economic security (reasonable pay, benefits, job security) will be less stressed and more satisfied. <br> 2. Workers with involvement at the workplace and control over task decisions will be less stressed and more satisfied. <br> 3. Workers with reasonable job demands will be less stressed and more satisfied. <br> 4. Workers who receive social support from peers and supervisors will be less stressed and more satisfied. |
| Importance of trust, respect, and fairness | Organizations need to engage in activities that develop trust, respect and fairness among workers, supervisors, and management. |

and spiritual fulfillment in work and life. Employee burnout is not good for an employee or an organization (Kalimo et al., 1997; Maslach et al., 2001; Smith, 1987), and people need to lead lives they can sustain in the long run if they are to be mentally and physically healthy. A healthy environment engages people's minds and hearts in meeting the needs of the organization. An organization cannot be sustainable unless it fully engages employees and their communities. It follows that a healthy environment is a prerequisite to a sustainable organization.

### 4.1 From Healthy Organizations to Sustainable Organizations

The term *sustainability* became widely known through the publication of *Our Common Future*, also known as the Brundtland Report, from the United Nations World Commission on Environment and Development (WCED, 1987). In the Brundtland Report (p. 24) sustainability is defined as "Development that meets the need of the present without compromising the ability of future generations to meet their own needs." Designing a healthy organization typically involves focusing on internal working conditions (see Table 1 and previous section) whereas designing a sustainable organization requires an additional focus on the environment in which the organization functions, in particular its social environment and the community at large (Delios, 2010; Pfeffer, 2010). This approach to healthy and sustainable organizations fits with the two (unfortunately distinct) approaches proposed by management researchers on business innovation and growth. Ahlstrom (2010) argues that the main objective of business is to develop new and innovative products and services. Disruptive technological innovations can lead to the development of new products and contribute to business growth. According to Ahlstrom (2010), it is through this innovation process that businesses can serve society by creating jobs, generating revenues, and allowing people greater access to cheaper products. This approach requires businesses to focus on their internal conditions and work organization in order to create opportunities for disruptive innovations and to foster entrepreneurial initiatives. This highlights the need for organizations to provide a good working environment for enhancing productivity and fostering innovation. Delios (2010) criticizes the approach proposed by Ahlstrom (2010) because it "ignores the external environment forces on an organization and it ignores the fact that organizations are social entities, populated by real communities of people" (p. 25). He proposes a broader view of business competitiveness that encompasses the issue of corporate social responsibility and the relationship between organizations and their larger social environment. In contrast to the strict transactional view of the relationship between people and businesses (such as Ahlstrom's (2010) approach), we need to understand the larger social role of business organizations. The transactional view focuses on the work that individuals perform in organizations. This is an important element of organizational design and management (see previous section); however, this is insufficient. The relationships between

employees and organizations have become more complex; often organizations play a central role in employees' lives and the communities in which employees live. Therefore, a comprehensive human factors approach to organizational design and management needs to include a focus on both internal organizational conditions for healthy and productive organizations and external conditions for sustainable organizations.

The World Health Organization (WHO) has adopted a similar broad approach to healthy workplaces (WHO, 2010). A renewed commitment to occupational safety and health led to the endorsement of the Workers' Health Global Plan for Action in 2007 (http://apps.who.int/gb/ebwha/pdf_files/WHA60/A60_R26-en.pdf). Out of this renewed effort for improving occupational safety and health, WHO (2010) developed and proposed a healthy workplace model that encompasses four elements:

1. Physical environment or the "structure, air, machinery, furniture, products, chemicals, materials and production processes in the workplace" (p. 9).

2. Psychosocial environment or the "organizational culture as well as attitudes, values, beliefs and daily practices in the enterprise that affect the mental and physical well-being of employees" (p. 10), also known as workplace stressors.

3. Personal health resources in the workplace or the "health services, information, resources, opportunities, flexibility and otherwise supportive environment an enterprise provides to workers to support or motivate their efforts to improve or maintain healthy personal lifestyles, as well as to monitor and support their physical and mental health" (p. 11).

4. Enterprise community involvement or the "activities in which an enterprise might engage, or expertise and resources it might provide, to support the social and physical wellbeing of a community in which it operates" (p. 13).

This model includes both the internal work environment (physical and psychosocial work environment and personal health resources in the workplace) and the linkage between the organization and its environment and the community. This broad approach goes beyond the workplace itself, which has been the target of many human factors efforts. It challenges the human factors professionals and researchers to think about larger social, economic, and environmental problems (Moray, 2000; Smith et al., 1994, 2009).

Pfeffer (2010) proposes to broaden the concept of sustainability by including not only natural and physical resources but also human and social resources. His paper, entitled "Building sustainable organizations: The human factor," clearly calls for organizations to go beyond environmental sustainability and to embrace the concept of "human sustainability." Organizations should, therefore, develop actions and programs aimed at social responsibility as well as invest in management practices that improve employee physical and

psychological well-being. Human sustainability can lead to business benefits, such as brand building and product differentiation (Pfeffer, 2010). Companies that invest in human sustainability may build a positive reputation that can attract additional customers.

## 4.2 Impact of Organizational Decisions on Human Sustainability

Pfeffer (2010) describes three types of decisions made by organizations that can affect human sustainability: (1) employer-funded health insurance, (2) layoffs and downsizing, and (3) work schedules. Employers who provide access to health insurance help their employees to improve their economic well-being as well as health and wellness. However, the empirical evidence for this relationship between employer-funded health insurance and human sustainability is rather weak. O'Brien (2003) identified various effects of offering health insurance on employee performance and human sustainability. First, the productivity of organizations depends on the quality of their employees and by offering health insurance employers can attract high-quality workers. Second, by offering access to health insurance, organizations may be able to retain workers. Third, offering health insurance can increase productivity because healthy workers are more productive than unhealthy workers. Finally, offering health insurance can increase employee satisfaction, and employees who do not have to worry about their own health or the health of their family members may be more productive (O'Brien, 2003). There is, however, limited evidence for these four effects of access to health insurance on employee performance. Buchmueller (2000) reviewed the economical literature to examine whether there are any spill-over effects of providing health insurance to employees. Based on his review he concluded that there is little evidence for the effect of health insurance on worker health and productivity, reduced turnover, or reduced employer costs associated with workers' compensation and absenteeism (Buchmueller, 2000). Some studies have shown that employees in jobs with health insurance coverage change jobs less frequently than workers without health insurance (Madrian, 1994; Monheit and Cooper, 1994), while other studies have shown that offering health insurance has no effect on turnover (Holtz-Eakin, 1994; Kapur, 1998). There is little evidence that having access to health insurance is related to lower turnover. According to O'Brien (2003), the existing literature has not taken into account the effect of ill-health on productivity. There is ample evidence that people with poor health or health conditions work less (Bartel and Taubman, 1979; Rizzo et al., 1998). Therefore, there is room for additional high-quality research to further investigate the relationship between employer-funded health insurance and human sustainability, in particular in the areas of employee satisfaction and well-being.

Decisions made by organizations such as layoffs and downsizing can have a major negative impact on workers affected by those changes as well as a profound impact on the community where the organization is located. A layoff can be considered a major life event and stressor that impacts the individual being laid off,

his/her family, his/her community, and society at large. Workers lose much more than a job when they are laid off. Jobs are an economic necessity and provide many psychological and social benefits. Losing one's job lessens worker feelings of self-worth and dignity. Losing one's job ranks alongside a death in the family with regard to stress because it leaves an emptiness that is difficult to fill (Hansen, 2009). Results of a case study assessing the impact of layoffs in a small community in Texas showed that 58% of study participants reported having increased health problems after the layoff (Virick, 2003). Unemployment has been linked to the following stress-related negative health outcomes: higher mortality rates, increased risk of heart attack, low-birthweight offspring, infectious diseases, chronic respiratory diseases, gastrointestinal disorders, depression, alcoholism, and suicide (Broadhead et al., 1983; Cassell, 1976; Dew et al., 1992; Hamilton et al., 1990; House, 1981; Kivimäki et al., 2003). In an interesting study, Dew et al. (1987) compared the long-term effects of two community-wide stressors: the Three Mile Island nuclear accident and widespread unemployment due to layoff in demographically comparable samples of women. Results of the study showed a remarkable similarity in the stressors' effect: Levels of various mental and physical health symptoms were elevated to similar degrees in both samples one year following the stressor, and symptoms remained elevated in both samples up to three years later. Based on studies on the effects of unemployment in several industrial countries, Brenner (1983, 1987a, 1987b, 1987c) calculated that for every 1% increase in the unemployment rate in the United States (an additional 1.5 million people out of work), an additional 47,000 deaths can be expected, including 26,000 from heart attacks, about 1200 from suicide, 831 murders, and 635 related to alcohol consumption. Results of a study by Benson and Fox (2004) showed that unstable employment increases the risk of intimate partner violence. For couples where the male was always employed, the rate of intimate partner violence was 4.7%. When men experienced one period of unemployment, the rate rose to 7.5%, and when men experienced two or more periods of unemployment, the rate of intimate partner violence rose to 12.3%.

Work hours can lead to work–family conflict if there is insufficient or inadequate time for workers to spend with family and friends (e.g., long work hours, night and week-end work) (Demerouti et al., 2004; Frone and Russell, 1992; Frone, 1997; Grant-Vallone and Donaldson, 2001; Kinnunen et al., 2004).

## 4.3 Sustainable Organizations from the ODAM Viewpoint

Historically, the human factors discipline has focused on the individual employee at the workplace level. Although some human factors research and practice have taken social and economic issues into account, they have typically not paid much attention to the environment (Steimle and Zink, 2006). Recently, macroergonomic approaches have embraced the broader system in which people work, including the environment (Hendrick and Kleiner, 2002). Macroergonomics is

concerned with the optimization of organizational and work system design through consideration of relevant personnel and technological and environmental variables and their interactions (Hendrick and Kleiner, 2001).

Zink and colleagues (2008) have proposed an ODAM approach to designing sustainable organizations that includes three elements: (1) focus on human needs (i.e., human factors and ergonomics), (2) inter- and intragenerational equity (i.e., every generation benefits from the heritage of previous generations and is concerned with the needs of future generations), and (3) three pillars of sustainable development with social, economic, and environmental objectives. Corporate social responsibility is therefore a key element of corporate sustainability: A company is more likely to grow and thrive over time (i.e., be sustained) if the company invests in its workers by providing good working conditions (internal focus) and contributes to its community (external focus).

Corporate sustainability can be defined as "meeting the needs of a firm's direct and indirect stakeholders (such as shareholders, employees, clients, pressure groups, communities, etc.) without compromising its ability to meet the needs of future stakeholders as well" (Dyllick and Hockerts, 2002, p. 131). Three concepts are important in achieving corporate sustainability: (1) a sustainable corporation considers not only economic but also social and environmental prerequisites and impacts of its actions as well as the interdependencies between them, (2) corporate sustainability requires a long-term business orientation as a basis for satisfying stakeholders' needs now and in the future, and (3) a sustainable corporation should generate income from its financial, natural, and social capital without depleting the capital (Dyllick and Hockerts, 2002).

Among a company's stakeholders are its employees, and, consequently, all efforts to achieve corporate sustainability should involve employees. Companies' policies and actions should be aimed at preserving its human capital. For example, many countries have a rapidly aging population, especially in Europe and to a lesser extent in the United states. In these countries, companies should develop policies and implement programs aimed at protecting their human capital, in particular older workers.

A sustainable organization has the flexibility and adaptability to respond readily to both internal and external influences. This might include shifts in the business climate, market opportunities, or the labor market as well as new ideas or challenges that may develop internally. Without full employee engagement, the organization will lack the flexibility and adaptability to respond in a timely manner to changing circumstances. A sustainable organization has employee stability and engagement as well as effective formal and informal communication channels and work processes that enable it to adjust to changing circumstances in an ongoing manner.

## 5 PRINCIPLES FOR WORK SYSTEM DESIGN AND ANALYSIS

In previous sections we described some of the theories and research that have been developed and tested over the past century about how organizations manage their workforce to provide benefits to the organization and the employees. In general, we have moved from command and control to a focus on educating employees and to recognizing the value of engaging employees' hearts and minds in addressing the challenges of the organization (Carayon and Smith, 2000; Hackman, 2002; Lawler, 1986, 1996, 2003; McGregor, 1960; Smith and Carayon-Sainfort, 1989; Smith and Carayon, 1995). To take it a step further, we believe the United States is now in the midst of a major societal shift that has implications for how effective organizations will be in attracting and retaining the most talented and creative people in the workforce. These individuals have high expectations for themselves; enough talent to find alternatives to the traditional career in a large organization; and a desire to find meaning, balance, authenticity, and spiritual fulfillment in their lives. Organizations that are viewed as hampering these human desires will find it difficult to attract and keep the best employees to the long-term detriment of the organization.

### 5.1 Model of Work System

In order to address these challenges, we propose an approach built on systems theory and work complexity. In 1989, Smith and Carayon-Sainfort (1989) proposed the work system model that defines various elements of work and the interactions between the work elements (see Figure 1). Because the work system model is anchored in the discipline of human factors and ergonomics, the person is at the center of the work system: The person has *physical*, *cognitive*, and *psychosocial characteristics and needs* that can influence his/her interactions with the rest of the work system. The person performs *tasks* using various *tools and technologies* in a specific *physical environment*. There are a number of *organizational conditions* that can influence the person and the rest of the work system. From an ODAM viewpoint, it is important to consider the cognitive, physical, and psychosocial needs and characteristics of the individual who is at the center of the work system. This is in line with the International Ergonomics Association (IEA) definition of human factors (or ergonomics) (IEA, 2000): The discipline of human factors (or ergonomics) includes three broad domains of specialization: (1) physical ergonomics, (2) cognitive ergonomics, and (3) organizational ergonomics. The work system model encompasses all three domains of human factors specialization. For instance, relevant topics for physical ergonomics include working postures: Working postures are defined by the interactions between the individual, the tasks, the physical environment, and the design of tools and technologies. In addition, there has been increasing recognition of the importance of psychosocial work factors as determinants of physical stress: A worker who is under psychosocial stress (e.g., time pressure) may increase his or her work pace, thus increasing the likelihood that awkward postures may lead to physical stress and health problems. Topics relevant to cognitive ergonomics include mental workload, which has been conceptualized as resulting from the lack of fit between task demands and individual
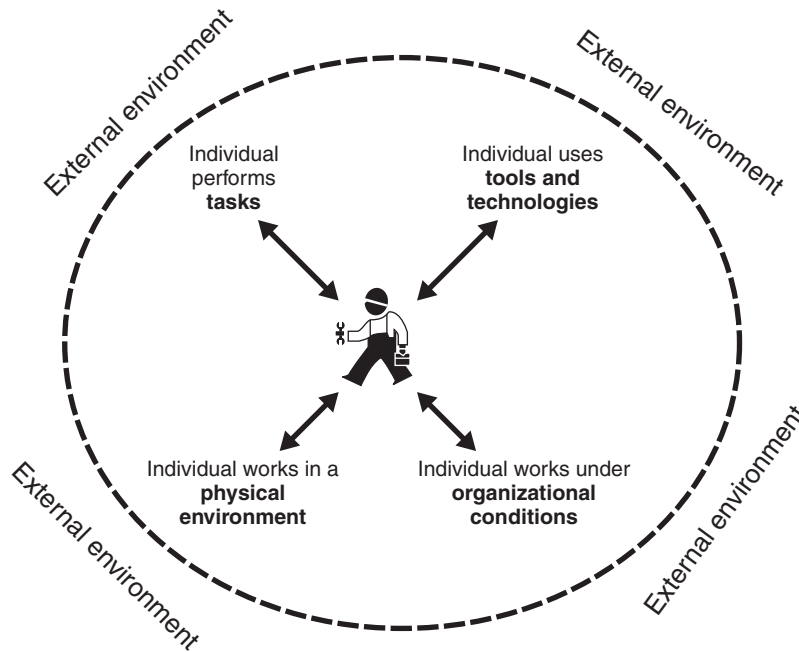
**Figure 1**  Elements of the work system and its external environment. (Adapted from Smith and Carayon-Sainfort, 1989).

resources and capabilities. However, it has become clear that mental workload is also influenced by larger organizational issues that can affect specific task demands (MacDonald, 2003). For instance, an organization may restrict the number of rest breaks taken by workers and increase the time pressure put on workers; this will definitely increase the task demands and will likely lead to increased mental workload. Therefore, from a human factors perspective, it is important to consider all elements of the work system and their interactions. The work system should be designed to accommodate the physical, cognitive, and psychosocial characteristics of the individual and meet their physical, cognitive, and psychosocial needs.

The work system model allows the consideration of all three groups of relevant human factors topics, that is, physical, cognitive, and organizational ergonomic issues. The work system model, however, is limited as it describes the work elements of an individual. This conceptualization of the work system for looking at single individuals has been recently expanded to consider how to design an organization. Carayon and Smith (2000) describe how an organization can be conceptualized as a collection of multiple work systems: The organization is comprised of individual employees with their own work systems. The work systems of the individual employees interact with each other; therefore, it is important to consider the interactions and interfaces between the different work systems. One way of designing the interactions between the work systems is to consider *organizational processes.* Any organization has multiple processes, including production processes, design processes, and support processes (e.g., human resources, supply chain

management). A process consists of a series of tasks that are temporally interdependent and that transform a range of inputs into an output. The connections between work systems have been examined for patient care processes, such as the outpatient surgery process (Carayon, 2009; Schultz et al., 2007). The work system elements (individual, task, tools/technologies, physical environment, and organizational conditions) can be used to describe the physical, cognitive, and organizational ergonomic issues related to care processes (Carayon et al., 2004; Schultz et al., 2007). Therefore, a care process can be considered as a series of tasks performed by various individuals; each task of the process is performed by one or several individuals who use multiple tools and technologies. The process tasks are performed in a physical environment. The organizational conditions of importance to process design include communication and coordination across tasks. In particular, in care processes, transitions of care across health care providers or settings can produce a range of communication and coordination issues related to information flow and care accountability (Carayon et al., 2004).

Our previous discussion of sustainable organizations emphasizes the need to consider the larger environment (including the social environment) in which organizations evolve. Therefore, a useful expansion of the work system model is to consider the larger environment in which the work system functions (Kleiner, 2008). According to the sociotechnical systems theory (Cherns, 1987; Clegg, 2000; Pasmore, 1988), there are two-way interactions between the system and its environment: (1) the environment influences the work system and (2) the work system influences the environment.

Figure 1 represents an expansion of the work system model that includes the environment; the dashed line between the work system and the environment implies that there are two-way interactions between the system and its environment. The external environment is comprised of the physical environment, the social environment, and the legal/regulatory/political/professional environment. Other human factors approaches have also considered the role of the larger environment in influencing work and workers. For instance, Rasmussen (2000) proposed a hierarchy of system levels that interact and influence each other. Moray (2000) has proposed a similar hierarchical approach to ergonomic system design.

A range of macroergonomic methods are available to analyze a work system, including both qualitative (e.g., interview and focus group) and quantitative (e.g., survey) methods. For additional information, see the section on macroergonomic methods in the *Handbook of Human Factors and Ergonomics Methods* by Stanton and colleagues (2004).

## 5.2 Continuous Improvement of Work System Design

Practitioners and researchers in organizational design and management have long recognized the importance of the continuous improvement process: The design of work systems should be a continuous process. For instance, when revisiting the principles of sociotechnical system (STS) design, Clegg (2000) emphasizes that the STS design process extends over time. New business and employee needs may arise, requiring the work system design to be reassessed and adapted. The dynamic nature of the work system and its environment clearly calls for a design approach based on continuous improvement. Carayon (2006) has discussed the need for continuous cycles of work system design, implementation, and continuous improvement/adaptation. Her principles for continuous work system improvement are summarized in Table 2.

**Table 2 Principles for Macroergonomic Continuous System Adaptation and Improvement**

| Principle | Description |
| --- | --- |
| Participation | Active participation of all stakeholders in system design activities (e.g., participatory ergonomics) |
| Interactions | Continuous interactions between multiple work systems and between work systems and their environment |
| Design | Continuous work system design and redesign |
| Adaptation | Adaptation of work system for long-term health, productivity, and sustainability |
| Learning | Support for both individual and organizational learning (e.g., collaborative problem definition, analysis, and modeling) |
| Sense making | Sense making of on-going changes and their impact |

*Source*: From Carayon, 2006.

The field of quality improvement and Total quality management has had a profound influence on organizational activities; many businesses routinely employ the PDCA, or plan−do−check−act, cycle (Deming, 1986). This continuous cycle of planning, implementation, assessment, and improvement can be recognized in the WHO (2010) model of healthy workplace continual improvement process. The WHO model involves the following eight steps:

1. Mobilization of employers, managers, and workers for work system change. This step requires a deep understanding of the needs, values, and priority issues of various members of the organization in order to identify the important issue(s) that will mobilize them.

2. Creation of a multidisciplinary team to work on implementing a work system change. The team should be provided with adequate resources to achieve its objectives. Professionals in the areas of occupational health and safety, human resources and engineering should be involved in this team as well as representatives from the employer and the employees.

3. Assessment of work system and employee and organizational health and performance. This will typically involve the use of various tools and methods to assess, for instance, worker health, workplace hazards, and turnover and productivity.

4. Priority setting for work system change. Several criteria are likely to be used to identify the work system change to be implemented, including limiting exposure to physical or psychosocial hazards, ease of implementing change, and likelihood of success.

5. Planning for implementation of work system change. The plan may be very simple or complex depending on the work system change and the size and complexity of the organization. Each action of the plan should have clear objectives and be assigned to specific members of the organization.

6. Implementation of work system change.

7. Evaluation of work system change. It is important to evaluate the pluses and minuses of the work system change, in particular with regard to the initial objectives. The evaluation should also include an evaluation of the planning phase and the implementation process.

8. Continuous improvement. This last step is actually the first step of the next cycle of work system changes. Additional changes may be necessary based on the evaluation results.

## 6 CONCLUSION

In this chapter, we have conducted an extensive historical review of various theories and approaches to organizational design and management. In the past two

decades, the human factors and ergonomics discipline has recognized the importance of organizational design and management; this concept is known as macroergonomics (Hendrick, 1991). Human factors researchers and professionals recognize the need to consider organizational and sociotechnical issues when designing work systems. For instance, when attempting to eliminate physical stressors such as awkward postures, ergonomists understand the need to consider organizational policies such as work schedules and rest−break schedules that can influence or mitigate the impact of awkward postures on workers. Other interactions between micro- and macroergonomic issues are discussed in this chapter and have been reviewed by others (Zink, 2000).

Our review of the literature also shows the need for human factors researchers and professionals to be aware of theoretical developments in connected fields and disciplines, in particular psychology, sociology, business, and occupational and public health. This need for multidisciplinary approaches to work system design aimed at enhancing performance, safety, quality of working life, and well-being has been discussed by other human factors researchers (Carayon, 2006; Moray, 2000; Rasmussen, 2000). As the world has become flat (Friedman, 2005), problems have become increasingly complex and multidimensional, requiring expertise in multiple areas and disciplines. Therefore, human factors researchers and professionals should be encouraged to team up with relevant domain experts in analyzing and improving system design.

The discipline of human factors and ergonomics has grown significantly in the past 50 years; this has led to specialization of human factors professionals and ergonomists in specific domains of human factors and ergonomics (Wilson, 2000). This specialization may have come at the expense of a true system design approach that recognizes interactions between system levels and the various levels of system design (Carayon, 2009; Karsh and Brown, 2010; Rasmussen, 2000). This chapter on human factors in organizational design and management proposes some directions for the human factors and ergononomics discipline to examine the larger organizational system as well as the environment in which organizations function and evolve. This should heighten the awareness of the organization and its environment among human factors professionals and researchers. We recommend that human factors professionals and researchers develop concepts, theories, and methods for building healthy and sustainable organizations.

## REFERENCES

Ahlstrom, D. (2010), "Innovation and Growth: How Business Contributes to Society," *Academy of Management Perspectives*, Vol. 24, No. 3, pp. 11–24.

Bartel, A. P., and Taubman, P. (1979), "Health and Labor Market Success: The Role of Various Diseases," *Review of Economics and Statistics*, Vol. 61, pp. 1–8.

Benson, M. L., and Fox, G. L. (2004), *Concentrated Disadvantage, Economic Distress, and Violence against Women in Intimate Relationships*, U.S. Department of Justice, Rockville, MD.

Black, J. S., and Porter, L. W. (2000), *Management: Meeting New Challenges*, Prentice-Hall, Upper Saddle River, NJ.

Brenner, M. H. (1983), "Mortality and Economic Instability: Detailed Analyses for Britain and Comparative Analyses for Selected Industrialized Countries," *International Journal of Health Services*, Vol. 13, No. 4, pp. 563–620.

Brenner, M. H. (1987a), "Economic Change, Alcohol Consumption and Heart Disease Mortality in Nine Industrialized Countries," *Social Science and Medicine*, Vol. 25, pp. 119–132.

Brenner, M. H. (1987b), "Economic Instability, Unemployment Rates, Behavioral Risks, and Mortality Rates in Scotland, 1952–1983," *International Journal of Health Services*, Vol. 17, No. 3, pp. 475–487.

Brenner, M. H. (1987c), "Relation of Economic Change to Swedish Health and Social Well-Being, 1950–1980," *Social Science and Medicine*, Vol. 25, No. 2, pp. 183–195.

Broadhead, W. E., et al. (1983), "The Epidemiological Evidence for a Relationship Between Social Support and Health," *American Journal of Epidemiology*, Vol. 117, pp. 521–537.

Buchmueller, T. C. (2000), *The Business Case for Employer-Provided Health Benefits: A Review of the Relevant Literature*, California HealthCare Foundation, Oakland, CA.

Caplan, R. D., Cobb, S., French, J. R. P., Harrison, R. V., and Pineau, S. R. (1975), *Job Demands and Worker Health*, U.S. Government Printing Office, Washington, DC.

Carayon, P. (2006), "Human Factors of Complex Sociotechnical Systems," *Applied Ergonomics*, Vol. 37, pp. 525–535.

Carayon, P. (2009), "The Balance Theory and the Work System Model . . . Twenty Years Later," *International Journal of Human-Computer Interaction*, Vol. 25, No. 5, pp. 313–327.

Carayon, P., and Smith, M. J. (2000), "Work Organization and Ergonomics," *Applied Ergonomics*, Vol. 31, pp. 649–662.

Carayon, P., Sainfort, F., and Smith, M. J. (1999), "Macroergonomics and Total Quality Management: How to Improve Quality of Working Life?" *International Journal of Occupational Safety and Ergonomics*, Vol. 5, No. 2, pp. 303–334.

Carayon, P., Schultz, K., and Hundt, A. S. (2004), "Righting Wrong Site Surgery," *Joint Commission Journal on Quality and Safety*, Vol. 30, No. 7, pp. 405–410.

Cassell, J. (1976), "The Contribution of the Social Environment to Host Resistance," *American Journal of Epidemiology*, Vol. 102, pp. 107–123.

Cherns, A. (1987), "Principles of Sociotechnical Design Revisited," *Human Relations*, Vol. 40, pp. 153–162.

Clegg, C. W. (2000), "Sociotechnical Principles for System Design," *Applied Ergonomics*, Vol. 31, pp. 463–477.

Cooper, C. L., and Marshall, J. (1976), "Occupational Sources of Stress: A Review of the Literature Relating to Coronary Heart Disease and Mental Ill Health," *Journal of Occupational Psychology*, Vol. 49, No. 1, pp. 11–25.

Cummings, T. G., Molloy, E. S., and Glen, R. H. (1975), "Intervention Strategies for Improving Productivity and the Quality of Work Life," *Organizational Dynamics*, Vol. 4, No. 1, pp. 52–68.

Delios, A. (2010), "How Can Organizations Be Competitive But Dare to Care?" *Academy of Management Perspectives*, Vol. 24, No. 3, pp. 25–36.

Demerouti, E., Bakker, A. B., and Bulters, A. (2004), "The Loss Spiral of Work Pressure, Work-Home Interface and Exhaustion: Reciprocal Relations in a Three-Wave Study," *Journal of Vocational Behavior*, Vol. 64, pp. 131–149.

Deming, W. E. (1986), *Out of the Crisis*, Massachussets Institute of Technology, Cambridge, MA.

Dew, M. A., Bromet, E. J., and Schulberg, H. C. (1987), "A Comparative Analysis of Two Community Stressors' Long-Term Mental Health Effects," *Americn Journal of Community Psychology*, Vol. 15, No. 2, pp. 167–184.

Dew, M. A., Bromet, E. J., and Penkower, L. (1992), "Mental Health Effects of Job Loss in Women," *Psychological Medicine*, Vol. 22, No. 3, pp. 751–764.

Drucker, P. F. (2001), *The Essential Drucker: Selections from the Management Works of Peter F. Drucker*, Harper-Collins, New York.

Dyllick, T., and Hockerts, K. (2002), "Beyond the Business Case for Corporate Sustainability," *Business Strategy and the Environment*, Vol. 11, No. 2, pp. 130–141.

Fisher, R., McPhail, R., and Menghetti, G. (2010), "Linking Employee Attitudes and Behaviors with Business Performance: A Comparative Analysis of Hotels in Mexico and China," *International Journal of Hospitality Management*, Vol. 29, pp. 397–404.

Frankenhaeuser, M., and Gardell, B. (1976), "Underload and Overload in Working Life. A Multidisciplinary Approach," *Journal of Human Stress*, Vol. 2, pp. 35–46.

French, J. R. P. J. (1963), "The Social Environment and Mental Health," *Journal of Social Issues*, Vol. 19, pp. 39–56.

Friedman, T. L. (2005), *The World Is Flat: A Brief History of the Twenty-First Century*, Farrar, Straus and Giroux, New York.

Frone, M. J., and Russell, M. (1992), "Antecedents and Outcomes of Work-Family Conflict: Testing a Model of the Work-Family Interface," *Journal of Applied Psychology*, Vol. 77, No. 1, pp. 65–78.

Frone, M. R. (1997), "Relation of Work-Family Conflict to Health Outcomes: A Four-Year Longitudinal Study of Employed Parents," *Jounal of Occupational and Organizational Psychology*, Vol. 70, pp. 325–335.

Gardell, B. (1981a), "Psychosocial Aspects of Industrial Production Methods," in *Society, Stress and Disease*, Vol. 4, L. Levi, Ed., Oxford University Press, London.

Gardell, B. (1981b), "Strategies for Work Reform Programmes on Work Organization and Work Environment," in *Working Life. A Social Science Contribution to Reform*, B. Gardell and G. Johansson, Eds., Wiley, Chichester.

Grant-Vallone, E. J., and Donaldson, S. I. (2001), "Consequences of Work-Family Conflict on Employee Well-Being Over Time," *Work and Stress*, Vol. 15, No. 3, pp. 214–226.

Grawitch, M. J., Gottschalk, M., and Munz, D. C. (2006), "The Path to a Healthy Workplace: A Critical Review Linking Healthy Workplace Practices, Employee Well-Being, and Organizational Improvements," *Consulting Psychology Journal: Practice and Research*, Vol. 58, No. 3, pp. 129–147.

Hackman, J. R. (2002), *Leading Teams: Setting the Stage for Great Performances*, Harvard Business School Press, Cambridge, MA.

Haims, M. C., and Carayon, P. (1998), "Theory and Practice for the Implementation of 'In-House', Continuous Improvement Participatory Ergonomic Programs," *Applied Ergonomics*, Vol. 29, No. 6, pp. 461–472.

Hamilton, V. L., Broman, C. L., Hoffman, W. S., and Renner, D. S. (1990), "Hard Times and Vulnerable People: Initial Effects of Plant Closing on Autoworkers' Mental Health," *Journal of Health and Social Behavior*, Vol. 31, No. 2, pp. 123–140.

Hansen, G. B. (2009), *A Guide to Worker Displacement: Some Tools for Reducing the Impact on Workers, Communities and Enterprises*, Skills and Employability Department, International Labor Office, Geneva, Switzerland.

Harter, J. K., Schmidt, F. L., and Keyes, C. L. (2002), "Well-Being in the Workplace and Its Relationship to Business Outcomes: A Review of the Gallup Studies," in *Flourishing: The Positive Person and the Good Life*, C. L. K. a. J. Haidt, Ed., American Psychological Association, Washington, DC, pp. 205–224.

Hendrick, H. W. (1986), "Macroergonomics: A Conceptual Model for Integrating Human Factors with Organizational Design," in *Human Factors in Organizational Design and Management*, Vol. 2, O. J. Brown and H. W. Hendrick, Eds., North Holland. Amsterdam, The Netherlands, pp. 467–478.

Hendrick, H. W. (1987), "Organizational Design," in *Handbook of Human Factors*, G. Salvendy, Ed., Wiley, New York, pp. 470–494.

Hendrick, H. W. (1991), "Human Factors in Organizational Design and Management," *Ergonomics*, Vol. 34, pp. 743–756.

Hendrick, H. W., and Kleiner, B. M. (2001), *Macroergonomics—An Introduction to Work System Design*, The Human Factors and Ergonomics Society, Santa Monica, CA.

Hendrick, H. W., and Kleiner, B. M. (2002), *Macro-Ergonomics: Theory, Methods, and Apllications*, Lawrence Erlbaum, Mahwah, NJ.

Herzberg, F. (1974, September/October), "The Wise Old Turk," *Harvard Business Review*, Vol. 52, pp. 70–80.

Holtz-Eakin, D. (1994), "Health Insurance Provision and Labor Market Efficiency in the United States and Germany," in *Social Protection versus Economic Flexibility: Is There a Trade-off?* R. M. Blank, Ed., University of Chicago Press, Chicago, IL.

House, J. (1981), *Work, Stress and Social Support*, Addison-Wesley, Reading, MA.

House, J. S. (1981), *Work Stress and Social Support*, Addison-Wesley, Reading, MA:

International Ergonomics Association (IEA) (2000), "The Discipline of Ergonomics," retrieved from www.iea.cc, August 22, 2004.

Kalimo, R., Lindstrom, K., and Smith, M. J. (1997), "Psychosocial Approach in Occupational Health," in *Handbook of Human Factors and Ergonomics*, G. Salvendy, Ed., Wiley, New York, pp. 1059–1084.

Kapur, K. (1998), "The Impact of Health on Job Mobility: A Measure of Job Lock," *Industrial and Labor Relations Review*, Vol. 51, pp. 282–297.

Karasek, R. A. (1979), "Job Demands, Job Decision Latitude, and Mental Strain: Implications for Job Redesign," *Administrative Science Quarterly*, Vol. 24, pp. 285–308.

Karsh, B.-T., and Brown, R. (2010), "Macroergonomics and Patient Safety: The Impact of Levels on Theory, Measurement, Analysis and Intervention in Patient Safety Research," *Applied Ergonomics*, Vol. 41, No. 5, pp. 674–681.

Katzell, R. A., Yankelovich, D., Fein, M., Ornati, O. A., and Nash, A. (1975), "Improving Productivity and Job Satisfaction," *Organizational Dynamics*, Vol. 4, No. 1, pp. 69–80.

Kinnunen, U., Geurts, S., and Mauno, S. (2004), "Work-to-Family Conflict and Its Relationship with Satisfaction and Well-Being: A One-Year Longitudinal Study on Gender Differences," *Work and Stress*, Vol. 18, No. 1, pp. 1–22.

Kivimäki, M., Vahtera, J., Elovainio, M., Pentti, J., and Virtanen, M. (2003), "Human Costs of Organizational Downsizing: Comparing Health Trends between Leavers and Stayers," *American Journal of Community Psychology*, Vol. 32, No. 1, pp. 57–67.

Kleiner, B. M. (2008), "Macroegonomics: Work System Analysis and Design," *Human Factors*, Vol. 50, No. 3, pp. 461–467.

Lawler, E. E., III (1986), *High Involvement Management: Participative Strategies for Improving Organizational Performance*, Jossey-Bass, San Francisco.

Lawler, E. E., III (2003), *Treat People Right: How Organizations and Individuals Can Propel Each Other into a Virtuous Spiral of Success*, Wiley, New York.

Lawler, E. E. I. (1996), *From the Ground Up: Six Principles for Building the New Logic Corporation*, Wiley (Jossey-Bass), New York.

MacDonald, W. (2003), "The Impact of Job Demands and Workload on Stress and Fatigue," *Australian Psychologist*, Vol. 38, No. 2, pp. 102–117.

Madrian, B. C. (1994), "Employment-Based Health Insurance and Job Mobility: Is There Evidence of Job-Lock?" *Quarterly Journal of Economics*, Vol. 109, pp. 27–54.

Maslach, C., Schaufeli, W. B., and Leiter, M. P. (2001), "Job Burnout," *Annual Review of Psychology*, Vol. 52, pp. 397–422.

McGregor, D. (1960), *The Human Side of Enterprise*, McGraw-Hill, New York.

Monheit, A. C., and Cooper, P. E. (1994), "Health Insurance and Job Mobility: Theory and Evidence," *Industrial and Labor Relations Review*, Vol. 46, No. 1, pp. 68–85.

Moray, N. (2000), "Culture, Politics and Ergonomics," *Ergonomics*, Vol. 43, No. 7, pp. 858–868.

Noro, K., and Imada, A. (1991), *Participatory Ergonomics*, Taylor and Francis, London.

O'Brien, E. (2003), "Employers' Benefits from Workers' Health Insurance," *The Milbank Quarterly*, Vol. 81, No. 1, pp. 5–43.

Parker, S., and Wall, T. (1998), *Job and Work Design*, Sage Publications, Thousand Oaks, CA.

Pasmore, W. A. (1988), *Designing Effective Organizations: The Sociotechnical Systems Perspective*, Wiley, New York.

Pfeffer, J. (2010), Building Sustainable Organizations: The Human Factor. *Academy of Management Perspectives*, Vol. 24, No. 1, pp. 34–45.

Porter, L. W., Steers, R., and Bigley, G. (2002), *Motivation and Work Behavior*, 7th ed., McGraw-Hill, New York.

Quick, J. C., and Quick, J. D. (2004), "Healthy, Happy, Productive Work: A Leadership Challenge," *Organizational Dynamics*, Vol. 33, No. 4, pp. 429–337.

Rasmussen, J. (2000), "Human Factors in a Dynamic Information Society: Where Are We Heading?" *Ergonomics*, Vol. 43, No. 7, pp. 869–879.

Rizzo, J. A., Abbott, I., T.A., and Berger, M. L. (1998), "The Labor Productivity Effects of Chronic Backache in the United States," *Medical Care*, Vol. 36, pp. 1471–1488.

Roethlisberger, F. J., and Dickson, W. J. (1942), *Management and the Worker: An Account of a Research Program Conducted by the Western Electric Company, Hawthorne works, Chicago, Il*, Harvard University Press, Cambridge, MA.

Sainfort, F., Taveira, A. D., Arora, N., and Smith, M. J. (2001), "Teams and Team Management and Leadership," in *Handbook of Industrial Engineering*, G. Salvendy, Ed., Wiley, New York, pp. 975–994.

Salas, E., Stagl, K. C., Burke, C. S., and Goodwin, G. F. (2004), "Fostering Team Effectiveness in Organizations: Toward an Integrative Theoretical Framework of Team Performance," in *Modeling Complex Systems: Motivation, Cognition, and Social Processes*, W. Spaulding and J. Flowers, Eds., University of Nebraska Press, Lincoln, NE.

Schultz, K., Carayon, P., Hundt, A. S., and Springman, S. (2007), "Care Transitions In the Outpatient Surgery Preoperative Process: Facilitators and Obstacles to Information Flow and Their Consequences," *Cognition, Technology and Work*, Vol. 9, No. 4, pp. 219–231.

Selye, H. (1950), *Stress*, Acta, Montreal.

Smith, K. U. (1965), *Behavior Organization and Work*, rev. ed., College Printing and Typing Co., Madison, WI.

Smith, M. J. (1987), "Occupational Stress," in *Handbook of Human Factors and Ergonomics*, G. Salvendy, Ed., Wiley, New York, pp. 844–860.

Smith, M. J., and Carayon, P. (1995), "New Technology, Automation, and Work Organization: Stress Problems and Improved Technology Implementation Strategies," *International Journal of Human Factors in Manufacturing*, Vol. 5, No. 1, pp. 99–116.

Smith, M. J., and Carayon-Sainfort, P. (1989), "A Balance Theory of Job Design for Stress Reduction," *International Journal of Industrial Ergonomics*, Vol. 4, No. 1, pp. 67–79.

Smith, M. J., Carayon, P., Smith, J., Cohen, W., and Upton, J. (1994), "Community Ergonomics: A Theoretical Model for Rebuilding the Inner City," in *The Human Factors and Ergonomics Society 38th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 724–728.

Smith, M. J., Derjani-Bayeh, A., and Carayon, P. (2009), "Community Ergonomics and Globalization: A Conceptual Model of Social Awareness," in *Industrial Engineering and Ergonomics*, C. M. Schlick, Ed., Springer, Heidelberg, Germany, pp. 57–66.

Stanton, N., Hedge, A., Brookhuis, K., Salas, E., and Hendrick, H. W., Eds. (2004), *Handbook of Human Factors and Ergonomics Methods*, CRC Press, Boca Raton, FL.

Steimle, U., and Zink, K. J. (2006), "Sustainable Development and Human Factors," in *International Encyclopedia of Ergonomics and Human Factors*, 2nd ed., W. Karwowski, Ed., Taylor and Francis, London, pp. 2258–2263.

Sundstrom, E., DeMeuse, K. P., and Futrell, D. (1990), "Work Teams: Applications and Effectiveness," *American Psychologist*, Vol. 45, No. 2, pp. 120–133.

Taris, T. W., and Schreurs, P. J. G. (2009), "Well-Being and Organizational Performance: An Organizational-Level Test of the Happy-Productive Worker Hypothesis," *Work and Stress*, Vol. 23, No. 2, pp. 120–136.

Taylor, F. (1911), *The Principles of Scientific Management*, Norton and Company, New York.

Virick, M. (2003), *Research Report: The Effect of Layoffs in the North Texas Region*, Center for Research on Organizational and Managerial Excellence, University of Texas, Arlington, TX.

World Commission on Environment and Development (WCED) (1987), *Our Common Future*, Oxford University Press, Oxford:.

Wilson, J. R. (2000), "Fundamentals of Ergonomics in Theory and Practice," *Applied Ergonomics*, Vol. 31, No. 6, pp. 557–567.

Wilson, J. R., and Haines, H. M. (1997), "Participatory Ergonomics," in *Handbook of Human Factors and Ergonomics*, G. Salvendy, Ed., Wiley, New York, pp. 490–513.

World Health Organization (WHO) (2010), *Healthy Workplaces: A Model for Action*, WHO, Geneva.

Wright, T. A., and Cropanzano, R. (2004), "The Role of Psychological Well-Being in Job Performance: A Fresh Look at an Age-Old Quest," *Organizational Dynamics*, Vol. 33, No. 4, pp. 338–351.

Zink, K. (2000), "Ergonomics in the Past and the Future: From a German Perspective to an International One," *Ergonomics*, Vol. 43, No. 7, pp. 920–930.

Zink, K. J., Steimle, U., and Fisher, K. (2008), "Human Factors, Business Excellence and Corporate Sustainability: Differing Perspectives, Joint Objectives," in *Corporate Sustainability as a Challenge for Comprehensive Management*, K. J. Zink, Ed., Physica-Verlag, Heidelberg, Germany, pp. 3–18.

# CHAPTER 19

# SITUATION AWARENESS

**Mica R. Endsley**
**SA Technologies, Inc.**
**Marietta, Georgia**

## 1 INTRODUCTION

As we move into the twenty-first century, the biggest challenge within most industries and the most likely cause of an accident receives the label of human error. This is a most misleading term, however, that has done much to sweep the real problems under the rug. It implies that people are merely careless or poorly trained or somehow not very reliable in general. In fact, in the vast majority of these accidents the human operator was striving against significant challenges. On a day-to-day basis, they cope with hugely demanding complex systems. They face both data overload and the challenge of working with a complex system. They are drilled with long lists of procedures and checklists designed to cope with some of these difficulties, but from time to time they are apt to fail. Industry's typical response to such failures has been more procedures and more systems, but unfortunately, this only adds to the complexity of the system. In reality, the person is not the cause of these errors but is the final dumping ground for the inherent problems and difficulties in the technologies we have created. The operator is usually the one who must bring it all together and overcome whatever failures and inefficiencies exist in the system.

So why are people having trouble coping with the present technology and data explosion? The answer lies in understanding how people process the vast amount of data around them to arrive at effective performance.

If these accidents are examined in detail, one finds that the operators generally have no difficulty in performing their tasks physically and no difficulty in knowing what is the correct thing to do, but they continue to be stressed by the task of *understanding what is going on in the situation*. Developing and maintaining a high level of *situation awareness* are the most difficult parts of many jobs and some of the most critical and challenging tasks in many domains today.

*Situation awareness* (SA) can be thought of as an internalized mental model of the current state of the operator's environment. All of the incoming data from the many systems, the outside environment, fellow team members, and others [e.g., other aircraft and air traffic control (ATC)] must all be brought together into an integrated whole. This integrated picture forms the central organizing feature from which all decision making and action take place (Figure 1).

A vast portion of the operator's job is involved in developing SA and keeping it up to date in a rapidly changing environment. This is a task that is not simple in light of the complexity and sheer number of factors that must be taken into account to make effective decisions. The key to coping in the information age is developing systems that support this process, yet this is where current technologies have left human operators the most vulnerable to error. Problems with SA were found to be the leading causal factor in a review of military aviation
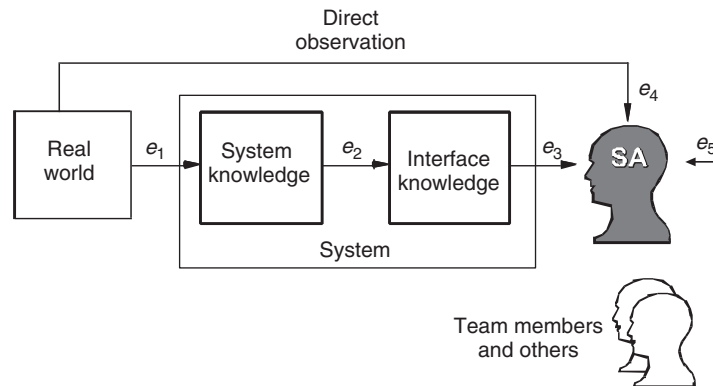
**Figure 1**   Sources of SA. (From Endsley, 1995d, 1997.)

mishaps (Hartel et al., 1991) and in a study of accidents among major air carriers; 88% of those involving human error could be attributed to problems with SA (Endsley, 1995c).

A similar review of errors in other domains, such as air traffic control (Rodgers et al., 2000) or nuclear power (Hogg et al., 1993; Mumaw et al., 1993), showed that this is not a problem limited to aviation but one faced by many complex systems.

Successful system designs must deal with the challenge of combining and presenting vast amounts of data now available from many technological systems in order to provide true SA (whether it is to a pilot, a physician, a business manager, or an automobile driver). An important key to the development of complex technologies is understanding that true SA exists only in the mind of the human operator. Therefore, presenting a ton of data will do no good unless the data are transmitted, absorbed, and assimilated successfully and in a timely manner by the human in order to form SA. Unfortunately, most systems fail in this regard, leaving significant SA problems in their wake (Figure 2).

## 1.1   Definition of Situation Awareness

Although much SA research (and the term) originated within the aviation domain, SA as a construct is widely studied and exists as a basis of performance across many different domains, including air traffic control,

military operations, education, driving, train dispatching, maintenance, and weather forecasting. One of the earliest and most widely applicable SA definitions describes it as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" (Endsley, 1988). SA therefore involves perceiving critical factors in the environment (level 1 SA), understanding what those factors mean, particularly when integrated together in relation to the operator's goals (level 2), and at the highest level an understanding of what will happen with the system in the near future (level 3). These higher levels of SA allow people to function in a timely and effective manner, even with very complex and challenging tasks. Each of these levels will be discussed in more detail.

### 1.1.1   Level 1: Perception of the Elements in the Environment

The first step in achieving SA is to perceive the status, attributes, and dynamics of relevant elements in the environment. A pilot needs to perceive important elements such as other aircraft, terrain, system status, and warning lights along with their relevant characteristics. In the cockpit, just keeping up with all of the relevant system and flight data, other aircraft, and navigational data can be quite taxing. An army officer needs to detect enemy, civilian, and friendly positions and actions, terrain features, obstacles, and weather. An air traffic controller or automobile driver has a different set of information that is needed for SA.

### 1.1.2   Level 2: Comprehension of the Current Situation

Comprehension of the situation is based on a synthesis of disjoint level 1 elements. Level 2 SA goes beyond simply being aware of the elements that are present to include an understanding of the significance of those elements in light of one's goals. The operators put together level 1 data to form a holistic picture of the environment, including a comprehension of the significance of objects and events. For example, upon seeing warning lights indicating a problem during takeoff, the pilot
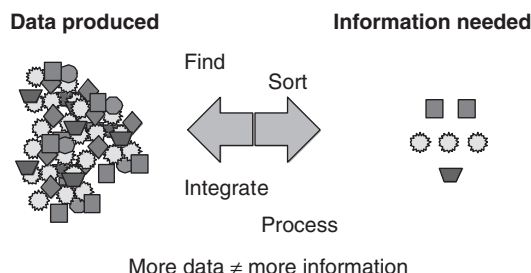


**Figure 2**   Information gap. (From Endsley, 2000b.)

must quickly determine the seriousness of the problem in terms of the immediate air worthiness of the aircraft and combine this with knowledge on the amount of runway remaining in order to know whether or not it is an abort situation. A novice operator may be capable of achieving the same level 1 SA as more experienced ones but may fall far short of being able to integrate various data elements along with pertinent goals in order to comprehend the situation.

### 1.1.3 Level 3: Projection of the Future Status

It is the ability to project the future actions of the elements in the environment, at least in the very near term, that forms the third and highest level of SA. This is achieved through knowledge of the status and dynamics of the elements and a comprehension of the situation (both levels 1 and 2 SA). Amalberti and Deblon (1992) found that a significant portion of experienced pilots' time was spent in anticipating possible future occurrences. This gives them the knowledge (and time) necessary to decide on the most favorable course of action to meet their objectives. This ability to project can similarly be critical in many other domains, including driving, plant control, and sports.

### 1.2 Elements of Situation Awareness

The "elements" of SA in the definition are very domain specific. Examples for air traffic control are shown in Table 1. These elements are clearly observable, meaningful pieces of information for an air traffic controller. Things such as aircraft type, altitude, heading, and flight plan, and restrictions in effect at an airport or conformance to a clearance each comprise meaningful elements of the situation for an air traffic controller. The elements that are relevant for SA in other domains can be delineated similarly. Cognitive task analyses have been conducted to determine SA requirements in commercial aviation (Farley et al., 2000), fighter aircraft (Endsley, 1993), bomber aircraft (Endsley, 1989b), and infantry operations (Matthews et al., 2004), among others.

## 2 DEVELOPING SITUATION AWARENESS

Several researchers have developed theoretical formulations for depicting the role of numerous cognitive processes and constructs on SA (Endsley, 1988, 1995d; Fracker, 1988; Taylor, 1990; Tenney et al., 1992; Taylor and Selcon, 1994; Adams et al., 1995; Smith and Hancock, 1995). There are many commonalties in these efforts, pointing to essential mechanisms that are important for SA. The key points are discussed here; however, more details on each model may be found in these readings. Reviews of these theoretical models of SA are also provided in Pew (1995), Durso and Gronlund (1999), and Endsley (2000b).

Endsley (1988, 1990b, 1995d) describes a theoretical framework model of SA which is summarized in Figure 3. In combination, the mechanisms of short-term sensory memory, perception, working memory, and long-term memory form the basic structures on which

**Table 1 Elements of SA for Air Traffic Control**

#### Level 1 SA

Aircraft
  Aircraft identification (ID), combat identification (CID), beacon code
  Current route (position, heading, aircraft turn rate, altitude, climb/descent rate, ground speed)
  Current flight plan (destination, filed plan)
  Aircraft capabilities (turn rate, climb/descent rate, cruising speed, max/min speed)
  Equipment on board
  Aircraft type
  Fuel/loading
  Aircraft status
    Activity (en route, arriving, departing, handed off, pointed out)
    Level of control, instrument flight rules (IFR), visual flight rules (VFR), flight following, VFR on top, uncontrolled object)
    Aircraft contact established
    Aircraft descent established
  Communications (present/frequency)
  Responsible controller
  Aircraft priority
    Special conditions
    Equipment malfunctions
    Emergencies
  Pilot capability/state/intentions
  Altimeter setting
Emergencies
  Type of emergency
  Time on fuel remaining
  Souls on board
Requests
  Pilot/controller requests
  Reason for request
Clearances
  Assignment given
  Received by correct aircraft
  Readback correct/complete
  Pilot acceptance of clearance
  Flight progress strip current
Sector
  Special airspace status
  Equipment functioning
  Restrictions in effect
  Changes to standard procedures
Special operations
  Type of special operation
  Time begin/terminate operations
  Projected duration
  Area and altitude effected
ATC equipment malfunctions
  Equipment affected

*(continued overleaf)*

**Table 1  Continued**

Alternate equipment available
Equipment position/range
Aircraft in outage area
Airports
Operational status
Restrictions in effect
Direction of departures
Current aircraft arrival rate
Arrival requirements
Active runways/approach
Sector saturation
Aircraft in holding (time, number, direction,
leg length)
Weather
Area affected
Altitudes affected
Conditions (snow, icing, fog, hail, rain, turbulence,
overhangs)
Temperatures
Intensity
Visibility
Turbulence
Winds
IFR/VFR conditions
Airport conditions

**Level 2 SA**

Conformance
Amount of deviation (altitude, airspeed, route)
Time until aircraft reaches assigned altitude, speed,
route/heading
Current separation
Amount of separation between
aircraft/objects/airspace/ground along route
Deviation between separation and prescribed
Limits
Number/timing aircraft on routes
Altitudes available
Timing
Projected time in airspace
Projected time until clear of airspace
Time until aircraft landing expected
Time/distance aircraft to airport
Time/distance until visual contact
Order/sequencing of aircraft
Deviations
Deviation aircraft/landing request
Deviation aircraft/flight plan
Deviation aircraft/pilot requests
Other sector/airspace
Radio frequency
Aircraft duration/reason for use

**Table 1  Continued**

Significance
Impact of requests/clearances on:
Aircraft separation/safety
Own/other sector workload
Impact of weather on:
Aircraft safety/flight comfort
Own/other sector workload
Aircraft flow/routing (airport arrival rates, flow rates,
holding requirements aircraft routes, separation
procedures)
Altitudes available
Traffic advisories
Impact of special operations on sector
Operations/procedures
Location of nearest capable airport for aircraft
type/emergency
Impact of malfunction on: routing, communications,
flow control, aircraft, coordination procedures,
other sectors, own workload
Impact on workload of number of aircraft sector
demand vs. own capabilities
Confidence level/accuracy of information
Aircraft ID, position, altitude, airspeed, heading
Weather
Altimeter setting

**Level 3 SA**

Projected aircraft route (current)
Position, fight plan, destination, heading, route,
altitude, climb/descent rate, airspeed, winds,
groundspeed, intentions, assignments
Projected aircraft route (potential)
Projected position $x$ at time $t$
Potential assignments
Projected separation
Amount of separation along route
(aircraft/objects/airspace/ground)
Deviation between separation and prescribed limits
Relative projected aircraft routes
Relative timing along route
Predicted changes in weather
Direction/speed of movement
Increasing/decreasing in intensity
Impact of potential route changes
Type of change required
Time and distance until turn aircraft amount of
turn/new heading, altitude, route change required
Aircraft ability to make change
Projected number of changes necessary
Increase/decrease length of route
Cost/benefit of new clearance
Impact of proposed change on aircraft separation

*Source*: Endsley and Rogers (1994a).
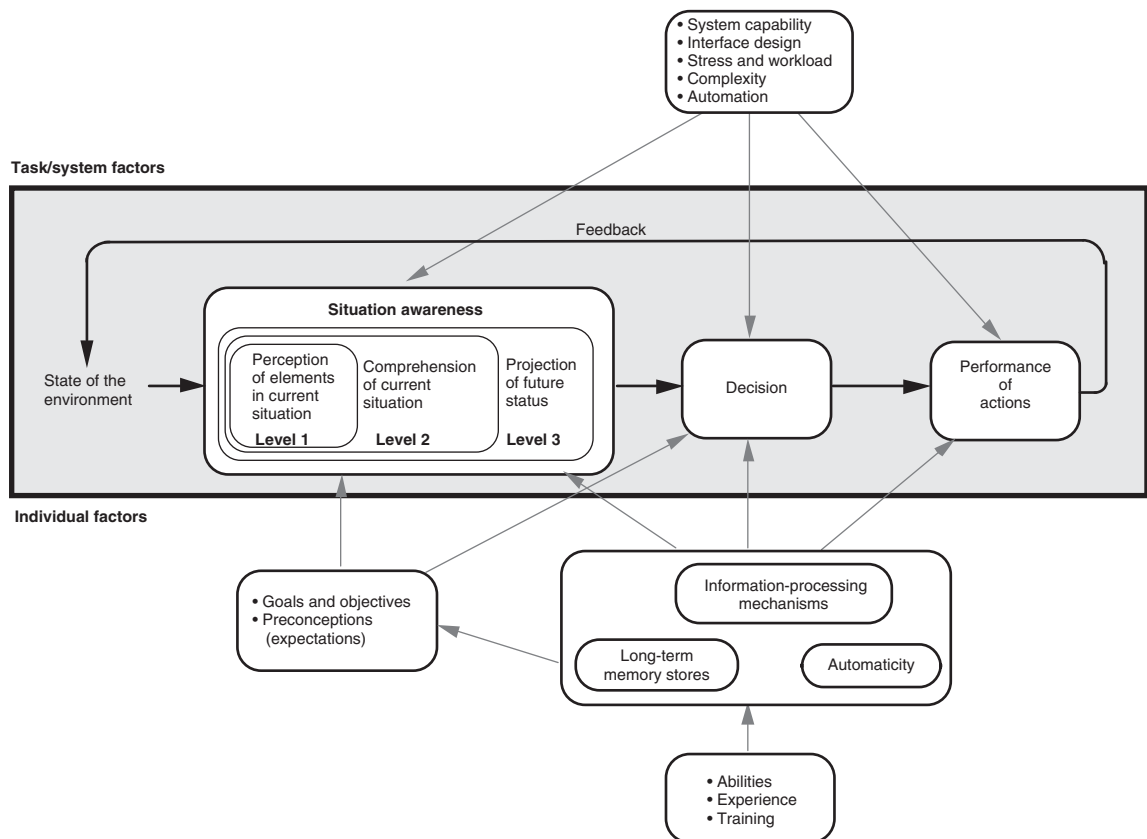
**Figure 3** Model of SA in dynamic decision making. (From Endsley, 1995d, 1997.)

SA is based. According to this model, which is formulated in terms of information-processing theory, elements in the environment may initially be processed in parallel through preattentive sensory stores, where certain properties are detected, such as spatial proximity, color, simple properties of shapes, and movement, providing cues for further focalized attention. Those objects that are most salient are processed further using focalized attention to achieve perception. Limited attention creates a major constraint on an operator's ability to perceive multiple items accurately in parallel and, as such, is a major limiting factor on a person's ability to maintain SA in complex environments.

The description thus far accurately depicts only simple data-driven processing; however, the model also shows a number of other factors that affect this process. First, attention and the perception process can be directed by the contents of both working and long-term memory. For instance, advance knowledge regarding the location of information, the form of the information, the spatial frequency, the color, or the overall familiarity and appropriateness of the information can all significantly facilitate perception. Long-term memory also serves to shape the perception of objects in terms of known

categories or mental representations. Categorization tends to occur almost instantly.

For operators who have not developed other cognitive mechanisms (novices and those in novel situations), the perception of the elements in the environment (the first level of SA) is significantly limited by attention and working memory. In the absence of other mechanisms, most of the operator's active processing of information must occur in working memory. New information must be combined with existing knowledge and a composite picture of the situation developed. Projections of future status and subsequent decisions as to appropriate courses of action will also occur in working memory. Working memory will be significantly taxed while simultaneously achieving the higher levels of SA, formulating and selecting responses, and carrying out subsequent actions.

In actual practice, however, goal-directed processing and long-term memory (often in the form of mental models and schema) can be used to circumvent the limitations of working memory and direct attention more effectively. First, much relevant knowledge about a system is hypothesized to be stored in mental models. Rouse and Morris (1985, p.7) define mental models as "mechanisms whereby humans are able to generate descriptions of system purpose and form, explanations

of system functioning and observed system states, and predictions of future states."

Mental models are cognitive mechanisms that embody information about system form and function; often, they are relevant to a physical system (e.g., a car, computer, or power plant) or an organizational system (e.g., how a university, company, or military unit works). They typically contain information about not only the components of a particular system but also how those components interact to produce various system states and events. Mental models can significantly aid SA as people recognize key features in the environment that map to key features in the model. The model then creates a mechanism for determining associations between observed states of components (comprehension) and predictions of the behavior and status of these elements over time. Thus, mental models can provide much of the higher levels of SA (comprehension and projection) without loading working memory.

Also associated with mental models are schema: prototypical classes of states of the system (e.g., an engine failure, an enemy attack formation, or a dangerous weather formation). These schema are even more useful to the formation of SA since these recognized classes of situations provide an immediate one-step retrieval of the higher levels of SA, based on pattern matching between situation cues and known schema in memory. Very often scripts, set sequences of actions, have also been developed for schema, so that much of the load on working memory for generating alternative behaviors and selecting among them is also diminished. These mechanisms allow the operator simply to execute a predetermined action for a given recognized class of situations (based on their SA). The current situation does not need to be exactly like the one encountered previously, due to the use of categorization mapping; as long as a close-enough mapping can be made into relevant categories, a situation can be recognized and comprehended in terms of the model, predictions made, and appropriate actions selected. Since people have very good pattern-matching abilities, this process can be almost instantaneous and produce a much lower load on working memory, which makes high levels of SA possible, even in very demanding situations.

Expertise therefore plays a major role in the SA process. For novices or those dealing with novel situations, decision making in complex and dynamic systems can be very demanding or impossible to accomplish successfully in that it requires detailed mental calculations based on rules or heuristics, placing a heavy burden on working memory. Where experience has allowed the development of mental models and schema, pattern matching between the perceived elements in the environment and existing schema/mental models can occur on the basis of pertinent cues that have been learned. Thus, the comprehension and future projection required for the higher levels of SA can be developed with far less effort and within the constraints of working memory. When scripts have been developed, tied to these schema, the entire decision-making process will be greatly simplified.

The operator's goals also play an important part in the process. These goals can be thought of as ideal states of the system model that the operator wishes to achieve. In what Casson (1983) has termed a *top-down* decision-making process, the operator's goals and plans will direct which environmental aspects are attended to in the development of SA. Goal-driven or top-down processing is very important in the effective information process and development of SA. Conversely, in a *bottom-up* or data-driven process, patterns in the environment may be recognized which will indicate to the operator that different plans will be necessary to meet goals or that different goals should be activated.

Alternating between "goal driven" and "data driven" is characteristic of much human information processing and underpins much of the SA development in complex worlds. People who are purely data driven are very inefficient at processing complex information sets; there is too much information, so they are simply reactive to the cues that are most salient. People who have clearly developed goals, however, will search for information that is relevant to those goals (on the basis of the associated mental model, which contains information on which aspects of the system are relevant to goal attainment), allowing the information search to be more efficient and providing a mechanism for determining the relevance of the information that is perceived. If people are only goal driven, however, they are likely to miss key information that would indicate that a change in goals is needed (e.g., no longer the goal "land the airplane" but the goal "execute a go-around"). Thus, effective information processing is characterized by alternating between these modes: using goal-driven processing to find and process efficiently the information needed for achieving goals and data-driven processing to regulate the selection of which goals should be most important at any given time.

The development of SA is a dynamic and ongoing process that is effected by these key cognitive mechanisms. Although it can be very challenging in many environments, with mechanisms that can be developed through experience (schema and mental models), we find that people are able to circumvent certain limitations (working memory and attention) to develop sufficient levels of SA to function very effectively. Nevertheless, developing accurate SA remains a very challenging feature in many complex settings and demands a significant portion of an operator's time and resources. Thus, developing selection batteries, training programs, and system designs to enhance SA is a major goal in many domains.

## 3  SITUATION AWARENESS CHALLENGES

Building and maintaining SA can be a difficult process for people in many different jobs and environments. Pilots report that the majority of their time is generally spent trying to ensure that their mental picture of what is happening is current and correct. The same can be said for people in many other domains, where systems are complex and there is a great deal of information to understand, where information changes rapidly, and where information is difficult to obtain. The reasons for this have been captured in terms of eight SA demons, factors that work to undermine SA in many systems and environments (Endsley et al., 2003).

### 3.1 Attentional Tunneling

Successful SA is highly dependent on constantly juggling one's attention between different aspects of the environment. Unfortunately, there are significant limits on people's ability to divide their attention across multiple aspects of the environment, particularly within a single modality, such as vision or sound, and thus attention sharing can occur only to a limited extent (Wickens, 1992). They can often get trapped in a phenomenon called *attentional narrowing* or *tunneling* (Bartlett, 1943; Broadbent, 1954; Baddeley, 1972).

When succumbing to attentional tunneling, they lock in on certain aspects or features of the environment they are trying to process and will intentionally or inadvertently drop their scanning behavior. In this case, their SA may be very good on the part of the environment of their concentration but will quickly become outdated on other aspects they are not watching. Attentional narrowing has been found to undermine SA in tasks such as flying and driving and poses one of the most significant challenges to SA in many domains.

### 3.2 Requisite Memory Trap

The limitations of working memory also create a significant SA demon. Many features of the situation may need to be held in memory. As a person scans different information from the environment, information accessed previously must be remembered and combined with new information. Auditory information must also be remembered, as it cannot be revisited in the way that visual displays can. Given the complexity and sheer volume of information required for SA in many systems, these memory limits create a significant problem for SA. System designs that necessitate that people remember information, even short term, increase the likelihood of SA error.

### 3.3 Workload, Anxiety, Fatigue, and Other Stressors

Stressors such as anxiety, time pressure, mental workload, uncertainty, noise or vibration, excessive heat or cold, poor lighting, physical fatigue, and working against one's circadian rhythms are unfortunately an unavoidable part of many work environments. These stressors can act to reduce SA significantly by further reducing an already limited working memory and reducing the efficiency of information gathering. It has been found that people may pay less attention to peripheral information, become more disorganized in scanning information, and are more likely to succumb to attentional tunneling when affected by these stressors. People are also more likely to arrive at a decision without taking into account all available information (premature closure).

### 3.4 Data Overload

Data overload is a significant problem in many systems. The volume of data and the rapid rate of change of that data create a need for information intake that quickly outpaces one's ability to gather and assimilate the data. As people can take in and process only a limited amount of information at a time, significant lapses in SA can occur. While it is easy to think of this problem as simply a human limitation, in reality it often occurs because data are processed, stored, and presented ineffectively in many systems. This problem is not just one of volume but also one of bandwidth, the bandwidth provided by a person's sensory and information-processing mechanisms. The rate that data can flow through the pipeline can be increased significantly based on the form of information presentation employed in the interface.

### 3.5 Misplaced Salience

The human perceptual system is more sensitive to certain features than others, including the color red, movement, and flashing lights. Similarly, loud noises, larger shapes, and things that are physically nearer have the advantage of catching a person's attention. These natural salient properties can be used to promote SA or to hinder it. When used carefully, properties such as movement or color can be used to draw attention to critical and highly important information and are thus important tools for designing to enhance SA. Unfortunately, these features are often overused or used inappropriately. The unnecessary distractions of misplaced salience can act to degrade SA of the other information the person is attempting to assimilate. Unfortunately, in many systems there is a proliferation of lights, buzzers, alarms, and other signals that work actively to draw people's attention, frequently either misleading or overwhelming them.

### 3.6 Complexity Creep

Over time, systems have become more and more complex, often through a misguided attempt to add more features or capabilities. Unfortunately, this complexity makes it difficult for people to form sufficient internal representations of how these systems work. The more features and the more complicated and branching the rules that govern a system's behavior, the greater the complexity. Although system complexity can slow down a person's ability to take in information, it works primarily to undermine the person's ability to correctly interpret the information presented and to project what is likely to happen (level 2 and 3 SA). A cue that should indicate one thing can be completely misinterpreted, as the internal mental model will be developed inadequately to encompass the full characteristics of the system.

### 3.7 Errant Mental Models

Mental models are important mechanisms for building and maintaining SA, providing key interpretation mechanisms for information collected. They tell a person how to combine disparate pieces of information, how to interpret the significance of that information, and how to develop reasonable projections of what will happen in the future. If an incomplete mental model is used, however, or if the wrong mental model is relied on for the situation, poor comprehension and projection (level 2 and 3 SA) can result. Also called a *representational*

*error*, it can be very difficult for people to realize that they are working on the basis of an errant mental model and break out of it. Mode errors, in which people misunderstand information because they believe that the system is in one mode when it is really in another, are a special case of this problem.

### 3.8   Out-of-the-Loop Syndrome

Automation creates a final SA demon. While in some cases automation can help SA by eliminating excessive workload, it can also act to lower SA by putting people *out of the loop*. In this state, they develop poor SA as to both how the automation is performing and the state of the elements the automation is supposed to be controlling. When the automation is performing well, being out of the loop may not be a problem, but when the automation fails or, more frequently, reaches situational conditions that it is not equipped to handle, the person is out of the loop and often unable to detect the problem, properly interpret the information presented, and intervene in a timely manner.

## 4   TRAINING TO SUPPORT SITUATION AWARENESS

There is some evidence that some people are significantly better than others at developing SA. In one study of experienced military fighter pilots, Endsley and Bolstad (1994) found a 10-fold difference in SA between the pilot with the lowest SA and the one with the highest SA. They also found this ability to be highly stable, with test–retest reliability rates exceeding 0.94 for those evaluated. Others (Secrist and Hartman, 1993; Bell and Waag, 1995) have similarly noted consistent individual differences, with some pilots routinely having better SA than their compatriots. These individual differences appear even when people operate with the same system capabilities and displays and in the same environment subject to the same demands.

A number of studies have sought to find the locus of these individual differences in SA abilities. Are they due simply to the effects of expertise and experience or are they indicative of the better cognitive mechanisms or capabilities that some people have? Endsley and Bolstad (1994) found that military pilots with better SA were better at attention sharing, pattern matching, spatial abilities, and perceptual speed. O'Hare (1997) also found evidence that elite pilots (defined as consistently superior in gliding competitions) performed better on a divided-attention task purported to measure SA. Gugerty and Tirre (1997) found evidence that people with better SA performed better on measures of working memory, visual processing, temporal processing, and time-sharing ability.

Although these studies have examined individual differences in only a few domains (e.g. piloting and driving), some of these attributes may also be relevant to SA differences in other arenas. If reliable markers can be found that differentiate those who will eventually be most successful at SA, more valid selection batteries can be developed for critical jobs such as air traffic controller, pilot, or military commander.

There has also been research to examine what skills differentiate those with high SA from those with low SA which might be trainable, thus significantly improving SA in the existing population of operators in a domain. For instance, SA differences between those at different levels of expertise have been examined in groups of pilots (Prince and Salas, 1998; Endsley et al., 2000), military officers (Strater et al., 2003), aircraft mechanics (Endsley and Robertson, 2000a), power plant operators (Collier and Folleso, 1995), and drivers (Horswill and McKenna, 2004). These studies have found many systematic differences, some of which may relate to underlying abilities, but many of which also point to learned skills or behaviors that may be trainable.

A number of training programs have been developed that seek to train knowledge and skills related to developing SA (at the individual or team level) in classroom settings, simulated scenarios or case studies, or through computer-based training. These include training programs for commercial aviation pilots (Robinson, 2000; Hormann et al., 2004), general aviation pilots (Prince, 1998; Endsley and Garland, 2000a; Bolstad et al., 2002; Endsley et al., 2002), drivers (Sexton, 1988; McKenna and Crick, 1994), aircraft mechanics (Endsley and Robertson, 2000a,b), and army officers (Strater et al., 2003, 2004). While some of these are classroom-based programs that seek to create more awareness in operators about the concept of SA and the many challenges that effect it, many more detailed programs have been created that seek to build up the critical knowledge and skills that underlie SA.

### 4.1   Interactive Situation Awareness Trainer (ISAT)

ISAT employs rapid experiential learning in support of mental model and schema development. In normal operations, over the course of many months and years, individuals will gradually build up the experience base to develop good mental models and schema for pattern matching upon which SA most often relies. ISAT attempts to boot strap this natural process by exposing the trainee to many, many situations in a very short period of time using computer-based training tools (Strater, et al., 2004). ISAT employs realistic mission scenarios with opportunities for complex operational decisions. It provides an increased opportunity for exposure to a variety of situations which (1) supports the development of situation-based knowledge stores, (2) trains the recognition of critical cues that signal prototypical situations, (3) supports information integration and decision making, and (4) promotes an understanding of the importance of consequences, timing, risk and capabilities associated with different events, behaviors, and decision options. Trainees learn what it means to develop SA in the environment, learn to build higher level SA out of data, and receive training on projecting future events in prototypical situations.

### 4.2   Virtual Environment Situation Awareness Review System (VESARS)

Feedback is critical to the learning process. In order to improve SA, individuals need to receive feedback on the

quality of their SA; however this often is lacking in the real world. For example, inexperienced operators may fail to appreciate the severity of threatening conditions because they have come through similar conditions in the past just by luck. Unfortunately, this also reinforces poor assessments. It is difficult for individuals to develop a good gage of their own SA in normal operations. Training through SA feedback allows trainees to fine tune critical behaviors and mental models based on a review of their SA performance (Endsley, 1989b).

The VESARS approach involves the use of SA measures that assess trainee SA in three areas: (1) a behavioral rating tool that assesses individual and team actions, (2) a communications rating tool that evaluates team communications, and (3) a SA query tool that allows direct and objective assessment of the quality of individual and team SA (Kaber et al., 2005, 2006). VESARS was specifically designed to work well with virtual and simulated training environments but it can also be employed in field exercises. SA training is provided after each simulated trial in which VESARS data are collected. Providing knowledge of results immediately following a simulation on the SA level achieved across the various SA requirements and relevant communications and behaviors allows trainees to understand the degree to which they were able to acquire SA and ways in which they need to modify their processes to improve their SA.

### 4.3 Situation Awareness Virtual Instructor (SAVI)

The SAVI trains warfighters on the behaviors that are consistent with and important to good SA by allowing trainees to play the role of the trainer as they rate the actions of others in vignettes provided through a computer and provide a rationale for their rating (Endsley et al., 2009). The SAVI approach leverages the exponential learning that occurs during peer instruction and in the transition to becoming a trainer. Trainees quickly learn what behaviors are appropriate for various operational situations, because they observe these aspects of performance and provide their assessments on the quality of the performance. Trainees are able to refine their mental models of good SA behaviors and communications as they also compare their assessments to those provided by domain experts. This allows trainees to fine tune their understanding of critical cues and behaviors associated with good SA.

The preliminary findings reported by the majority of these efforts show initial successes in improving SA and performance in their respective settings. In general, more longitudinal studies are needed to ascertain the degree to which such efforts can be successful in improving the SA of persons in the wide variety of challenging situations that are common in these domains.

## 5 SYSTEM DESIGN TO SUPPORT SITUATION AWARENESS

In addition to training to improve SA at the individual level, efforts to improve SA through better sensors,

information processing, and display approaches have characterized much of the past 20 years. Unfortunately, a significant portion of these efforts have stopped short of really addressing SA; instead, they simply add a new sensor or black box that is purported to improve SA. While ensuring that operators have the data needed to meet their level 1 SA requirements is undoubtedly important, a rampant increase in such data may inadvertently hurt SA as much as it helps. Simply increasing the amount of data available to an operator instead adds to the information gap, overloading the operator without necessarily improving the level of SA the person can develop and maintain.

As a construct, however, SA provides a key mechanism for overcoming this data overload. SA specifies how all the data in an environment need to be combined and understood. Therefore, instead of loading the operator down with 100 pieces of miscellaneous data provided in a haphazard fashion, SA requirements provide guidance as to what the real comprehension and projection needs are. Therefore, it provides the system designer with key guidance on how to bring the various pieces of data together to form meaningful integrations and groupings of data that can be absorbed and assimilated easily in time-critical situations. This type of systems integration usually requires very unique combinations of information and portrayals of information that go far beyond the black-box technology-oriented approaches of the past. In the past it was up to the operator to do it all. This task left him or her overloaded and susceptible to missing critical factors. If system designers work to develop systems that support the SA process, however, they can alleviate this bottleneck significantly.

So how should systems be designed to meet the challenge of providing high levels of SA? Over the past decade a significant amount of research has been focused on this topic, developing an initial understanding of the basic mechanisms that are important for SA and the design features that will support those mechanisms. Based on this research, the SA-oriented design process has been established (Endsley et al., 2003) to guide the development of systems that support SA (Figure 4). This structured approach incorporates SA considerations into the design process, including a determination of SA requirements, design principles for SA enhancement, and measurement of SA in design evaluation.

### 5.1 SA Requirements Analysis

The problem of determining what aspects of the situation are important for a particular operator's SA has frequently been approached using a form of cognitive task analysis called *goal-directed task analysis*, illustrated in Figure 5. In such analysis, the major goals of a particular job class are identified, along with the major subgoals necessary for meeting each goal. Associated with each subgoal, the major decisions that need to be made are then identified. The SA needed for making these decisions and carrying out each subgoal are identified. These SA requirements focus not only on what data the operator needs but also on how that information is integrated or combined to address each decision. In this analysis process, SA requirements are defined as
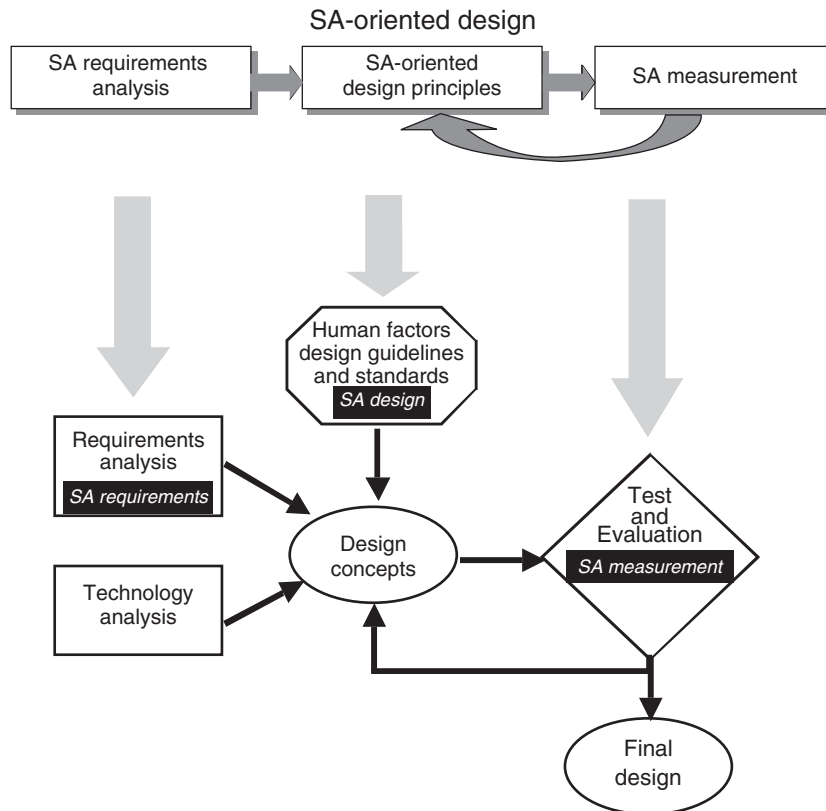
## SA-oriented design



**Figure 4**   SA-oriented design process. (From Endsley et al., 2003.)
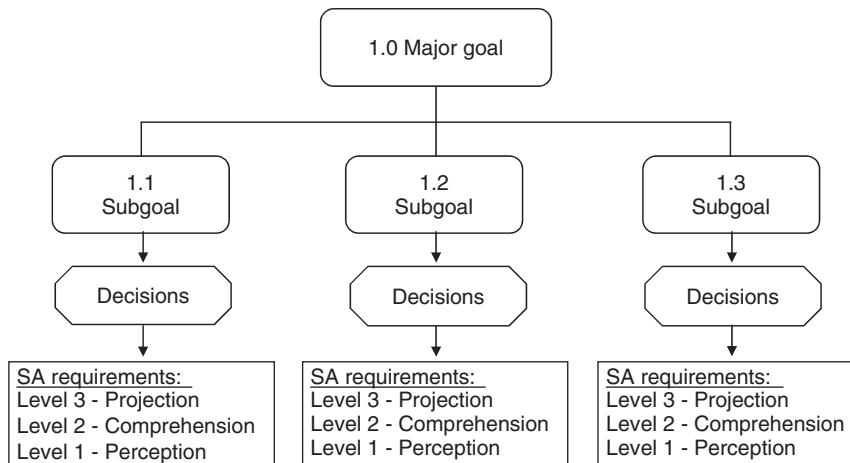


**Figure 5**   Goal-directed task analysis for determining SA requirements.

those dynamic information needs associated with the major goals or subgoals of the operator in performing his or her job (as opposed to more static knowledge, such as rules, procedures, and general system knowledge). This type of analysis is based on goals or objectives, not tasks (as a traditional task analysis might). This is because goals form the basis for decision making in many complex environments. Conducting such an analysis is usually carried out using a combination of cognitive engineering procedures. Expert

elicitation, observation of operator performance of tasks, verbal protocols, analysis of written materials and documentation, and formal questionnaires have formed the basis for the analyses. In general, the analysis has been conducted with a number of operators, who are interviewed, observed, and recorded individually, with the resulting analyses pooled and then validated overall by a larger number of operators.

An example of the output of this process is shown in Table 2. This example shows the SA requirements analysis for the subgoal "maintain aircraft conformance" for the major goal "avoid conflictions" for an air traffic controller. In this example, the subgoal is divided even further into lower level subgoals prior to the decisions and SA requirements being listed. In some cases,

### Table 2  Example of Goal-Directed Task Analysis for En Route Air Traffic Control

1.3 Maintain aircraft conformance

1.3. Assess aircraft conformance to assigned parameters

- Aircraft at/proceeding to assigned altitude?
- Aircraft proceeding to assigned altitude fast enough?
- Time until aircraft reaches assigned altitude
- Amount of altitude deviation
- Climb/descent
- Altitude (current)
- Altitude (assigned)
- Altitude rate of change (ascending/descending)
- Aircraft at/proceeding to assigned airspeed?
- Aircraft proceeding to assigned airspeed fast enough?
- Time until aircraft reaches assigned airspeed
- Amount of airspeed deviation
- Airspeed (indicated)
- Airspeed (assigned)
- Groundspeed
- Aircraft on/proceeding to assigned route?
- Aircraft proceeding to assigned route fast enough?
- Aircraft turning?
- Time until aircraft reaches assigned route/heading
- Amount of route deviation
- Aircraft position (current)
- Aircraft heading (current)
- Route/heading (assigned)
- Aircraft turn rate (current)
- Aircraft heading (current)
- Aircraft heading (past)
- Aircraft turn capabilities
- Aircraft type
- Altitude
- Aircraft groundspeed
- Weather
- Winds (direction, magnitude)

*Source*: Endsley and Rodgers (1994b).

addressing a particular subgoal occurs through reference to another subgoal in other parts of the analysis, such as the need to readdress aircraft separation in this example. This shows the degree to which a particular operator's goals and resultant SA needs may be very interrelated. The example in Table 2 shows just one major subgoal out of four that are relevant for the major goal "avoid conflictions," which is just one of three major goals for an air traffic controller.

This analysis defines systematically the SA requirements (at all three levels of SA) that are needed to effectively make the decisions required by the operator's goals. Many of the same SA requirements appear throughout the analysis. In this manner, the way in which pieces of data are used together and combined to form what the operator really wants to know is determined. Although the analysis will typically include many goals and subgoals, they may all be active at once. In practice, at any given time more than one goal or subgoal may be operational, although they will not always have the same prioritization. The analysis does not indicate any prioritization among the goals (which can vary over time) or that each subgoal within a goal will always be active. Unless particular events are triggered (e.g., the subgoal of assuring aircraft conformance in this example), a subgoal may not be active for a given controller.

The analysis strives to be as technology free as possible. How the information is acquired is not addressed, as this can vary considerably from person to person, from system to system, and from time to time. In some cases it may be through system displays, verbal communications, other operators, or internally generated from within the operator. Many of the higher level SA requirements fall into this category. The way in which information is acquired can vary widely between persons, over time, and between system designs.

The analysis seeks to determine what operators would ideally like to know to meet each goal. It is recognized that they often must operate on the basis of incomplete information and that some desired information may not be available at all with today's system. However, for purposes of design and evaluation of systems, we need to set the yardstick to measure against what they ideally need to know, so that artificial ceiling effects, based on today's technology, are not induced in the process. Finally, it should be noted that static knowledge, such as procedures or rules for performing tasks, is outside the bounds of an SA requirements analysis. The analysis focuses on the dynamic situational information that affects what the operators do.

To date, these analyses have been completed for many domains of common concern, including en route air traffic control (Endsley and Rodgers, 1994b), TRACON air traffic control (Endsley and Jones, 1995), fighter pilots (Endsley, 1993), bomber pilots (Endsley, 1989a), commercial transport pilots (Endsley et al., 1998b), aircraft mechanics (Endsley and Robertson, 1996), and airway facilities maintenance (Endsley and Kiris, 1994). A similar process was employed by Hogg et al. (1993) to determine appropriate queries for a nuclear reactor domain.

## 5.2   SA-Oriented Design Principles

The development of a system design for successfully providing the multitude of SA requirements that exist in complex systems is a significant challenge. Design principles have been developed based on the theoretical model of the mechanisms and processes involved in acquiring and maintaining SA in dynamic complex systems (Endsley, 1988, 1990b, 1995d; Endsley et al., 2003). The 50 design principles include (1) general guidelines for supporting SA, (2) guidelines for coping with automation and complexity, (3) guidelines for the design of alarm systems, (4) guidelines for the presentation of information uncertainty, and (5) guidelines for supporting SA in team operations. Some of the general principles include the following: (1) direct presentation of higher level SA needs (comprehension and projection) is recommended, rather than supplying only low-level data that operators must integrate and interpret manually; (2) goal-oriented information displays should be provided, organized so that the information needed for a particular goal is co-located and answers directly the major decisions associated with the goal; (3) support for global SA is critical, providing an overview of the situation across the operator's goals at all times (with detailed information for goals of current interest) and enabling efficient and timely goal switching and projection; (4) critical cues related to key features of schemata need to be determined and made salient in the interface design (in particular, those cues that will indicate the presence of prototypical situations will be of prime importance and will facilitate goal switching in critical conditions); (5) extraneous information not related to SA needs should be removed (while carefully ensuring that such information is not needed for broader SA needs); and (6) support for parallel processing, such as multimodal displays, should be provided in data-rich environments.

SA-oriented design is applicable to a wide variety of system designs. It has been used successfully as a design philosophy for systems involving remote maintenance operations, medical systems, flexible manufacturing cells, and command and control for distributed teams.

## 5.3   Design Evaluation for SA

Many concepts and technologies are currently being developed and touted as enhancing SA. Prototyping and simulation of new technologies, new displays, and new automation concepts are extremely important for evaluating the actual effects of proposed concepts within the context of the task domain and using domain-knowledgeable subjects. If SA is to be a design objective, it is critical that it be evaluated specifically during the design process. Without this, it will be impossible to tell if a proposed concept actually helps SA, does not affect it, or inadvertently compromises it in some way. A primary benefit of examining system design from the perspective of operator SA is that the impact of design decisions on SA can be assessed objectively as a measure of the quality of the integrated system design when used within the actual challenges of the operational environment.

SA measurement has been approached in a number of ways. See Endsley and Garland (2000b) for details on these methods. A review of the advantages and disadvantages of these methods may be found in Endsley (1996) and Endsley and Smolensky (1998). In general, direct measurement of SA can be very advantageous in providing more sensitivity and diagnosticity in the test and evaluation process. This provides a significant addition to performance measurement and workload measurement in determining the utility of new design concepts. Whereas workload measures provide insight into how hard an operator must work to perform tasks with a new design, SA measurement provides insight into the level of understanding gained from that work.

Direct measurement of SA has generally been approached either through subjective ratings or by objective techniques. Although subjective ratings are simple and easy to administer, research has shown that they correlate poorly with objective SA measures, indicating they more closely capture a person's confidence in his or her SA rather than the actual level or accuracy of that SA (Endsley et al., 1998a).

One of the most widely used objective measures of SA is the SA global assessment technique (SAGAT) (Endsley, 1988, 1995b, 2000a). SAGAT has been used successfully to measure operator SA directly and objectively when evaluating avionics concepts, display designs, and interface technologies (Endsley, 1995b). Using SAGAT, a simulated test scenario employing the design of interest is frozen at randomly selected times, the system displays are blanked, and the simulation is suspended while operators quickly answer questions about their current perceptions of the situation. The questions correspond to their SA requirements as determined from an SA requirements analysis for that domain. Operator perceptions are then compared to the real situation based on simulation computer databases to provide an objective measure of SA.

Multiple "snapshots" of operators' SA can be acquired in this way, providing an index of the quality of SA provided by a particular design. The collection of SA data in this manner provides an objective, unbiased assessment of SA that overcomes the problems incurred when collecting such data after the fact and minimizes biasing of controller SA due to secondary task loading or cuing the controller's attention artificially, which real-time probes may do. By including queries across the full spectrum of an operator's SA requirements, this approach minimizes possible biasing of attention, as subjects cannot prepare for the queries in advance since they could be queried over almost every aspect of the situation to which they would normally attend. The primary disadvantage of this technique involves the temporary halt in the simulation.

The method is not without some costs, however, as a detailed analysis of SA requirements is required in order to develop the battery of queries to be administered. SAGAT is a global tool developed to assess SA across all of its elements based on a comprehensive assessment of operator SA. As a global measure, SAGAT includes queries about all operator SA requirements, including level 1 (perception of data), level 2 (comprehension of

meaning), and level 3 (projection of the near future) components. This includes a consideration of system functioning and status as well as relevant features of the external environment.

SAGAT has also been shown to have predictive validity, with SAGAT scores indicative of pilot performance in a combat simulation (Endsley, 1990a). It is also sensitive to changes in task load and to factors that affect operator attention (Endsley, 2000a), demonstrating construct validity. It has been found to produce high levels of reliability (Endsley and Bolstad, 1994; Collier and Folleso, 1995; Gugerty, 1997). Studies examining the intrusiveness of the freezes to collect SAGAT data have generally found there to be no effect on operator performance (Endsley, 1995a, 2000a)

An example of the use of SAGAT for evaluating the impact of new system concepts may be found in Endsley et al. (1997a). A totally new form of distributing roles and responsibilities between pilots and air traffic controllers was examined. Termed *free flight*, this concept was originally developed as a major change in the operation of the national airspace. It may include pilots filing direct routes to destinations rather than along predefined fixed airways and authority for the pilot to deviate from that route either with air traffic controllers' permission or perhaps even fully autonomously (RTCA, 1995). As it was felt that such changes could have a marked effect on the ability of the controller to keep up as monitor in such a new system, a study was conducted to examine this possibility (Endsley et al., 1997b).

Results showed a trend toward poorer controller performance in detecting and intervening in aircraft separation errors with these changes in the operational concept and poorer subjective ratings of performance. Finding statistically significant changes in separation errors during ATC simulation testing is quite rare, however. More detailed analysis of the SAGAT results provided more diagnostic detail as well as backing up this finding. As shown in Figure 6, controllers were aware of significantly fewer aircraft in the simulation under free-flight conditions. Attending to fewer aircraft under a higher workload has also been found in other studies (Endsley and Rodgers, 1998).

In addition to reduced level 1 SA, however, controllers had a significantly reduced understanding (level 2 SA) of what was happening in the traffic situation, as evidenced by lower SA regarding which aircraft weather would affect the situation and a reduced awareness of those aircraft that were in a transitionary state. They were less aware of which aircraft had not yet completed a clearance and, for those aircraft, whether the instruction was received correctly and whether they were conforming. Controllers also demonstrated lower level 3 SA with free flight. Their knowledge of where the aircraft was going (to the next sector) was significantly lower under free-flight conditions.

These findings were useful in pinpointing whether concerns over this new and very different concept were justified or whether they merely represented resistance to change. The SAGAT results showed not only that the new concept did indeed induce problems for controller SA that would prevent them from performing effectively
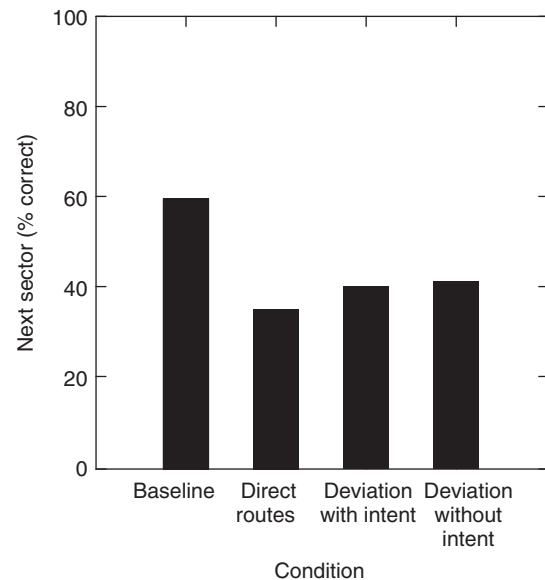


**Figure 6**  SAGAT results. (From Endsley et al., 1997b.)

as monitors to back up pilots with separation assistance but also in what ways these problems were manifested. This information is very useful diagnostically in that it allows one to determine what sort of aid might be needed for operators to assist them in overcoming these deficiencies.

For instance, in this example, a display that provides enhanced information on flight paths for aircraft in transitionary states may be recommended as a way of compensating for the lower SA observed. Far from just providing a thumbs-up or thumbs-down input on a concept under evaluation, this rich source of data is very useful in developing iterative design modifications and making trade-off decisions.

## 6  CONCLUSIONS

A firm theoretical foundation has been laid for understanding the factors that affect SA in complex environments. This foundation can be used to guide the development of training programs and the development of system designs that go beyond data presentation to provide higher levels of SA. In either case, validation of the effectiveness of the proposed solutions through detailed, objective testing is paramount to ensuring that the approach is actually successful in improving SA.

The need to process and understand large volumes of data is critical for many endeavors, from the cockpit to military missions, from power plants to automobiles, and from space stations to day-to-day business operations. It is likely that the potential benefits of the information age will not be realized until we come to grips with the challenges of managing this dynamic information base to provide people with the SA they need on a real-time basis. Doing so is the primary challenge of the next decade of technology.

## REFERENCES

Adams, M. J., Tenney, Y. J., and Pew, R. W. (1995), "Situation Awareness and the Cognitive Management of Complex Systems," *Human Factors*, Vol. 37, No. 1, pp. 85–104.

Amalberti, R., and Deblon, F. (1992), "Cognitive Modeling of Fighter Aircraft Process Control: A Step Towards an Intelligent On-Board Assistance System," *International Journal of Man–Machine Systems*, Vol. 36, pp. 639–671.

Baddeley, A. D. (1972), "Selective Attention and Performance in Dangerous Environments," *British Journal of Psychology*, Vol. 63, pp. 537–546.

Bartlett, F. C. (1943), "Fatigue Following Highly Skilled Work," *Proceedings of the Royal Society (B)*, Vol. 131, pp. 147–257.

Bell, H. H., and Waag, W. L. (1995), "Using Observer Ratings to Assess Situational Awareness in Tactical Air Environments," in *Experimental Analysis and Measurement of Situation Awareness*, D. J. Garland and M. R. Endsley, Eds., Embry-Riddle Aeronautical University Press, Daytona Beach, FL, pp. 93–99.

Bolstad, C. A., Endsley, M. R., Howell, C., and Costello, A. (2002), "General Aviation Pilot Training for Situation Awareness: An Evaluation," in *Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 21–25.

Broadbent, D. E. (1954), "Some Effects of Noise on Visual Performance," *Quarterly Journal of Experimental Psychology*, Vol. 6, pp. 1–5.

Casson, R. W. (1983), "Schema in Cognitive Anthropology," *Annual Review of Anthropology*, Vol. 12, pp. 429–462.

Collier, S. G., and Folleso, K. (1995), "SACRI: A Measure of Situation Awareness for Nuclear Power Plant Control Rooms," in *Experimental Analysis and Measurement of Situation Awareness*, D. J. Garland and M. R. Endsley, Eds., Embry-Riddle Aeronautical University Press, Daytona Beach, FL, pp. 115–122.

Durso, F. T., and Gronlund, S. D. (1999), "Situation Awareness," in *Handbook of Applied Cognition*, F. T. Durso, R. Nickerson, R. Schvaneveldt, S. Dumais, S. Lindsay, and M. Chi, Eds., Wiley, New York, pp. 284–314.

Endsley, M. R. (1988), "Design and Evaluation for Situation Awareness Enhancement," in *Proceedings of the Human Factors Society 32nd Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 97–101.

Endsley, M. R. (1989a), "Final Report: Situation Awareness in an Advanced Strategic Mission," NOR DOC 89-32, Northrop Corporation, Hawthorne, CA.

Endsley, M. R. (1989b), "Pilot Situation Awareness: The Challenge for the Training Community," in *Proceedings of the Interservice/Industry Training Systems Conference (I/ITSC)*, American Defense Preparedness Association), Forth Worth, TX, pp. 111–117.

Endsley, M. R. (1990a), "Predictive Utility of an Objective Measure of Situation Awareness," in *Proceedings of the Human Factors Society 34th Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 41–45.

Endsley, M. R. (1990b), "Situation Awareness in Dynamic Human Decision Making: Theory and Measurement," Ph.D. dissertation, University of Southern California, Los Angeles.

Endsley, M. R. (1993), "A Survey of Situation Awareness Requirements in Air-to-Air Combat Fighters," *International Journal of Aviation Psychology*, Vol. 3, No. 2, pp. 157–168.

Endsley, M. R. (1995a), "Direct Measurement of Situation Awareness in Simulations of Dynamic Systems: Validity and Use of SAGAT," in *Experimental Analysis and Measurement of Situation Awareness*, D. J. Garland and M. R. Endsley, Eds., Embry-Riddle Aeronautical University Press, Daytona Beach, FL, pp. 107–113.

Endsley, M. R. (1995b), "Measurement of Situation Awareness in Dynamic Systems," *Human Factors*, Vol. 37, No. 1, pp. 65–84.

Endsley, M. R. (1995c), "A Taxonomy of Situation Awareness Errors," in *Human Factors in Aviation Operations*, R. Fuller, N. Johnston, and N. McDonald, Eds., Avebury Aviation, Ashgate Publishing, Aldershot, Hampshire, England, pp. 287–292.

Endsley, M. R. (1995d), "Toward a Theory of Situation Awareness in Dynamic Systems," *Human Factors*, Vol. 37, No. 1, pp. 32–64.

Endsley, M. R. (1996), "Situation Awareness Measurement in Test and Evaluation," in *Handbook of Human Factors Testing and Evaluation*, T. G. O'Brien and S. G. Charlton, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 159–180.

Endsley, M. R. (1997), "Communication and Situation Awareness in the Aviation System," presentation to the Aviation Communication: A Multi-cultural Forum, Embry-Riddle University, Prescott, AZ.

Endsley, M. R. (2000a), "Direct Measurement of Situation Awareness: Validity and Use of SAGAT," in *Situation Awareness Analysis and Measurement*, M. R. Endsley and D. J. Garland, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 147–174.

Endsley, M. R. (2000b), "Theoretical Underpinnings of Situation Awareness: A Critical Review," in *Situation Awareness Analysis and Measurement*, M. R. Endsley and D. J. Garland, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 3–32.

Endsley, M. R., and Bolstad, C. A. (1994), "Individual Differences in Pilot Situation Awareness," *International Journal of Aviation Psychology*, Vol. 4, No. 3, pp. 241–264.

Endsley, M. R., and Garland, D. G. (2000a), "Pilot Situation Awareness Training in General Aviation," in *Proceedings of the 14th Triennial Congress of the International Ergonomics Association and the 44th Annual Meeting of the Human Factors and Ergonomics Society*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 357–360.

Endsley, M. R., and Garland, D. J., Eds. (2000b), *Situation Awareness Analysis and Measurement*, Lawrence Erlbaum Associates, Mahwah, NJ.

Endsley, M. R., and Jones, D. G. (1995), "Situation Awareness Requirements Analysis for TRACON Air Traffic Control," TTU-IE-95-01, Texas Tech University, Lubbock, TX.

Endsley, M. R., and Kiris, E. O. (1994), "Situation Awareness in FAA Airway Facilities Maintenance Control Centers (MCC): Final Report," Texas Tech University, Lubbock, TX.

Endsley, M. R., and Robertson, M. M. (1996), *Team Situation Awareness in Aircraft Maintenance*, Texas Tech University, Lubbock, TX.

Endsley, M. R., and Robertson, M. M. (2000a), "Situation Awareness in Aircraft Maintenance Teams," *International Journal of Industrial Ergonomics*, Vol. 26, pp. 301–325.

Endsley, M. R., and Robertson, M. M. (2000b), "Training for Situation Awareness in Individuals and Teams," in *Situation Awareness Analysis and Measurement*, M. R. Endsley and D. J. Garland, Eds., Lawrence Erlbaum Associates, Mahwah, NJ.

Endsley, M. R., and Rodgers, M. D. (1994a), "Situation Awareness Information Requirements for En Route Air Traffic Control," DOT/FAA/AM-94/27, Federal Aviation Administration Office of Aviation Medicine, Washington, DC.

Endsley, M. R., and Rodgers, M. D. (1994b), "Situation Awareness Information Requirements for En Route Air Traffic Control," in *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 71–75.

Endsley, M. R., and Rodgers, M. D. (1998), "Distribution of Attention, Situation Awareness, and Workload in a Passive Air Traffic Control Task: Implications for Operational Errors and Automation," *Air Traffic Control Quarterly*, Vol. 6, No. 1, pp. 21–44.

Endsley, M. R., and Smolensky, M. (1998), "Situation Awareness in Air Traffic Control: The Picture," in *Human Factors in Air Traffic Control*, M. Smolensky and E. Stein, Eds., Academic, New York, pp. 115–154.

Endsley, M. R., Mogford, R., Allendoerfer, K., Snyder, M. D., and Stein, E. S. (1997a), "Effect of Free Flight Conditions on Controller Performance, Workload and Situation Awareness: A Preliminary Investigation of Changes in Locus of Control Using Existing Technology," DOT/FAA/CT-TN 97/12, Federal Aviation Administration William J. Hughes Technical Center, Atlantic City, NJ.

Endsley, M. R., Mogford, R. H., and Stein, E. S. (1997b), "Controller Situation Awareness in Free Flight," in *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 4–8.

Endsley, M. R., Selcon, S. J., Hardiman, T. D., and Croft, D. G. (1998a), "A Comparative Evaluation of SAGAT and SART for Evaluations of Situation Awareness," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 82–86.

Endsley, M. R., Farley, T. C., Jones, W. M., Midkiff, A. H., and Hansman, R. J. (1998b), "Situation Awareness Information Requirements for Commercial Airline Pilots," ICAT-98-1, MIT International Center for Air Transportation, Cambridge, MA.

Endsley, M. R., Garland, D. J., Shook, R. W. C., Coello, J., and Bandiero, M. (2000), "Situation Awareness Problems in General Aviation," SATech 00–01, SA Technologies, Marietta, GA.

Endsley, M. R., et al. (2002), "Situation Awareness Training for General Aviation Pilots: Final Report," SATECH 02–04, SA Technologies, Marietta, GA.

Endsley, M. R., Bolte, B., and Jones, D. G. (2003), *Designing for Situation Awareness: An Approach to Human-Centered Design*, Taylor & Francis, London.

Endsley, M. R., Riley, J. M., and Strater, L. D. (2009), "Leveraging Embedded Training Systems to Build Higher Level Cognitive Skills in Warfighters," in *Proceedings of the NATO Conference on Human Dimensions in Embedded Virtual Simulation*, NATO, Fort Worth, TX.

Farley, T. C., Hansman, R. J., Amonlirdviman, K., and Endsley, M. R. (2000), "Shared Information between Pilots and Controllers in Tactical Air Traffic Control," *Journal of Guidance, Control and Dynamics*, Vol. 23, No. 5, pp. 826–836.

Fracker, M. L. (1988), "A Theory of Situation Assessment: Implications for Measuring Situation Awareness," in *Proceedings of the Human Factors Society 32nd Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 102–106.

Gugerty, L., and Tirre, W. (1997), "Situation Awareness: A Validation Study and Investigation of Individual Differences," in *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 564–568.

Gugerty, L. J. (1997), "Situation Awareness during Driving: Explicit and Implicit Knowledge in Dynamic Spatial Memory," *Journal of Experimental Psychology: Applied*, Vol. 3, pp. 42–66.

Hartel, C. E., Smith, K., and Prince, C. (1991), "Defining Aircrew Coordination: Searching Mishaps for Meaning," paper presented at the 6th International Symposium on Aviation Psychology, Columbus, OH.

Hogg, D. N., Torralba, B., and Volden, F. S. (1993), "A Situation Awareness Methodology for the Evaluation of Process Control Systems: Studies of Feasibility and the Implication of Use," 1993-03-05, OECD Halden Reactor Project, Storefjell, Norway.

Hormann, H. J., Blokzijl, C., and Polo, L. (2004), "ESSAI: A European Training Solution for Enhancing Situation Awareness and Threat Management on Modern Aircraft Flight Decks," in *Proceedings of the 16th Annual European Aviation Safety Seminar of the Flight Safety Foundation and European Regions Airline Association*, Flight Safety Foundation, Barcelona, Spain.

Horswill, M. S., and McKenna, F. P. (2004), "Drivers Hazard Perception Ability: Situation Awareness on the Road," in *A Cognitive Approach to Situation Awareness: Theory, Measurement and Application*, S. Banbury and S. Tremblay, Eds., Ashgate Publishing, Aldershot, Hampshire, England, pp. 155–175.

Kaber, D. B., Riley, J. M., Lampton, D., and Endsley, M. R. (2005), "Measuring Situation Awareness in a Virtual Urban Environment for Dismounted Infantry Training," in *Proceedings of the 11th International Conference on HCI*, Lawrence Erlbaum, Mahwah, NJ.

Kaber, D. B., Riley, J. M., Sheik-Nainar, M. A., Hyatt, J. R., & Reynolds, J. P. (2006), "Assessing Infantry Soldier Situation Awareness in Virtual Environment-based Training of Urban Terrain Operations," in *Proceedings of the International Ergonomics Association*, IEA, Maastricht, The Netherlands.

Matthews, M. D., Strater, L. D., and Endsley, M. R. (2004), "Situation Awareness Requirements for Infantry Platoon Leaders," *Military Psychology*, Vol. 16, No. 3, pp. 149–161.

McKenna, F., and Crick, J. L. (1994), *Developments in Hazard Perception*, Department of Transport, London.

Mumaw, R. J., Roth, E. M., and Schoenfeld, I. (1993), "Analysis of Complexity in Nuclear Power Severe

Accidents Management," in *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 377–381.

O'Hare, D. (1997), "Cognitive Ability Determinants of Elite Pilot Performance," *Human Factors*, Vol. 39, No. 4, pp. 540–552.

Pew, R. W. (1995), "The State of Situation Awareness Measurement: Circa 1995," in *Experimental Analysis and Measurement of Situation Awareness*, M. R. Endsley and D. J. Garland, Eds., Embry-Riddle Aeronautical University Press, Daytona Beach, FL, pp. 7–16.

Prince, C. (1998), *Guidelines for Situation Awareness Training*, Naval Air Warfare Center Training Systems Division, Orlando, FL.

Prince, C., and Salas, E. (1998), "Situation Assessment for Routine Flight and Decision Making," *International Journal of Cognitive Ergonomics*, Vol. 1, No. 4, pp. 315–324.

Robinson, D. (2000), "The Development of Flight Crew Situation Awareness in Commercial Transport Aircraft," in *Proceedings of the Human Performance, Situation Awareness and Automation: User-Centered Design for a New Millennium Conference*, SA Technologies, Marietta, GA, pp. 88–93.

Rodgers, M. D., Mogford, R. H., and Strauch, B. (2000), "Post-hoc Assessment of Situation Awareness in Air Traffic Control Incidents and Major Aircraft Accidents," in *Situation Awareness Analysis and Measurement*, M. R. Endsley and D. J. Garland, Eds., Lawrence Erlbaum Associates, Mahwah, NJ.

Rouse, W. B., and Morris, N. M. (1985), "On Looking into the Black Box: Prospects and Limits in the Search for Mental Models," DTIC AD-A159080, Center for Man–Machine Systems Research, Georgia Institute of Technology, Atlanta, GA.

RTCA (1995), *Report of the RTCA Board of Directors Select Committee on Free Flight*, RTCA, Washington, DC.

Secrist, G. E., and Hartman, B. O. (1993), "Situational Awareness: The Trainability of Near-Threshold Information Acquisition Dimension," *Aviation, Space and Environmental Medicine*, Vol. 64, pp. 885–892.

Sexton, G. A. (1988), "Cockpit-Crew Systems Design and Integration," in *Human Factors in Aviation*, E. L. Wiener and D. C. Nagel, Eds., Academic, San Diego, CA, pp. 495–526.

Smith, K., and Hancock, P. A. (1995), "Situation Awareness Is Adaptive, Externally Directed Consciousness," *Human Factors*, Vol. 37, No. 1, pp. 137–148.

Strater, L. D., Jones, D., and Endsley, M. R. (2003), "Improving SA: Training Challenges for Infantry Platoon Leaders," in *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 2045–2049.

Strater, L. G., et al. (2004), "PC-Based Tools to Improve Infantry Situation Awareness," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 668–672.

Taylor, R. M. (1990), "Situational Awareness Rating Technique (SART): The Development of a Tool for Aircrew Systems Design," in *Situational Awareness in Aerospace Operations*, AGARD-CP-478, NATO-AGARD, Neuilly sur Seine, France, pp. 3/1–3/17.

Taylor, R. M., and Selcon, S. J. (1994), "Situation in Mind: Theory, Application and Measurement of Situational Awareness," in *Situational Awareness in Complex Settings*, R. D. Gilson, D. J. Garland, and J. M. Koonce, Eds., Embry-Riddle Aeronautical University Press, Daytona Beach, FL, pp. 69–78.

Tenney, Y. T., Adams, M. J., Pew, R. W., Huggins, A. W. F., and Rogers, W. H. (1992), "A Principled Approach to the Measurement of Situation Awareness in Commercial Aviation," NASA Contractor Report 4451, NASA Langley Research Center, Langley, VA.

Wickens, C. D. (1992), *Engineering Psychology and Human Performance*, 2nd ed., HarperCollins, New York.

# CHAPTER 20

# AFFECTIVE ENGINEERING AND DESIGN

**Martin G. Helander**
**Nanyang Technological University**
**Singapore**


**Halimahtun M. Khalid**
**Damai Sciences Sdn Bhd**
**Kuala Lumpur, Malaysia**

## 1 INTRODUCTION

Recently there has been a rapid growth in affective and pleasurable engineering and design. This development is in great contrast to earlier design traditions in engineering design. The affective and emotive aspects of design have largely been ignored. One exception was Titchener (1910), who considered pleasure an irreducible fundamental component of human emotion.

Design decision making and cognition were considered first by Herbert Simon's (1969) book *The Science of the Artificial*. His expression "The proper study of mankind is the science of design" remains a challenge, since ergonomics has not produced much research in structuring design problems, perhaps because design problems are often ill-defined (Goel and Pirolli, 1992). This presents a challenge also for the present authors because our focus is primarily limited to human factors; we will need to draw on articles published from other fields, and our view may not have been as comprehensive as it could have been.

The term *affect* has different meanings in psychology and human factors. Affect is the general term for the judgmental system, and *emotion* is the conscious experience of affect. Much of human behavior is subconscious, beneath conscious awareness (Leontjev, 1978). Consciousness came late in the evolution of humans and also in the way the brain processes information (Norman, 2005). The affective system makes quick and efficient judgments which help in determining if an environment is dangerous—shall I fight or flight? For instance, I may have an uneasy feeling (affect) about a colleague at work, but I don't understand why, since I am not conscious about what I am reacting to.

*Pleasure*, on the other hand, is a good feeling coming from satisfaction of homeostatic needs like hunger, sex, and bodily comfort (Seligman and Csikszentmihalyi, 2000). This is differentiated from *enjoyment*, which is a good feeling coming from breaking through the limits of homeostasis of people's experiences, for example, performing in an athletic event or playing in a string quartet. Enjoyment could lead to more personal growth and long-term happiness than pleasure, but people usually prefer pleasure over enjoyment, maybe because it is less effortful. Although each discipline has a unique definition, their goals are quite similar. We elaborate further in a later part when we discuss relevant theories of affect and pleasure.

Affective human factors design has a great future because it sells. Examples include affective design of mobile phones (Khalid, 2004; Seva and Helander, 2009) and cars (Khalid et al., 2011). But much more research is needed in this emerging field. Advances in psychology research on affect were elegantly summarized by Kahneman et al. (1999) in their edited volume *Well-Being: Foundations of Hedonic Psychology.* Helander et al. (2001) published the first conference proceedings on affective human factors design. Then Jordan (2002) wrote on pleasurable design while Norman (2005) later wrote on emotional design. In human–computer interaction (HCI), there is the book *Affective Computing* by Picard (1997) and a review by Brave and Nass (2003). *Funology* is a new trend in HCI design (Carroll, 2004; Blythe et al., 2004; Bardzell, 2009). *Hedonomics,* a new term coined by Hancock (2000), has entered human factors engineering (Helander and Tham, 2003; Khalid, 2004; Hancock et al., 2008). Monk et al. (2002) noted that design of seductive and fun interfaces is one important challenge in theory as well as in application.

The expression of emotions is important in product semantics and design. The question of which emotions are invoked naturally follows the question of what the artifact could or would mean to the users (Krippendorff, 2006). In emotional design, pleasure and usability should go hand in hand as well as aesthetics, attractiveness, and beauty (Norman, 2005). The interplay between user-perceived usability (i.e., pragmatic attributes), hedonic attributes (e.g., stimulation, identification), goodness (i.e., satisfaction), and beauty was considered in the design of MP3 player skins (Hassenzahl, 2004, 2010). Hassenzahl found that satisfaction depended on both perceived usability and hedonic attributes.

Jordan (2002) noted that a product or service offering should engage the people for whom it is designed at three abstraction levels: First, it has to be able to perform the task for which it was designed. For example, a car has to be able to take the user from point A to point B. The product's functionality should work well and it should be easy to use. The second level relates to the emotions associated with the product or service. These emotions are part of the "user experience." For example, when using an automated teller machine, feelings of trust and security might be appropriate. Driving a sports car should be exciting, but there should also be a sense of safety. The third level reflects the aspirational or inspirational qualities of the product or service. What does owning the product or using the service say about the user? For example, owning the latest, smallest mobile phone may suggest a "pretty cool" person. These observations make a case for ergonomics as well as for emotional design and social status.

Our premise is that people have affective reactions toward tasks, artifacts, and interfaces. These are caused by design features that operate either through the perceptual system (looking at) or from a sense of controlling (touching and activating) or from reflection and experience. Affective reactions are difficult if not impossible to control; the limbic system in the brain is in operation whether we want it or not. They are in operation whenever we look at objects (beautiful or ugly) and

they are particularly obvious when we try "emotional matching," such as buying clothes for a friend or selecting a birthday card for someone close to us.

Approaches to emotions and affect have been studied with the purpose of understanding (1) how one can measure and analyze human reactions to affective and pleasurable design and (2) how one can produce affective design features of products. Affect is said to be the customer's psychological response to design details of a product, while pleasure is the emotion that accompanies the acquisition or possession of something good or desirable (Demirbilek and Sener, 2003).

Desmet (2003) proposed that products can elicit four categories of emotional responses: instrumental, aesthetic, social, and surprise. *Instrumental emotions* address the use of the product. *Aesthetic emotions* relate to the beauty of a product. *Social emotions* are a result of the product being admired by a group of users. *Surprise emotions* relate to novelty in a design which can amaze users. Each of these emotions results from an appraisal or user experience of a product. With regard to visual perception, this appraisal is based on the aesthetic impressions; the pleasure in using a product will be affected by the pleasure in use; the pleasure of owning the product depends on the importance of the product to the owner. Hence, there are both semantic interpretations of product use (what does it mean to me) and symbolic associations (e.g., pride) that come with ownership (what does my ownership mean to others).

According to Dong et al. (2009), there are three different aspects to consider in affective design: aesthetic impression, semantic interpretation, and symbolic association. *Aesthetic impression* conveys a message about how the product is perceived in terms of attractiveness. This corresponds to Norman's (2005) "visceral level" in design. *Semantic interpretation* relates to what message a product communicates about its function, mode of use, and qualities. This corresponds to Norman's "behavioral level" in design. *Symbolic association* sends a message about the owner or user. It is the personal and social significance attached to the product and its design.

## 1.1 Neurological Basis of Emotions

In the brain the thalamus directs sensory information to the neocortex, which is the thinking part of the brain. The neocortex is the top layer of the cerebral hemispheres. It is 2–4 mm thick and is part of the cerebral cortex. The cortex routes signals to the amygdala (the "emotional brain") for the proper emotional reaction. The amygdala then triggers excretion of peptides and hormones which create emotion and action. However, if there is a potential threat, the thalamus will bypass the cortex and signal directly to the amygdala, which is the trigger point for the primitive fight-or-flight response. When the amygdala feels threatened, it can react irrationally and destructively (Phelps, 2006). Goleman (1995) noted that emotions make us pay attention immediately. There is a sense of urgency and an action plan is prepared without having to think twice. The emotional component evolved very early in the development of mankind. For someone who is being threatened, the

emotional response can take over the rest of the brain in a split second.

The amygdala exhibits three signs: strong emotional reaction, sudden onset, and postepisode realization that the reaction was inappropriate. The soccer player Zinedine Zidane of France head butted Marco Materazzi of Italy in the 2006 World Cup Soccer finals. As a result, France lost the World Cup to Italy and Zidane's career ended in disgrace; his surprising and aggressive response demonstrated the three signs of "amygdala hijack," that is, strong emotional reaction, sudden onset, and regret for your actions when you reflect later. Figure 1 illustrates the neurological mechanisms of the amygdala.

In Figure 1, there are three main areas: the thalamus, the amygdala, and the neocortex. The thalamus receives sensory input from the environment, which is then sent to the cortex for fine analysis. It is also sent to the limbic system, the main location for emotions, where the relevance of the information is determined (LeDoux, 1995). The amygdala evaluates the relevance of much of the information. There are two principal routes to the amygdala (LeDoux, 1993). The most common route passes through the cerebral cortex. The information from the senses will then arrive at the thalamus, and from there it goes to the corresponding primary sensory cortex, which will extract auditory, visual, and tactile information. The stimulus is then elaborated in different parts of the associative cortex, where complex characteristics and global properties are analyzed. The results are sent to the amygdala, as well as to the areas associated with the hippocampus, which is a neighboring structure to the amygdala and communicates directly with the amygdala. As the amygdala receives this information, it will evaluate the desirability or danger of the stimulus. For example, the sight of a tiger at close quarters is a highly alarming stimulus if we are in the jungle but completely harmless if it is in a cage at

the zoo. This context-related information appears to be provided by the hippocampus.

The limbic system coordinates the physiological response and directs the attention (in the cortex) and various cognitive functions. Primitive emotions (e.g., the startle effect) are handled directly through the thalamus–limbic pathway. In this case the physiological responses are mobilized, such as for fight and flight. Reflective emotions, such as pondering over a beautiful painting, are handled by the cortex. In this case physiological responses are not necessary; the situation is harmless and there is no requirement to deal any further with the situation. According to Kubovy (1999), these types of *pleasures of the mind* do not give rise to a physiological response or to any facial expressions.

## 1.2 Affective Design

"The reptilian always wins" (Rapaille, 2006). An object with a strong identity—good quality, design and functionality—makes a reptilian. However, many affective needs are unspoken and need to be identified by probing user emotions. To increase a product's competitiveness, emotions are incorporated into design (Helander and Khalid, 2006).

Designers are expanding the semantic approach to design by utilizing affective design parameters. By so doing, objects take on meanings that were previously not present. Indeed, semantic design prescribes that objects should have a meaning that goes beyond their functional outlook (Krippendorff, 2006). Designers need to identify components of affective as well as functional design and integrate these components in product design. Functional requirements are easy to understand, but affective requirements are subtle and puzzling and often difficult to identify. They vary over time, and customers often have difficulties in explaining
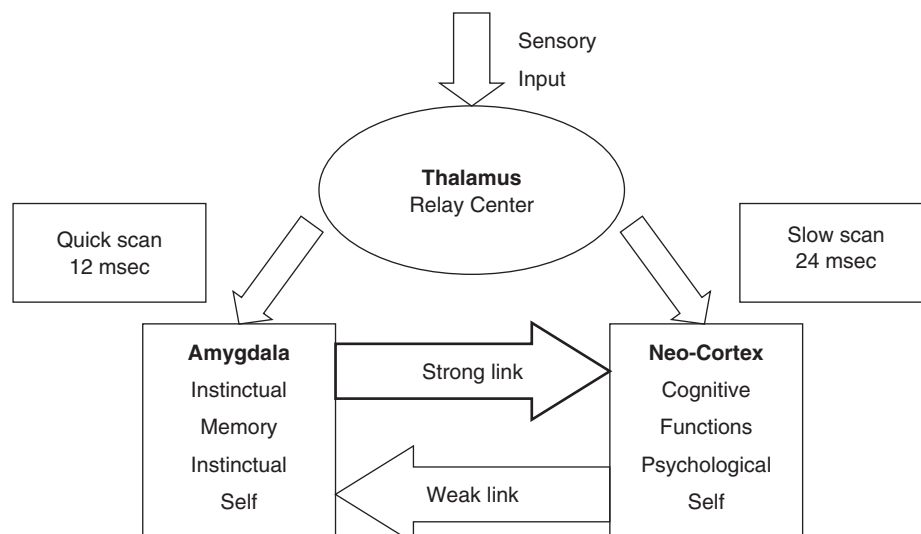


**Figure 1** Role of amygdala.

what the requirements are (Khalid et al., 2011). With the growing emphasis on emotional design, some companies have created design which may not directly cater to customer needs but rather support the company's standing as a premier design house. Hence, there are many reasons for affective design.

## 1.3 Integration of Affective and Cognitive Systems

Cognition must consider affect or emotion and vice versa; human decision making and behavior are guided by cognition as well as emotions. Figure 2 denotes the relationship between affect and cognition. Whereas affect refers to feeling responses, cognition is used to interpret, make sense of, and understand user experience. To do so, symbolic, subjective concepts are created that represent the personal interpretations of the stimuli. Cognitive interpretations may include a deeper, symbolic understanding of products and behaviors.

One of the most important accounts of affect in decision making comes from Damasio (1994). In his book *Descartes' Error*, he described observations of patients with damage to the ventromedial frontal cortex of the brain. This left their intelligence and memory intact but impaired their emotional assessments. The patients were socially incompetent, although their intellect and ability to analyze and reason about solutions worked well. Damasio argued that thought is largely made up from a mix of images, sounds, smells, words, and visual impressions. During a lifetime of learning, these become "marked" with affective information—positive or negative feelings. These *somatic markers* are helpful in predicting decision making and behavior.

Affect plays a central role in dual-process theories of thinking, knowing, and information processing (Epstein, 1994). There is hence much evidence that people perceive reality in at least two ways; one is affective (intuitive and experiential) and the other is cognitive (analytical and rational). Formal decision making must consider a combination of affective and cognitive factors; the affective system is much quicker than the cognitive system. When a person seeks to respond to an emotional event, he or she will search the affective system automatically for past experiences. This is like searching a memory bank for related events, including their emotional contents (Epstein, 1994), see Figure 3.

Emotions do not cause thinking to be nonrational; they can motivate a passionate concern for objectivity, such as anger at injustice. There is a cross-coupling so that rational thinking facts as well as feelings and affective thinking entail cognition. Rational thinking is perhaps more precise, comprehensive, and insightful than nonrational thinking. However, it is just as emotional.

Separating emotion from cognition has been considered a major weakness of psychology and cognitive science (Vygotsky, 1962). Recent research using functional magnetic resonance imaging (fMRI) has validated the assertions that cognition and emotions are at least partly unified and contribute to the control of thought and behavior (Vul et al., 2009). William James wrote about this already in the 1890s. However, there are some exceptions. If a person is facing fear, the bodily reaction may come first, for example in a tsunami disaster (Khalid et al., 2010). The person will start running instinctively before there is any cognitive evaluation. The order of response would be (1) affect, (2) behavior, and (3) cognition.

Cognition also contributes to the regulation of emotion. Contemporary views in artificial intelligence and psychology are embracing an integrated view of emotion and cognition. In *Emotion Machine*, Minsky (2007) noted that the traditional idea that ''thinking" is contaminated by "emotions" is not correct. He claimed that emotions and thinking cannot be separated; they are unified.

Combining the description from contemporary psychology and neuroscience, Camerer et al. (2005) illustrated the two distinctions between controlled and automatic processes (Schneider and Shiffrin, 1977) and between cognition and affect, as in Table 1.

As described in Table 1, controlled processes have several characteristics. They tend to (1) be serial (employing a step-by-step logic or computations); (2) be invoked deliberately by an agent when encountering a challenge or surprise; (3) be associated with a subjective feeling of effort; and (4) typically occur consciously. As such, people have reasonably good introspective access
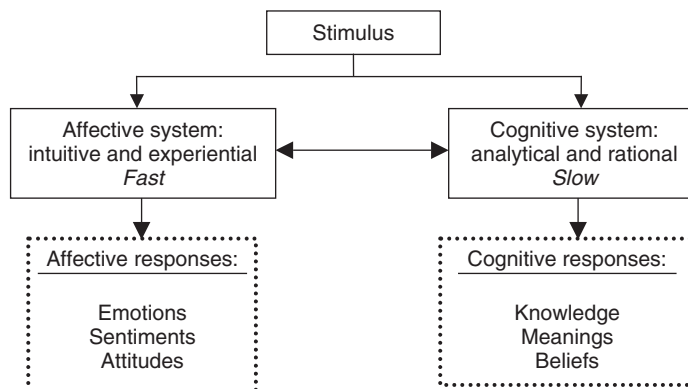


**Figure 2**　Integration of affect and cognition.
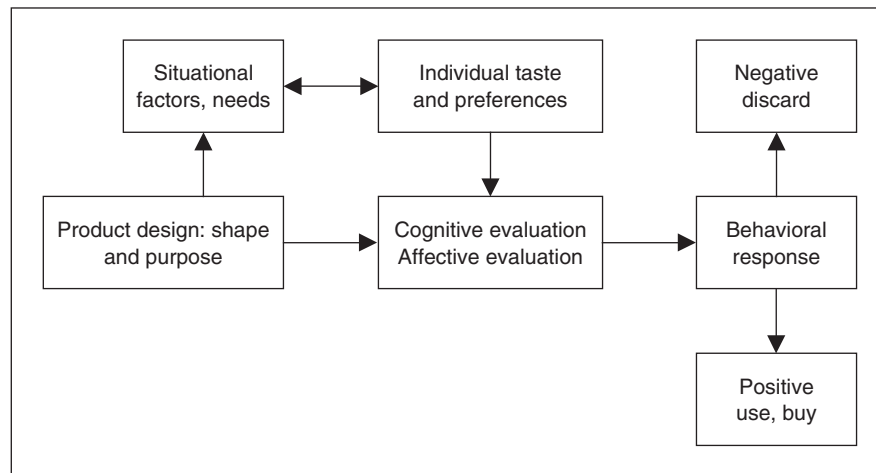
**Figure 3** Comparison and evaluation of product shape and individual preferences.

**Table 1  Two-Dimensional Characterization of Neural Functioning**

| Type of Process | Cognitive | Affective |
|---|---|---|
| Controlled Processes<br>• Serial<br>• Evoked deliberately<br>• Effortful<br>• Occurs consciously | I | II |
| Automatic Processes<br>• Parallel<br>• Effortless<br>• Reflexive<br>• No introspective access | III | IV |

*Source*: Camerer et al. (2005)

to controlled processes. If they are asked how they solved a math problem or chose a new car, they can usually provide a good account of the decision-making process.

Automatic processes are the opposite of controlled processes. Automatic processes tend to (1) operate in parallel; (2) not be associated with any subjective feeling of effort; and (3) operate outside of conscious awareness. As a result, people often have little introspective access as to why the automatic choices or judgments were made. For example, a product can be perceived automatically and effortlessly as "attractive"; it is only in retrospect that the controlled system may reflect on the judgments and try to substantiate it logically.

The second distinction, represented by the two columns of Table 1, is between cognitive and affective processes. This distinction is pervasive in psychology (e.g., Zajonc, 1998) and neuroscience (Damasio, 1994; LeDoux, 1995). Zajonc (1998) defined *cognitive processes* as those that answer true–false questions and *affective processes* as those that motivate approach–avoidance behavior. Affective processes

include emotions, such as anger, sadness, and shame, as well as "biological affects" such as hunger, pain, and the sex drive (Buck, 1999).

Elaborating this further, quadrant I, for example, is in charge when one considers purchase of an expensive machine. Quadrant II can be used by "actors," such as salespersons, who replay previous emotional experiences to motivate customers that they are experiencing these emotions. Quadrant III deals with motor control and governs the movements of the limbs, such as a tennis player when he or she returns a serve. Quadrant IV applies when a person wins a surprising award. The four categories are often not so easy to distinguish because most behavior results from a combination of several quadrants.

## 1.4  Understanding Affect and Pleasure in Different Disciplines

There are many definitions and classifications of affect and pleasure in the literature. We mention a few that have relevance to human factors design.

***Marketing***  Peter and Olson (1996), with a background in marketing, defined four different types of affective responses: *emotions, feelings, moods,* and *evaluations*, and offered a classification (Table 2). These responses are associated with different levels of physiological arousal as well as different intensities of feeling. Emotions are associated with physiological arousal, while evaluations (e.g., reflections) of products typically encompass weak affective responses with low level of arousal.

***Product Design***  Tiger (1992) identified four conceptually distinct types of pleasure from a product. These were further elaborated by Jordan (see Blythe, et al. 2003). We extended the taxonomy to five. Whether they are used as a source for pleasure depends on the needs of the individual.

**Table 2  Types of Affective Responses**

| Type of Affective Response | Examples of Positive and Negative Affect | Level of Physiological Arousal | Intensity or Strength of Feeling |
|---|---|---|---|
| Emotions | Joy, love<br>Fear, guilt, anger | Higher arousal and activation | Stronger |
| Specific feelings | Warmth, appreciation,<br>Disgust, sadness | | |
| Moods | Alert, relaxed, calm<br>Blue, listless, bored | | |
| Evaluations | Like, good, favorable<br>Dislike, bad, unfavorable | Lower arousal and activation | Weaker |

*Source*: Peter and Olson (1996)

- *Physical or physio-pleasure* has to do with the body and the senses. It includes things like feeling good physically (e.g., eating, drinking), pleasure from relief (e.g., sneezing, sex), as well as sensual pleasures (e.g., touching a pleasant surface).

- *Socio-pleasures* include social interaction with family, friends, and co-workers. This includes the way we are perceived by others, our persona, and our status.

- *Psychological pleasure* has to do with pleasures of the mind—reflective as well as emotional. It may come from doing things that interest and engage us (e.g., playing in an orchestra or listening to a concert), including being creative (e.g., painting) or enjoying the creativity of other people.

- *Reflective pleasure* has to do with reflection on our knowledge and experiences. The value of many products comes from this and includes aesthetics and quality.

- *Normative pleasure* has to do with societal values such as moral judgment, caring for the environment, and religious beliefs. These can make us feel better about ourselves when we act in line with the expectation of others as well as our beliefs.

Jordan (1998) defined pleasure with products as the emotional and hedonic *benefits* associated with product use. Coelho and Dahlman (2000) defined displeasure as the emotional and hedonic *penalties* associated with product use. Chair comfort, for example, has to do with feeling relaxed, whereas chair discomfort has to do with poor biomechanics. The two entities should be measured on different scales as they are two different dimensions (Helander and Zhang, 1997). Both discomfort and displeasure operate like design constraints; we know what to avoid but that does not mean we understand how to design a pleasurable product. Fixing poor biomechanics

and getting rid of displeasure do not necessarily generate a sense of relaxation and comfort.

With increasing age and personal experiences, our repertoire of emotions expands. Researchers believe that only the startle reflex is innate, while most emotions are learned over time, especially sentiments. People are often attracted to complexity in music and in art. One can listen to a piece of music many times; each time one discovers something new. Likewise in a painting; many modern paintings are difficult to comprehend—each time you look, the interpretation changes. Some pleasures are hard to appreciate, and therefore they become challenging.

## 2  FRAMEWORK FOR AFFECTIVE DESIGN

Design is a problem-solving discipline. Design addresses not only the appearance of the designed product but also the underlying structure of the solution and its anticipated reception by users. A design theory helps designers to identify the problem and develop their "instincts" in selecting the right solutions (Cross, 2000).

A product design represents a solution to a set of design goals and constraints which are formulated by the designer. The design constraints include performance objectives constraints, ergonomic constraints, product and cost constraints, and regulatory and legal constraints (Ullman, 2009).

### 2.1  How Designers Design

Crilly et al. (2009) interviewed 21 UK-based industrial designers with significant professional experience and they also held senior design positions. The purpose was to investigate the goals and procedures that drive designers of products. The outcome of the research is summarized in Table 3. Clearly, a designer's choice of strategies would depend on the type of product and customer requirements. For example, in emotional design, designers elicit emotional responses in consumers by designing products that surprise, satisfy, or delight.

**Table 3  Design Strategies among Product Designers**

| Stages | Strategies |
| --- | --- |
| 1. Attention | Draw consumer attention away from other alternatives. |
| 2. Recognition | Incorporate product brand, tradition, and style since consumers recognize this. |
| 3. Attraction | Make product attractive/elegant. There are no formulas. It is solely intuitive design. |
| 4. Comprehension | Improve usability by using product form to inform customers how it should be used or how it works. |
| 5. Attribution | Carefully manipulate product form to enhance impressions of usability and reliability. |
| 6. Identification | Generate forms that imply lifestyles. |
| 7. Emotion | Elicit emotional responses in consumers by designing products that surprise, satisfy, or delight. |
| 8. Action | Encourage purchase and usage behaviors which promote consumer satisfaction and commercial success. For example, buy and then frequently update a model for children's toothbrush. |

*Source*: Crilly et al. (2009).

Until recently, the affective aspects of design and design cognition were substantially absent from formal theories of design. However, as Rosenberg and Hovland (1960) noted 50 years ago, affect is a prerequisite to the formation of human beliefs, values, and judgments, and design models that do not include affect (emotions, feelings) are inadequate.

Much of the ability to render attractive products is attributed to the personal experiences and creativity of designers. It is a question of making the design as simple and as clean and aesthetic as possible; much hard work goes into creating a clean image.

A product design represents a solution to a set of design goals and constraints which are realized by the designer (Bloch, 1995), as mentioned above.

## 2.2  Consumer Process

The process of buying a product is influenced by two affective processes: (1) affective matching of needs and (2) affective matching of personal utility; see Figure 4.

In the first instance a person tries to match features of several alternative products to his or her perceived needs. At the same time the customer has constraints that eliminate many products due to price, suitability, and aesthetics design. Assume that you are buying a shirt or a blouse for a friend. You will consider the price, size, style, and color. You will try to imagine how well it fits to his or her personal needs and if the shirt will be appreciated. This "emotional matching" also occurs when you buy a shirt or a blouse for yourself, except that the process is more automatic, and you may not consciously reflect on all the details. So, the personal evaluation process is quicker and sometimes subconscious. While you are aware of why you like something, you may not have reflected on why you reject an item. Consider, for example, going through a rack of blouses or shirts in a store. The rejection of an item takes only a second. The affective matching of a blouse is a pattern-matching process with well-developed criteria

for aesthetics and suitability. The constraint filter helps in decision making by eliminating alternatives. It operates in a similar fashion to the decision heuristic "elimination by aspect" (Tversky, 1972). Some blouses are rejected at an early stage. This can happen for many reasons, such as high price, ugly color, and poor quality. A quick decision is then made to reject the product and consider the next (Seva and Helander, 2009).

If the blouse is accepted, there will be a trial adoption. A customer may try the blouse. A second affective matching takes place, where the personal utility and the benefit–cost ratio of the purchase are judged. There can be three decision outcomes: reject (search for another blouse), accept (pay and leave), or give up (walk out of the store).

Based on the sales pattern and customer surveys, the product mix in a store will be modified. However, for a new product, it may not be possible to predict customer emotions and the sales. Below we focus on methods for measuring emotion response to artifacts, an important factor in determining the success of a product.

## 2.3  Affective User–Designer Model

Taking the designer and consumer into consideration, the systems model in Figure 5 provides a framework for issues that must be addressed in affective engineering and design (Helander et al., 2010), based on a novel concept called *citarasa*.

*Citarasa* originated from Sanskrit; the term is widely used in regions where the Malay language is spoken. The word *Cita* means "intent, aspiration," while *Rasa* means "feelings, taste" (Khalid, 2006). Citarasa presupposes that people have an *emotional intent*, that is, an explicit goal-driven desire, when they want a product, and design must therefore address customers' emotional intent and affective needs.

To conceptualize the framework, we used Khoo's (2007) investigation of how emotional intent developed when a person bought a new car. First, he made contacts
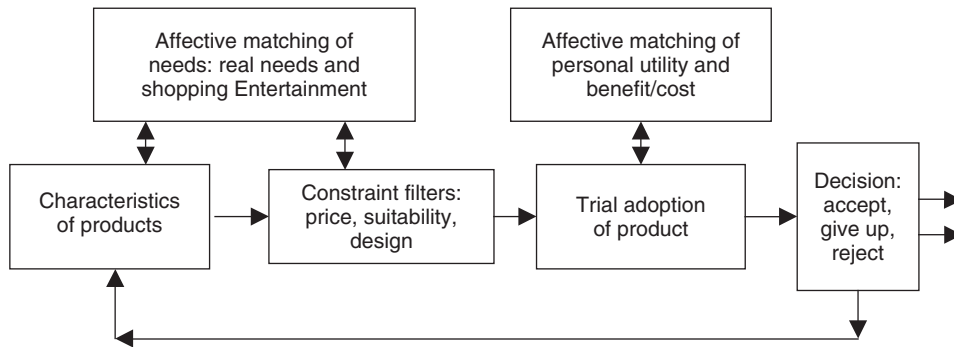
**Figure 4** Consumer process.

with car buyers in a showroom in Singapore. Using cognitive task analysis (Crandall et al., 2006), he probed their reasons for buying a car to identify tacit information such as buyer expertise. Each customer was interviewed several times during a six-month period in three stages:

- *Stage* 1. Initially, a customer maintains beliefs about different cars. This stage deals primarily with *affective requirements*. A customer first talks to his friends and reads technical reviews and literature about cars. He may consider the "dream car" (such as a Porsche) but soon realizes that this would be unrealistic.

- *Stage* 2. About six months before the purchase he visits showrooms and test drives cars. This is where the researchers made contacts with the customers. During this stage customers talk to their family about functional needs—what the car will be used for, how many persons travel, fuel consumption, and anticipated repair record. This stage basically deals with *functional requirements*.

- *Stage* 3. Just before making the purchase, the buyer will consider the "quality" of the car, and several customers upgraded their purchase to a car with higher price in order to improve quality. However, the higher price also brings greater prestige, and it was difficult to distinguish which of the two motives inspired the car buyer. This stage combines affective and functional requirements.

During this period customers went through motivational stages, following the model by Fishbein and Ajzen (1975): from belief (about personal needs including luxury) to attitude (personal preferences) to intention (sorting out the real needs of the family) to behavior (purchase).

With this information we developed a model of emotional intent as shown in Figure 5. To investigate intent we identified customers' functional and affective requirements when buying a product such as a car. A need for an "elegant" car can relate to several interior or exterior characteristics, such as color, shape, size, and capacity that can be manipulated by designers to satisfy customer needs.

Citarasa descriptors were generated from the visceral, behavioral, and reflective needs of the customer (Norman, 2005; Helander and Khalid, 2009). The needs were elicited by probing the customer's intent using a laddering technique called the why-why-why interview method (Goh, 2008; Helander et al., in press; Khalid et al., 2011).

In Figure 5, there are two subsystems: the designer environment (for product development) and the customer/user environment (for evaluation of functional and affective design factors).

***Designer Environment***   In designing a new vehicle, a product planner/designer uses information from a variety of sources, including *marketing, context of use/activity* of a vehicle, and *society norms and fashions*. This information broadly determines the design goals (what one should design) and the constraints (what one should not design). The Marketing Department will offer information about customer needs and future markets which will determine design goals. In addition, Marketing will often determine economic parameters, including the sales price, thereby imposing important design goals and constraints.

Affect is elicited not only by perceptual features of the product but also by the activity with the product and context of use. Hence, product planners must analyze the context of activity, that is, primarily driving. This will determine not only affective requirements but also implicitly functional requirements of the vehicle. Some design factors are influenced by social norms and fashions. For example, functional requirements concerning fuel efficiency will be regulated by governments as well as by customers. After they are adopted by Marketing, they appear to the designer as design constraints. Trends and fashions in design are important and play a decisive role in affective design.

Product planners and designers need to understand how affective design can be derived. Norman (2005) noted that there are three types of affective design features—visceral, behavioral, and reflective:
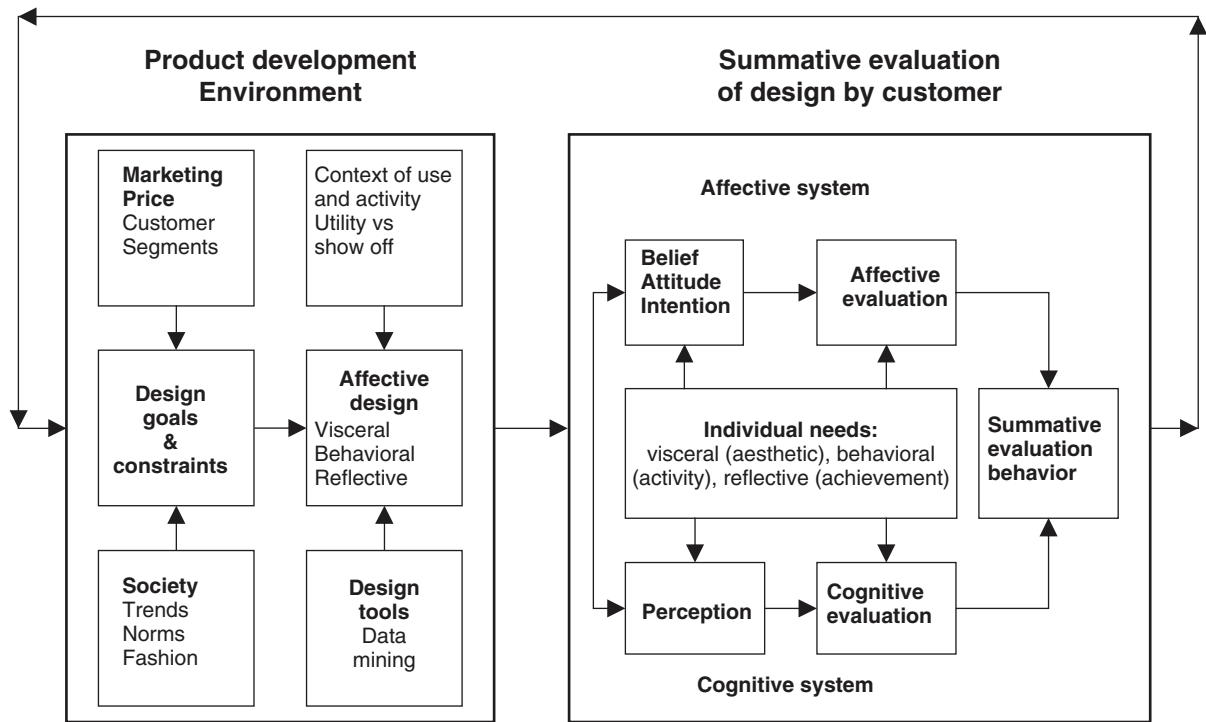
**Figure 5**   Framework for evaluation of affective engineering and design (Helander and Khalid, 2009).

- *Visceral design* refers to the visual aspects of the design, such as shape, color, materials, ornamentation, and texture. Visceral design is applied to many designed objects with aesthetic value. For vehicle design, there are several basic factors, including the vehicle exterior and interior and elements such as dashboard, seats, and controls. The information on affective needs of buyers is then translated into affective and functional requirements by designers.

- *Behavioral design* has to do with the pleasure in using the object, such as steering a smooth and well-balanced vehicle, driving at a very high speed, and intuitively finding controls. Steering a vehicle with a well-designed steering system along a curvy road can similarly be very satisfying.

- *Reflective design* has to do with things that have been learned over the years. A vehicle buyer may take much interest in an aesthetic design, because it was a tradition in the family to be surrounded by beautiful objects from early childhood. The interest and preferences for aesthetic design are learned and will typically increase with age (Norman, 2005).

There are also cultural differences in learned preferences, for example, Chinese will buy red items because it symbolizes happiness, while Indians associate red with power and strength (Helander et al., 2007; Chang et al., 2006). Designers should consider first-hand customer's conscious reflective needs, which would comply with the current fashion trends and a person's cultural background. Reflective design assumes that the emotional intent (affective requirements) drive customer choice, together with functional requirements. Customers reflect on the suitability of various design options and select a vehicle that suits both types of requirements.

Design features of a car will be addressed by the designer one by one, whereas the customer uses a holistic evaluation—for example, "I like it" without reflecting on what exactly triggered this reaction. It is important that designers are well informed about these different and complementary design options and understand how to implement them. This will require training, so that designers can make conscious decisions and fully understand when and how to utilize visceral, behavioral, and reflective design.

***Customer/User Environment***  This subsystem is made up of a user's affective and cognitive systems. The affective system is based on the capability of the product to elicit affect. Due to uniqueness in style and personality some products are more capable of evoking affect than others (Seva and Helander, 2009). The prospect of owning such a product generates a variety of emotions that are not experienced when confronted with standardized products. In essence, deep-seated desires of users for individuality, pleasure, and aesthetics cause emotion in a user's evaluation of the product.

Products also elicit cognitive responses, appraised according to fulfillment of identified functions. For example, a car must not only be capable of transportation but must also be reliable. In the process of evaluation, the user employs previous knowledge to determine if the product is acceptable to one's standard or not. There are very clear criteria that have been set forth by the user that can be rationally measured to arrive at a decision. The cognitive system shown in Figure 5 follows the human information-processing model that explains the psychological processes involved in interacting with a system. The system begins with the perception of the artifact's attributes. The attributes trigger cognition and memory recall. Information stored in memory about a product is retrieved and used for evaluation and decision making. When evaluating a car, for example, knowledge obtained from past experiences, published material, and other customers' opinion come to mind and are used to make the best decision. The decision is then used as a basis for one's action—to make or forego a purchase.

Customers' individual needs influence information processing. A person's attention is attracted by visual items that stand out because of color, brightness, and size (Triesman and Souther, 1985). Attention is also drawn by unique design features that lead customers to feel awe, surprise, and excitement. Emotions are experienced because some needs are fulfilled just by the sight of a product or the prospect of owning it. Visual aesthetics affect the consumer's perception of a product in many ways and influence the evaluation of a product (Bloch et al., 2003). Like aesthetics, achievement drives people's emotion and influence perception. People feel pleasure at the thought of accomplishing something worthy (Kubovy, 1999). The need for achievement is comparable to the need to satisfy one's curiosity by gathering more information. A customer who wants to buy a car or a truck suddenly becomes aware of the models and brands that are available when he or she had been oblivious of this information in the past.

The need for power can draw a person to products that can enhance his or her image. Color is one product attribute that elicits strong emotion and association (Chang et al., 2006). It may communicate complex information and symbolism as well as simple messages. Automobile design is an area where color seems to have fairly consistent associations. In Western countries, black is a color associated with status and sophistication. We found black to be associated with "elegance" in Europe as well as Asia (Khalid et al., 2011).

To summarize, pleasure with products is viewed from three theory-based perspectives: (1) the context of use and activity; (2) categories of pleasure with products, including visceral, behavioral, and reflective; and (3) the centrality of human needs structure in driving both the cognitive and affective evaluation systems. With reference to Figure 5, pleasure with products should be considered in the context of product use—the activity context. The same product can bring forth different levels of pleasure, depending on the goals and expectations of the user and the activity that is being performed.

Users' requirements of designed products have frequently been compared to Maslow's hierarchy of needs. This suggests that, once issues of utility, safety, and

comfort have been satisfied, emphasis may shift toward the decorative, emotional, and symbolic attributes of design (Yalch and Brunel, 1996). Therefore, depending on motivation and context, a product's perceived attributes may be of greater importance than what its actual performance would suggest. This is because appearance is important; consumers do not just buy a product, they buy value in the form of entertainment, experience, and identity.

## 3 THEORIES OF AFFECT AND PLEASURE

Several theories in psychology support the notions that we have raised, while some provide directions for future research and methods development. These theories are summarized below.

### 3.1 Activity Theory

Activity theory employs a set of basic principles and tools—object orientedness, dual concepts of internalization/externalization, tool mediation, hierarchical structure of activity, and continuous development—which together constitute a general conceptual system (Bannon, 1993). In human activity theory, the basic unit of analysis is human (work) activity. Human activities are driven by certain needs, where people wish to achieve a certain purpose (Bannon and Bødker, 1991). The activities are usually mediated by one or more instruments or tools, such as a photographer using a camera. Thus, the concept of mediation is central to activity theory.

Leontjev (1978) distinguished between three different types of cognitive activities: (1) simple activity, which corresponds to automated stimulus–response; (2) operational activity, which entails perception and an adaptation to the existing conditions; and (3) intellectual activity, which makes it possible to evaluate and consider alternative activities. Note that these activities are in agreement with Rasmussen's model of skill-based, rule-based, and knowledge-based behavior (Rasmussen et al., 1994). For each of the cognitive stages there are corresponding emotional expressions: affect, emotion, and sentiments.

*Affect* is an intensive and relatively short-lasting emotional state. For instance, as I walk down colorful Orchard Road in Singapore and look at items displayed in the shop windows, there are instantaneous reactions to the displayed items, most of these reactions are unconscious, and I have no recollection of them afterward. Through affect, we can monitor routine events. Many events are purely perceptual and do not require decision making, but there is an affective matching of events that are stored in memory (see Figure 4). This helps in understanding and interpreting their significance.

*Emotions* are conscious. When I stop to look at some item in one of the shopping windows, I am aware of why I stopped. Emotions go beyond the single situation and typically remain in memory for one or several days.

*Sentiments* according to Leontjev (1978) are longer lasting and include intellectual and aesthetic sentiments, which also affect my excursion along Orchard Road. I know from experience that some stores are impossible;

on the other hand, there are a few that are clearly very interesting. Sentiments are learned responses.

Feelings are an integral aspect of human activity and must be investigated as psychological processes that emerge in a person's interaction with his or her objective world. Their processes and states guide people toward achieving their goals (Aboulafia and Bannon, 2004). Feelings should not be viewed merely as perturbances of underlying cognitive processes. Predicting affect is likely to be easier than predicting emotions or sentiments. To evoke affective reactions in a user, the artifact could be designed to provide people with a variety of sudden and unexpected changes (visual or auditory) that cause excitement and joy or alarm. Designing toys for children can give us ideas about such design space.

Predicting emotional responses that extend over several situations can be more difficult. Emotions are not dependent on the immediate perceptual situation. The emotional state of a computer user is not usually oriented toward the mediating device itself but to the overall activity in general (either work activity or pleasure). The artifact is merely a mediating tool between the motive and the goal of the user (Aboulafia and Bannon, 2004).

Leontjev (1978) emphasized that emotions are relevant to activity, not to the actions or operations that realize it. In other words, several work or pleasure situations influence the emotion of the user. Even a successful accomplishment of one action or another does not always lead to positive emotions. For example, the act of sneezing in itself usually evokes satisfaction. However, it may also evoke fear of infecting another person. Thus, the affective and emotional aspects of objects are capable of changing, depending on the nature of the human activity (the overall motive and goal). As such, stressed Aboulafia and Bannon (2004, p. 12), "objects or artifacts—in and of themselves—should not be seen as affective, just as objects in and of themselves should not be defined as 'cognitive' artifacts, in Norman's (1991) sense. The relation between the object (the artifact) and the human is influenced by the motive and the goal of the user, and hereby the meaning or personal sense of the action and operation that realize the activity."

We note that Norman (2005) would object to these notions. In fact, he proposed that domestic robots need affect in order to make complex decisions, and Velásquez (1998) wrote about robots that weep. Equipped with only pure logical functions, a robot would not be able to make decisions—just like Damasio's (1994) patients.

### 3.2 Emotions versus Pleasures of the Mind

Ekman (1992, 1994) stated that there are five fundamental emotions that differ from one another in important ways: anger, fear, sadness, disgust, and happiness. Evolution played an important role in shaping the features of these emotions as well as their current function.

The pleasures of the mind have been neglected by contemporary psychology (Cabanac, 1992). Kubovy (1999) argued that pleasures of the mind are different from basic emotions. Pleasures of the mind are not accompanied by any distinctive facial expression. Take, for example, a person viewing a painting in a museum.

She may feel elated, but nothing is revealed on her face, and there is no distinctive physiological response pattern. This is very different from social interaction, such as a conversation with a colleague at work, where half of the message is in the person's face. Since one may not be able to use either physiological measures or facial expressions, one is left with subjective measures. There is nothing wrong with asking people; subjective methods, interviews, questionnaires, and verbal protocols provide valuable information. The problem is what questions should be asked in order to differentiate between products?

The notion of pleasures of the mind dates back to Epicurus (341–270 B.C.), who regarded pleasures of the mind as superior to pleasures of the body because they were more varied and durable. Kubovy (1999) also noted that pleasures of the mind are quite different from pleasures of the body, for example, tonic pleasures and relief pleasures. Table 4 summarizes Ekman's eight features of emotion in the left-hand column; the right-hand column shows pleasures of the mind.

### 3.3 Reversal Theory: Relationship between Arousal and Hedonic Tone

Arousal is a general drive which has its roots in the central nervous system. According to common arousal theories, organisms fluctuate slightly about a single reference point. Reversal theory, on the other hand, focuses on the subjective experiences of humans. The central concept of reversal theory is that the preferred arousal level fluctuates (Apter, 1989). Reversal theory claims that people have two preferred points, and they

**Table 4  Features of Emotions and Pleasures of the Mind**

| Emotions . . . | Pleasures of the mind . . . |
| --- | --- |
| Have a distinctive universal signal (such as a facial expression) | Do not have a distinctive universal (facial) signal |
| Are almost all present in other primates | At least some of them may be present in other primates |
| Are accompanied by a distinctive physiological response | Are not accompanied by a distinctive physiological response |
| Give rise to coherent responses in the autonomic and expressive systems | Do not give rise to coherent responses |
| Can develop rapidly and may happen before one is aware of them | Are relatively extended in time |
| Are of brief duration (on the order of seconds) | Are usually not of brief duration |
| Are quick and brief and imply the existence of an automatic appraisal mechanism | Even though neither quick nor brief, may be generated by an automatic appraisal mechanism |

*Source*: Kubovy (1999, p. 137)

frequently switch or reverse between them. The theory therefore posits bistability rather than homeostasis. In the first state, which is called *telic*, low arousal is preferred, whereas high arousal is experienced as unpleasant. In the telic state, calmness (low arousal, pleasant) is contrasted with anxiety (high arousal, unpleasant). The opposite is true when the person is in the *paratelic* state. In the paratelic state, low arousal is experienced as boredom (unpleasant) and high arousal as excitement (pleasant).

A given level of arousal may therefore be experienced as either positive or negative. A quiet Sunday afternoon can be experienced as serene or dull. One may also experience a crowded and noisy party as exciting or anxiety provoking. The perceived level of pleasantness, called *hedonic tone*, is different for the two states. The paratelic state is an arousal-seeking state and the telic state arousal avoiding. When in the telic state, people are goal oriented; they are serious minded and try to finish their current activity to attain their goal. On the other hand, to have a good time, the paratelic state is appropriate. Goals and achievements are not of interest; rather, this is the time to play, have fun, and be spontaneous.

### 3.4 Theory of Flow

*Flow* is a state of optimal experience, concentration, deep enjoyment, and total absorption in an activity (Csikszentmihalyi, 1997). Csikszentmihalyi (1990) described that when people become totally absorbed by an activity time flies quickly. They forget other things around them and focus on the activity. He referred to this as the state of flow. Examples could be playing a game, or a musical instrument. This state is characterized by a narrowing of the focus of awareness so that irrelevant perceptions are filtered out. People focus on the goals of the task that they are performing and they perceive a sense of control over the environment.

The experience of flow is associated with positive affect—people remember these situations as pleasurable. It may be participation as a violin player in an orchestra, solving math problems, or playing chess. All of these cases may involve a sense of total attention and accomplishment which the person thinks of as a pleasurable experience.

Flow has been studied in a broad range of contexts, including sports, work, shopping, games, hobbies, and computer use. It has been found useful by psychologists, who study life satisfaction, happiness, and intrinsic motivations; by sociologists, who see in as the opposite of anomie and alienation; by anthropologists, who are interested in the phenomenon of rituals.

Webster et al. (1993) suggested that flow is a useful construct for describing human-computer interactions. They claimed that flow has the effect when users perceive a sense of control over the interactions with technology. They also perceive that their attention is focused on the interaction. As a result their curiosity is aroused during the interaction, and they find the interaction interesting.

In e-commerce, a compelling design of a website should facilitate a state of flow for its customers.

Hoffman and Novak (1996) defined flow as a mental state that may occur during network navigation. It is characterized by a seamless sequence of responses facilitated by machine interactivity. It is enjoyable and there is a loss of self-consciousness. To experience flow while engaged in any activity, people must perceive a balance between their skills and the challenges of the interaction, and both their skills and challenges must be above a critical threshold.

Games promote flow and positive affect (Johnson and Wiles, 2003). The study of games can also provide information on the design of nonleisure software to achieve positive affect. Bergman (2000) noted that the pleasure of mastery can only occur by overcoming performance obstacles, so that the user can work without interruptions. He will then obtain a sense of accomplishment. Thus, an interface may be designed that can improve a user's attention and make the user feel in total control as well as free of distractions from nonrelated tasks, including poor usability.

### 3.5 Endowment Effect

Research on the *endowment effect* has shown that people tend to become attached to objects they are endowed with, even if they did not have any desire to own the object before they got possession of it (Thaler, 1980). Once a person comes to possess a good, he or she values it more than before possessing it. This psychology works well for companies that sell a product and offer a two-week return policy. Very few return the product. Put simply, this means that people place an extra value on the product once they own it.

Lerner et al. (2004) extended the endowment effect by examining the impact of negative emotions on the assessment of goods. As predicted by appraisal-tendency theory, disgust induced by a prior, irrelevant situation carried over to unrelated economic decisions, thereby reducing selling and choice prices and eliminating the endowment effect. Sadness also carried over, reducing selling prices but increasing choice prices. In other words, the feeling of sadness produces a *reverse endowment effect* in which choice prices exceeded selling prices. Their study demonstrates that incidental emotions can influence decisions even when real money is at stake and emotions of the same valence can have opposing effects on such decisions.

### 3.6 Hierarchy of Needs

According to Maslow (1968), people have hierarchies of needs that are ordered from physiological to safety, love/belonging, esteem, and self-actualization. They are usually depicted using a pyramid or a staircase, such as in Figure 6. The hierarchy affects how needs are prioritized. Once a person has fulfilled a need at a lower level, he or she can progress to the next level. To satisfy the need for self-actualization, a person would have to fulfill the lower four needs, which Maslow (1968) referred to as *deficiency needs*. These needs are different in nature to self-actualization. Many authors have pointed out that the hierarchy is not a strict progression. For example, some individuals may deemphasize safety but emphasize the needs for love/belonging.

Hancock et al. (2005) presented a hierarchy of needs for ergonomics and hedonomics (Figure 6). The ergonomic needs address safety, functionality, and usability; in Maslow's reasoning these would be referred to as deficiency needs. The two upper levels, pleasure and individuation, deal with self-actualization. Individuation, at the top of the pyramid, is concerned with ways in which a person customizes his or her engagement and priorities, thereby optimizing pleasure as well as efficiency.

One may question if there is really a hierarchy or if the elements of Figure 6 are independent of each other. If so, the progression would not be from bottom to top but rather in parallel. Helander and Zhang (1997) found that comfort and discomfort are orthogonal concepts, and it is necessary to use two different scales to measure them. Similarly, it may be necessary to use several scales in Figure 6 to measure each of the five concepts. Essentially, a combination of subjective and objective measures is needed to capture the various dimensions of emotion.

## 4 MEASUREMENT OF AFFECT EMOTION, AND PLEASURE

The measurement of affect in human factors research is complex and challenging. Although a number of methods for measuring affect have been developed, their applicability and effectiveness in different contexts remain questionable. Evaluation of affective design in product development is much more difficult than evaluation of technical (e.g., performance requirements) and business-related (e.g., market shares, sales) matters. Affect in design is sometimes fuzzy and may be better understood from an intuitive and personal viewpoint. As such, it is difficult to evaluate affective design from an objective and systematic approach (Khalid, 2008).

However, affective design has become increasingly important in industry. The impact on customer satisfaction, sales figures, and corporate core values demands greater treatment of the issues for integration into the product development processes. Moreover, affective design measures need to be derived and communicated explicitly to the designers. For this reason, there is a need to identify methods and tools that support a systematic approach to affective design.

Several methods have been developed for measuring and analyzing affect in HCI (Shami et al., 2008; Ashkanasy and Cooper, 2008) and product design (Khalid et al., 2011; Xu et al., 2009; Hekkert and Schifferstein, 2008). However, these methods are sometimes difficult to be used directly in industry. First, knowledge on methods that are appropriate and useful for design of specific products in industry is still limited. Second, many methods are developed in academia and may not be directly adaptable to the constraints and requirements in industry. Third, most methods require knowledge of how to use them in order to be reliable and valid. Figure 7 provides an overview of the existing methods that have been employed in product design.
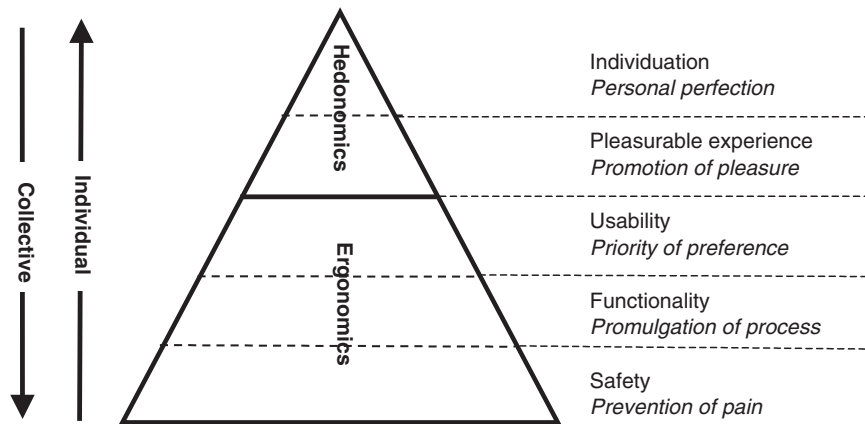
**Figure 6**   Hierarchy of ergonomics and hedonomic needs (Hancock et al., 2005).
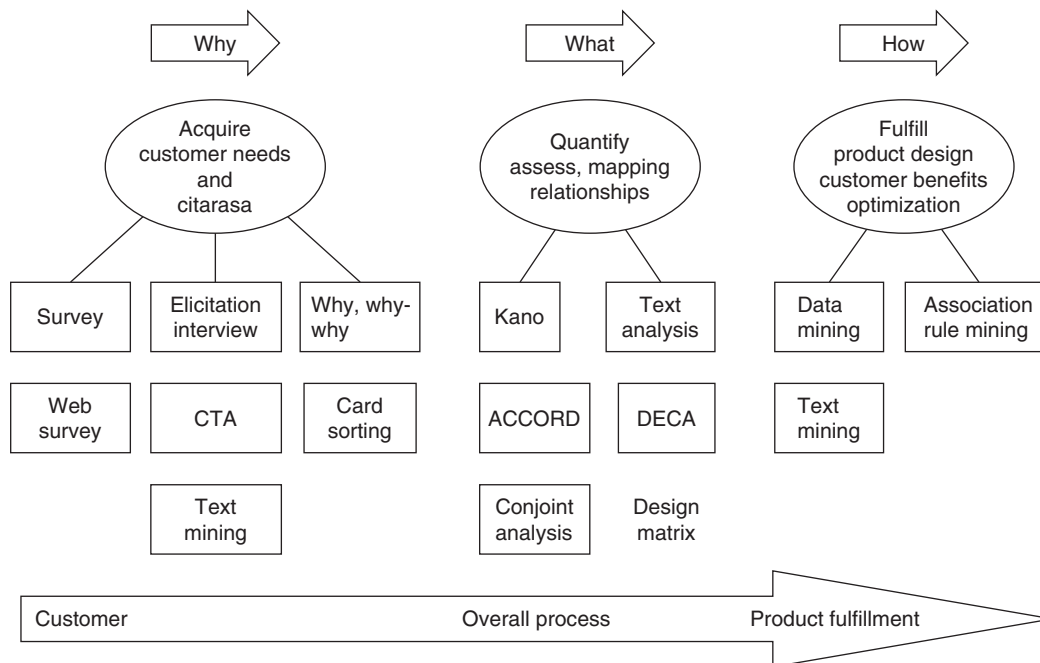


**Figure 7**   Overview of methods for affective engineering and design.

The why-why-why method has proven effective in eliciting customers' affect (Goh, 2008; Helander and Khalid, 2009; Khalid et al., 2011). The method was applied in the CATER project (Khalid et al., 2007a) to generate customers' needs and requirements as mentioned above.

### 4.1   Measurement Issues

To measure affect, emotions, or pleasure of users in relation to a product or system, we need to consider five

pertinent issues: dynamics, context, reliability, validity, and error (Larsen and Fredrickson, 1999).

### 4.1.1   Dynamics

Emotions are generated by different systems in the brain with different timing mechanisms, and they evolve over time. They are difficult to capture, which raises three critical issues that need to be addressed: (1) how to identify the onset and end of a particular emotion, (2) how to ensure that a measure of emotion can capture the duration, and (3) how to compare over

time the subjective emotion experience to the measured experience.

### 4.1.2 Context

Emotions occur in a context. Activity theory, for example, emphasizes the ongoing work activity. Therefore, it is important to capture the context and the peculiarities of the scenario in which the emotions were generated. Emotions also vary from person to person and are related to personality, experience, mood, and physiological arousal. In addition, the 24-h circadian rhythm influences the emotion experience.

### 4.1.3 Reliability

Finding measurements that are stable from time to time may prove problematic because a person's mood changes frequently and it may be difficult to reproduce the emotive experience a second time for a retest. For some situations, a test–retest correlation is a good estimate of reliability. Emotion can also be measured for members in a group. The interest here may be differences between people in their reactions to emotion-provoking events such as disasters (Khalid et al., 2010).

### 4.1.4 Validity

Determining whether a measure that we use to evaluate emotion(s) measures what we intend to measure is always a concern due to the fact that emotions are complex responses. As such, the measurement of an emotion cannot be reduced to one single measure (Larsen and Fredrickson, 1999).

Defining measure(s) linked to a theory can enhance construct validity. But it also simplifies measurement since the focus may be on only a few types of measures. For example, measuring the "pleasures of the mind" construct may be linked to a theory of emotional expression that restricts selection of dependent variables since these do not necessarily generate a facial expression or physiological response. Therefore, we would not consider either physiological variables or facial measures.

### 4.1.5 Measurement Error

There are two types of measurement error: random error and systematic error. To overcome random error, one can take many measures instead of a single measure and estimate a mean value. Therefore, multiple items or mathematical measurement models can be used to control or eliminate random measurement error. However, this approach is not suitable for methods which require assessments at certain times, such as experience sampling.

Another problem is that some types of assessments are intrusive (Schimmack, 2003). By asking a person to respond to a question, the contextual scenario of the emotional experience is disrupted, which may reduce the validity of the data. To minimize disruption, one can reduce the number of questions. Another way is to seek measures that are less intrusive, for example, physiological responses and facial expressions.

For heterogeneous scales that sample a broad range of affects (e.g., PANAS scales), many items are needed.

Watson et al. (1988) used 10 items and obtained item–factor correlations ranging from 0.75 to 0.52. Systematic measurement error does not pose a problem for within-subject analysis because the error is constant across repeated measurements. However, it can be misleading to use average values for calculation of correlation coefficients.

Below we present some methods that have been developed in human factors, product design and computing for the measurement of affect and emotions.

## 4.2 Methods in Affective Engineering

The methods for measuring affect or emotions are no longer entrenched in human factors alone. Research in consumer behavior, marketing, and advertising has developed instruments for measuring emotional responses to advertisement and consumer experiences of products. Since this is an area of great activity, we focus on methods that relate to affective design of products. The methods are classified into four broad categories: (1) subjective, (2) objective, (3) physiological, and (4) performance. The subjective methods are further categorized into three classes of measures: (1) user ratings of product characteristics, (2) user ratings of emotions and/or reporting of user experience without specific reference to an artifact, and (3) user ratings of emotions as induced by artifacts. Table 5 summarizes the methods.

### 4.2.1 Subjective Measures

***Ratings of Product Characteristics*** These subjective methods involve user evaluations of products. There are two established techniques: kansei engineering (Nagamachi, 2010; Nagamachi and Lokman, 2010) and semantic scales (Osgood et al., 1967). We introduce Citarasa engineering (Khalid et al., 2011) to complement these methods, and it is presented as a case at the end of this section.

***Kansei Engineering*** The method centers on the notion of *kansei* or customer's feelings for a product (Nagamachi, 2001, 2008; Khangura, 2009; Schutte et al., 2008). The word "kansei" encompasses various concepts, including sensitivity, sense, sensibility, feeling, aesthetics, emotion, affection, and intuition—all of which are conceived in Japanese as mental responses to external stimuli, often summarized as psychological feelings (Krippendorff, 2006).

Kansei engineering may be conceived as a five-stage process (Kato, 2006):

1.  *Perception Kansei*. The process of perceiving information received from objects or media via the human five senses at the physical, physiological, psychological, and cognitive levels. Different levels of appreciation may result in information being processed differently by each individual.

2.  *Situation Kansei*. The process of subjective interpretation of the situation in which the person is placed. Although people can be in the

**Table 5 Overview of Affective Engineering Methods**

| Methods | Techniques | Research Examples |
|---|---|---|
| **Subjective Measures** | | |
| Subjective rating of emotional product attributes | Kansei engineering | Nagamachi (2010); Kuang and Jiang (2009); Bouchard et al. (2009); Ishihara et al. (2009); Bahn et al. (2009); Lin et al. (2007); Jiao et al. (2006); Schutte and Eklund (2005) |
| | Citarasa engineering | Khalid et al. (2011); Golightly et al. (2011); Helander and Khalid (2009) |
| | Semantic scales | Wellings et al. (2010); Chen et al. (2009); Chuang and Chen (2008); Alcantara et al. (2005); Khalid and Helander (2004); Karlsson et al. (2003); Chen et al. (2003); Chen and Liang (2001); Küller (1975) |
| Subjective rating of emotions — general | Self-report | Lottridge (2008); Isen and Erez (2006); Sandvik et al. (1993); Brown and Schwartz (1980). |
| | Experience sampling method | Vastenburg and Hererra (2010); Lew (2009); Kapoor and Horvitz (2008); Consolvo et al. (2007); Hektner et al. (2007); Schimmack (2003) |
| | Affect grid | Warr (1999); Killgore (1998); Russell et al. (1989) |
| | Checklist (MACL) | Hodgetts and Jones (2007); Nowlis and Green (1957) |
| | Multiple-affect adjective check list | King and Meiselman (2009); Izard (1997); McNair et al. (1971); Zuckerman and Lubin (1965) |
| | Interview | Demir et al., (2009); Goh (2008); Jordan (2000); Housen (1992) |
| Subjective rating of emotions induced by artifact | PANAS scale | Saerbeck and Bartneck (2010); Turner et al. (2008); Watson et al. (1988) |
| | Questionnaire | Khalid et al. (2007b); Jordan (2000) |
| | Product emotion measurement instrument | Desmet and Schifferstein (2008) |
| **Objective Measures** | | |
| | Facial electromyography, facial action and affect coding system | Davis et al. (1995); Ekman (1982); Izard (1979); Ekman and Friesen (1976) |
| | Emotion judgment in speech; psychoacoustics and psychophonetics | Maffiolo et al. (2002); Biever (2005); Larsen and Fredrickson (1999); Scherer (1986) |
| **Psychophysiological Measures** | | |
| | Thermography; galvanic skin response, and wearable sensors | Jenkins et al. (2009); Picard (2000); Larsen and Fredrickson (1999) |
| **Performance Measures** | | |
| | Judgment task involving probability estimates, and lexical decision task | Mayer and Bremer (1995); Niedenthal and Setterlund (1994); Ketelaar (1989); Challis and Krane (1988) |

same location at the same time, they can feel differently of the situation due to differences in lifestyles and behaviors of the day.

3. *Knowledge Kansei*. The process of organizing the knowledge base within one's own memory in terms of, for example, association rules. Differences may be due to an individual's personal interests, resulting in differences in vocabulary to express the experience and the associated organization.

4. *Action/Expressive Kansei*. The process of taking specific action or expressing information, in various media, to the external world through the physical body or an electronic gadget. Differences in behavior habits (rules) may account for skill differences and also for differences in action patterns.

5. *Intention Kansei*. The process of selecting action and expression (output) from sensory images and interpretations of a situation (input). This

process is based upon a person's internalized objective. Differences in lifestyles form the differences in response relationships between input and output. This process, which is akin to goal driven, links the above processes together.

Some aspects of the above processes are translated into the following basic procedure in kansei engineering (Nagamachi, 2001): (1) collect kansei words; (2) correlate design characteristics with kansei words (e.g., using Osgood's semantic differential technique); (3) perform factor analysis on kansei words to determine similarity; and (4) analyze product features to predict emotions.

The method has been expanded to use statistical visualization techniques (Ishihara et al., 2009), fuzzy logic (Lin et al., 2007), and data mining (Jiao et al., 2006). Besides automobiles (Bahn et al., 2009; Schutte and Eklund, 2005), other products that have applied the kansei method include mobile phones (Kuang and Jiang, 2009), footwear (Bouchard et al., 2009), home design (Nagamachi, 2008), and robotics (Saerbeck and Bartneck, 2010). Kansei engineering has also been applied to services (Hartono and Tan, 2009) and surface roughness (Choi and Jun, 2007).

In product platform development of mobile phones (Kuang and Jiang, 2009), the procedure involves (1) identifying platform and individual parameters, then quantifying the relationship between the product's perceptual image and the design parameters by using regression analysis from an affective evaluation survey; (2) grouping customers' responses according to their preference similarity coefficients by a cluster analysis of the preference evaluation survey in which the values of the platform parameters are fixed, then determining the number of platforms based on the clusters; (3) establishing the quantified relationship between the average preference and the individual parameters for each cluster using the regression method; and (4) determining the values of the individual parameters based on the satisfaction of each customer group. The product platform developed by the proposed method can achieve customer satisfaction, and a company can combine the simple individual form elements to the platform to rapidly develop a customized product form to meet a certain customer's affective need.

In short, the kansei engineering method has expanded extensively beyond the Japanese borders since it was first introduced in 1970s. A society by the same name is operating actively in Japan, with members across the globe. The Kansei Engineering Society also hosts conferences with the most recent in France called KEER 2010 (Kansei Engineering and Emotion Research).

**Semantic Scales** Semantic scales are similar to scales used in kansei engineering. The main difference is that the scales rely on the methodology proposed in Osgood's semantic differential (SD) technique (Osgood et al., 1967). This technique makes it possible to assess semantic differences between objects. Adjective pairs of opposite meanings are created, such as light–heavy, open–closed, and fun–boring. Subjects then rate objects using, for example, a five-point scale, such as 1 (very fun), 2 (fun), 3 (neutral), 4 (boring), and 5 (very boring). A main problem is to validate the word pairs. In the first place it is not trivial to assess if the two words constitute semantic opposites. One would also need to demonstrate that the chosen word pair is appropriate to evaluate the artifact in question. In our opinion it is easier to use scales that are unipolar; this is because it is difficult to find semantic opposites of words.

Chen et al. (2009) explored relationships between touch perception and surface physical properties using a semantic differential questionnaire to assess responses to touching the textures of confectionary packaging such as cardboards, laminate boards, and flexible material against six word pairs: warm–cold, slippery–sticky, smooth–rough, hard–soft, bumpy–flat, and wet–dry. In addition, they obtained four physical measurements to characterize the surfaces' roughness, compliance, friction, and rate of cooling of an artificial finger when touching the surface. Results of correlation and regression analyses showed that touch perception is associated with more than one physical property and the regression model can represent both strength and form.

Karlsson et al. (2003) used Küller's (1975) method that was developed for architectural appreciation in design; however, they applied to the automobile. They obtained significant results that discriminated between the designs of four passenger cars: BMW 318 (more complex and potent), Volvo S80 (more original and higher social status), Audi A6 (less enclosed), and VW Bora (greater affect). Considering the significant results, one may debate whether formal validation is necessary; the significant results carried much face validity. Obviously, this methodology works well with cars as well as architecture.

Chen et al. (2003) proposed a framework for understanding how product shapes evoke affective responses. For a set of representative product shapes, they first conduct a survey to evaluate the affective characteristics of each product shape. They then compute a spatial configuration that summarizes the affective responses toward the set of shapes by applying perceptual mapping techniques to the survey data. A series of new product shapes that smoothly interpolate among product shapes were then generated by using image-morphing techniques. With data from a follow-up survey, they inserted the new shapes into the spatial configuration. The trajectory or distribution of the interpolated shapes provided visualization of how affective characteristics changed in response to varying shapes. They found the relationship between the shapes and the affective characteristic to be nonlinear and nonuniform.

Wellings et al. (2010) applied the SD rating scale to measure the "feel" of push-switches in five luxury saloon cars both in context (in-car experience) and out of context (on a bench). Besides semantic scales they also obtained hedonic data on subjective liking. Factor analysis showed that perceived characteristics of switch haptics can be explained by three factors: *image, build quality*, and *clickiness*.

Khalid and Helander (2004) developed a rating tool to measure user responses to four future electronic

devices [cell phone, personal digital assistant, radio, and geographical positioning system (GPS)] for an instrument panel of a car. Users had to imagine the products and rated their affective preferences for 15 product attributes on 10-point SD scales. These attributes comprised functional and affective customer needs derived from a customer survey. Using factor analysis, three generic factors were extracted: *holistic attributes, styling*, and *functional design*. Depending upon the familiarity of the device there were clear differences among users. Devices that were unfamiliar to the test persons, such as GPS, were assessed using holistic attributes. Familiar designs, such as car radio and cell phone, were assessed using styling and functionality attributes.

More recently, Chuang and Chen (2008) introduced the hierarchical sorting method and divide-and-conquer method to improve the efficiency of rating a large number of visual stimuli, such as armchairs, derived from multiple attribute scales. The attribute data collected by both methods were quite consistent with an average correlation of 0.73. As such, the methods were proven to be more efficient than the manual card-sorting method. Depending on the objective of the perceptual task, the hierarchical sorting method is effective for distinguishing details in visual differences among similar stimuli relative to the divide-and-conquer method.

Alcantara et al. (2005) applied SD to structure the semantic space of casual shoes. Sixty-seven volunteers evaluated 36 shoe models on 74 adjectives that formed the "reduced semantic universe." Factorial analysis of principal components was used to identify the semantic axes. A statistical index was introduced to measure the subject's consensus and then used to analyze the influence of the number of volunteers in the semantic evaluation results. The results showed that comfort and quality were independently perceived by consumers; whereas comfort was clearly identified by users, quality was not.

### Subjective Ratings of Emotions

These include techniques that report a person's subjective experience, such as self-reports and experience sampling method, or rating of one's own emotions in the form of affect grid, checklist, and interview. These methods have more general applicability to products as well as tasks and scenarios.

### Self-Reports

This technique requires participants to document their subjective experiences of the current situation. A self-report can reflect on one's present state and compare it to the past state. As such, the self-reporting technique relies on the participant's ability to report experiences and to reflect accurately on their experiences. The measures may be instantaneous or retroactive. Instantaneous reports refer to the emotion as first experienced, while retrospective reports refer a situation after the fact. Such assessments can be complemented using a video as a reminder.

Self-report measures involve a plethora of affect inventories: verbal descriptions of an emotion or emotional state, rating scales, standardized checklists, questionnaires, semantic and graphical differentials, and projective methods. Criticisms of self-report methods include the possibility that they draw attention to what the experimenter is trying to measure, that they fail to measure mild emotions, and that they are construct valid (Isen and Erez, 2006).

Lottridge (2008) measured emotional responses based on the model of valence and arousal. Reactions to storyboard and video prototypes in a pilot study motivated the need for continuous affective self-reports. An experiment then compared the relative ease of use and the cognitive complexity of different methods of emotional measurement.

A major limitation of the self-report is that it relies exclusively on the person's cognitive labels of emotions to remember and summarize their experiences over longer or shorter intervals of time. But emotion, as argued above, is a multichannel phenomenon and is not limited to the cognition of emotion. In addition, there are physiological, facial, nonverbal, behavioral, and experiential elements (Diener, 1994). Self-reports of emotional well-being, such as happiness, tend to reveal greater consistency than many other types of emotion (Brown and Schwartz, 1980). There is also agreement between self-reports of emotional well-being and interview ratings, peer reports, and memory for pleasant events (Sandvik et al., 1993).

### Experience Sampling Method (ESM)

Coined by Larson and Csikszentmihalyi (1983), ESM measures people's self-reported experiences close in time to the occurrence of the scenario that evoked the emotion. Typically ESM uses a combination of online and short-term retrospective question formats in which people report what is presently or recently occurring (e.g., "How do you feel *right now*? How did you feel *this past hour*?"). As such, these procedures measure subjective experience that is episodic in nature.

ESM was designed to capture user experiences in the field. Initially ESM took advantage of the popularity of earlier mobile devices (e.g., pagers) to ask people for feedback at random times during the day. This configuration aimed to reduce problems that participants might have when recalling events, a problem underlying many self-report techniques (Hektner et al., 2007). With ESM, participants make a quick record close to the moment of interest, rather than having to recall what they did in the past.

As mobile technology evolves, the sampling process becomes more intelligent and sensitive to the context of the product use (Lew, 2009). Nowadays, researchers can use selective sampling (Consolvo et al., 2007), a sampling technique that links the timing and questions to relevant events using a portable device that collects user feedback. The device's ESM controller uses sensor data that could capture contextual as well as user product events to select relevant sampling moments. In addition, the controller, based on the same information, may decide what question or flow of questions should be asked and how they should be presented together with the format of the answers.

A common challenge when using ESM is to maximize the quality and quantity of the samples while

minimizing interruptions and maintaining the motivation of the participants. The development of strategies to optimize sampling, interruptions, and motivation has been the key focus in recent ESM research (Kapoor and Horvitz, 2008). Using these strategies, timing and content of questions can be adapted to the actual state and context of the participants and the product.

Experience sampling techniques can be used to study user experiences with products in a natural setting and over time (Vastenburg and Hererra, 2010). Currently, researchers can use selective sampling to link the timing and questions to relevant product events and contextual events. Existing research has focused on maximizing the quality and quantity of feedback, while at the same time minimizing interruptions and maintaining the motivation of the participants. Adaptive experience sampling is a method that enables researchers and designers to change the focus of their experience sampling study on the fly.

**Affect Grid** Developed by Russell et al. (1989) the technique measures single-item affect in the form of a square grid anchored horizontally with pleasure/displeasure (valence) and vertically with arousal/sleepiness (activation). On the basis of subjective feelings, a subject places an X along two dimensions: pleasantness and arousal. Both aspects will be rated; if both are rated highly, the subject feels great excitement. Similarly, there are feelings of depression, stress, and relaxation. The affect grid displays strong evidence of discriminant validity between the dimensions of pleasure and arousal. Studies that used the affect grid to assess mood provided further evidence of construct validity. However, the scale is not an all-purpose scale and is slightly less reliable than a multiple-item questionnaire for self-reported mood.

Studies that measured more subtle affective states than the main dimensions of valence, arousal, and dominance have had some but more limited success—for example, mapping subtle affective attributes to a defined subregion of a two-dimensional (2D) valence and arousal grid (Killgore, 1998). Similarly, Warr (1999) used the same scale to measure well-being along a 2D framework of well-being in terms of the location in this 2D space of arousal and pleasure. A particular degree of pleasure or displeasure may be accompanied by high levels of mental arousal or a low level of arousal (sleepiness), and a particular level of mental arousal may be either pleasurable (pleasant) or unpleasurable (unpleasant) (Warr, 1999).

**Checklists.** Mood checklists comprise lists of adjectives that describe emotional states. Subjects are required to check their emotions. The mood adjective checklist (MACL) developed by Nowlis and Green (1957) contains 130 adjectives with a four-point scale: "definitely like it," "slightly," "cannot decide," and "definitely not."

Hodgetts and Jones (2007) used a mood checklist as an interrupting task in an on-going performance of Tower of London (ToL) problems. A list of six statements along a mood continuum (e.g., "extremely happy" to "extremely sad") was presented in the center of the screen and participants were asked to select the one that best applied to them. This task was irrelevant to the main ToL problem. The number of mood checklists to be presented was manipulated—in three, five, or seven—but the precise timing of these was controlled by the participant rather than the computer program. The mood checklist interruptions were brief and undemanding in content, comparable to many types of pop-ups that increasingly invade our computer screens. Despite its brief duration, the interruption affected task performance.

Zuckerman and Lubin (1965) developed the multiple-affect adjective checklist (MAACL) comprising 132 items, which they revised in 1985 (MAACL-R). The revised version allowed scoring of several pleasant emotions, taking into account global positive and negative affect as well as sensation seeking. The revised MAACL-R scale is an alternative to the profile of mood states (POMS) scale (McNair et al., 1971). The POMS is an assessment of transient mood states, measuring six factors: tension–anxiety, depression–dejection, anger–hostility, vigor–activity, fatigue–inertia, and confusion–bewilderment. Of the six factors, only one represents positive expressions (vigor–activity). Izard (1992) developed the multi-item differential emotional scale (DES) with the purpose of assessing multiple discrete emotions.

More recently, King and Meiselman (2009) developed EsSense Profile™ using the adjectives from POMS and MAACL-R scales. Terms were validated based on the clarity and usage frequency to ensure the application to a range of products. The final scale consisted of 39 emotions to represent consumer affective responses toward food.

**Interviews.** Interviews may be used to assess product pleasure and pleasure from activities or tasks. It is a versatile method and can be performed face to face or through phone conversation. Questions can be structured or unstructured (Jordan, 2000). A structured interview has a predetermined set of questions, whereas an unstructured interview uses a series of open-ended questions.

Housen (1992) proposed a nondirective, stream-of-consciousness interview. Participants are asked simply to talk about anything that comes to their mind as they look at a work of art. It is called the "aesthetic development interview." There are no directed questions that can influence the viewer's statement. It provides a window into a person's thinking and minimizes researcher biases or assumptions. To insure reliability and consistency in the evaluation, the interviews are often examined by two independent coders, and the coding is then charted graphically by computer to enable a comprehensive representation of all thoughts that went through the subject's head.

A combination of in-depth interviews and experience sampling was used to capture the causes and emotions experienced when interacting with products (Demir et al., 2009). The appraisal patterns elicited emotions of product users for four emotion groups: happiness/joy, satisfaction/contentment, anger/irritation, and disappointment/dissatisfaction.

**Subjective Ratings of Emotions Induced by Artifacts** These rating scales have been used to document how artifacts make a person feel. By asking a question such as "What does the look of this car make you feel?" the user is expected to evaluate his or her emotions in relation to the artifact.

**PANAS Scales** Watson et al. (1988) developed the positive affect–negative affect schedule (PANAS). The purpose of PANAS is to measure positive and negative mood states of a person during different times or contexts: the current day, a week, or a year. PANAS evaluates mood adjectives using a five-point scale: "not at all," "slight," "a little," "moderately," "quite a bit," and "very much." *Positive affect* (PA) refers to feelings of enthusiasm, alertness, and activeness. A high PA score reflects a state of "high energy, full concentration and pleasurable engagement" (Watson et al., 1988). *Negative affect* (NA), on the other hand, refers to feelings of distress and unpleasurable engagement.

To describe PA, 10 descriptors were used: attentive, interested, alert, excited, enthusiastic, inspired, proud, determined, strong, and active. NA is measured using the following 10 descriptors: distressed, upset, hostile, irritable, scared, afraid, ashamed, guilty, nervous, and jittery. The 10-item scales have been shown to be highly internally consistent, largely uncorrelated, and stable at appropriate levels over a period of two months. When used with short-term instructions (e.g., *right now* or *today*) they are sensitive to fluctuations in mood, but when longer term instructions are used (e.g., *past year* or *general*) the responses are stable.

Turner et al. (2008) used the PANAS schedule to explore the relationships between comfort in making and receiving mobile phone calls in different social contexts, their affective responses to public mobile phone use by others, and how such factors relate to personal attributes and specific beliefs about calling behavior. The results of factor analyses revealed "context" as the most important in mobile phone use, and users differed in the extent to which they felt comfortable making and receiving calls in different locations.

To investigate the relation between a robot motion and the perceived affective state of the robot, Saerbeck and Bartneck (2010) applied two scales: PANAS and SAM (self-assessment mannikins). They used two motion characteristics for the perceived affective state: acceleration and curvature.

**Questionnaire.** Philips Corporate Design developed a questionnaire for measuring pleasure from products (Jordan, 2000). The questionnaire has 14 questions, focusing on several feelings that a user may have: stimulated, entertained, attached, sense of freedom, excited, satisfaction, rely, miss, confidence, proud, enjoy, relax, enthusiastic, and looking after the product. Using a five-point scale, ranging from disagree (0) to neutral (2) and strongly agree (4), the questionnaire covered most of a user's possible responses. To measure pleasure, open-ended items were added as an option. This was particularly useful to inform the product developers about the users' evaluation of pleasure.

Khalid et al. (2007b) developed two questionnaires to elicit citarasa in vehicle customers. The CARE (car evaluation) questionnaire was created for car customers and the TNT (truck needs tracker) questionnaire for truck customers. The CARE tool is an open-ended questionnaire with 60 questions which required detailed responses from the participant. In addition, there were close-ended questions for customer personal data and knowledge. The questions were driven by the *citarasa* model (see Figure 5). There were four parts:

*Part I: Customer Experience and General Needs.* Elicited customer experiences with the car, such as driving experience, previous experience, and expected criteria. The information was needed for the "marketing" section in the model, which focused on factors such as price, and customer information. Questions were designed to probe customers about cars driven currently and previously and whether it constituted a good buy. To measure customer citarasa relating to visceral, behavioral, and reflective design, questions addressed sensory characteristics (tactile, visual, olfactory, and auditory), functional requirements, and needs for trends, status, and style.

*Part II: Specific Requirements for Design.* Measured affective and functional requirements of existing cars and images of cars. The purpose was to generate a set of citarasa descriptors for developing the citarasa engineering database. Questions addressed both exterior and interior requirements of cars, and they were partially based on samples drawn from the Consumer Reports (2007) Annual Car Reliability Survey.

*Part III: Customer Expertise.* Evaluated customer knowledge, expertise, goals, and affordance. This relates to the "design goals, constraints" and "marketing" aspects of the citarasa model. Questions include how to obtain information on cars and how to decide it is a good car.

*Part IV: Customer Demographics.* Recorded driving experience, handedness, living arrangement, education level, and income. The data collected in this part were used to construct the "society" constraints in the model.

A pilot study was conducted to test the usability of the questionnaire. Questions that were found to be difficult or ambiguous were modified. The results were also used for creating response categories.

**Emotion Rating Scales** Several studies have developed methods and tools for measuring emotions in products. Desmet and Schifferstein (2008) proposed a method to measure complex emotions to product design using a nonverbal, cross-cultural tool called PrEmo® (Product Emotion Measurement Tool). PrEmo® consists of 14 animated characters expressing seven positive and seven negative emotions.

*Citarasa engineering* was developed in the CATER project. The method is driven by the concept of *citarasa* or *emotional intent* (Khalid et al., 2007a, 2011). Citarasa differs from kansei engineering as it has a theoretically

driven basis for elicitation and analysis of affective needs for vehicle design. The method has five steps as summarized in Figure 8.

A model of emotional intent was conceptualized. A semantic framework of citarasa words was then developed to map words to specific vehicle components to form citarasa ontology. Customer citarasa were then elicited in the field using probe interview technique. Affective needs were refined through Web surveys; finally the elicited citarasa were analyzed using data-mining techniques and the citarasa analysis tool. The tool is linked to the citarasa database, which enables analysis of affective needs in several countries in Europe and Asia. The system has been technically verified, validated, and tested for usability with consumers and automotive end users.

Unlike other methods that relied solely on subjective rating, this methodology uses a multimethod approach to elicit affective descriptors—from interview to questionnaire, rating scale, ranking, and mood board. It was used to validate the descriptors in a Web-based tool called the Citarasa System (Khalid et al., 2009). The results were analyzed with WEKA, a data-mining method, and other statistical analyses, including correlation analysis, cluster analysis, and factor analysis.

### 4.2.2 Objective Measures

Two objective methods to record emotions are analysis of facial expressions and vocal content of speech or voice expressions.

*Facial Expressions* Numerous methods exist for measuring facial expressions (Ekman, 1982). Facial expressions provide information about (1) *affective state*, including emotions such as fear, anger, enjoyment, surprise, sadness, disgust, and more enduring moods such as euphoria, dysphoria, or irritableness; (2) *cognitive state*, such as perplexity, concentration, or boredom; and (3) *temperament and personality*, including such traits as hostility, sociability, or shyness.

Ekman and Friesen (1976) identified five types of messages conveyed by rapid facial signals: (1) *emotions*, including happiness, sadness, anger, disgust, surprise, and fear; (2) *emblems*, culture-specific symbolic communicators such as the wink; (3) *manipulators*, self-manipulative associated movements such as lip biting; (4) *illustrators*, actions accompanying and highlighting speech such as a raised brow; and (5) *regulators*, non-verbal conversational mediators such as nods or smiles.

Measurement of facial expressions may be accomplished by using the facial action coding system (FACS). The method, proposed by Ekman and Friesen (1976), captures the facial changes that accompany an emotional response to an event. FACS was developed by determining how the contraction of various facial muscles (singly and in combination with other muscles) changes the appearance of the face. Videotapes of more than 5000 different combinations of muscular actions were examined to determine the specific changes in appearance and how to best differentiate one appearance from another.

Measurement with FACS is done in terms of action units (AUs) rather than muscular units for two reasons. First, for a few changes in appearance, more than one muscle is used to produce a single AU. Second, FACS distinguishes between two AUs for the activity of the frontalis muscle that produces wrinkles on the forehead. This is because the inner and outer portion of
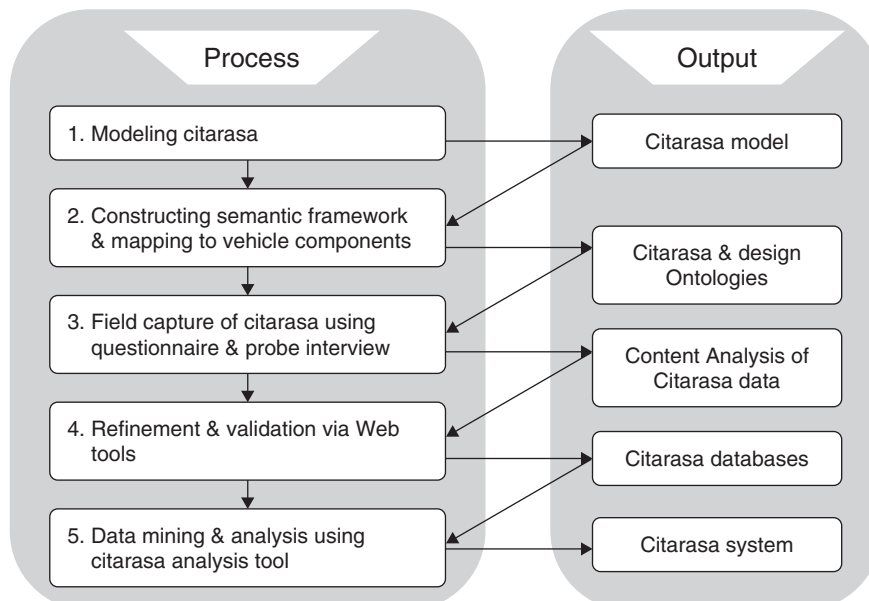


**Figure 8** Methodology for citarasa elicitation and analysis (Khalid et al., 2011).

this muscle can act independently, producing different changes in appearance. There are 46 AUs which account for changes in facial expression and 12 AUs which describe gross changes in gaze direction and head orientation. To use FACS the investigator must learn about the appearance and the muscles of the face for each AU. This demands much time and effort.

The maximally discriminative affect coding system (MAX) developed by Izard (1979) measures visible appearance changes in the face. The MAX units are formulated in terms of facial expressions that are relevant to eight specific emotions, rather than in terms of individual muscles. Unlike FACS, MAX does not measure all facial actions but scores only facial movements that relate to the eight emotions.

Facial changes can also be registered using electromyography (EMG). EMG measures nerve impulses to muscles which produce facial changes or expressions. This measure assumes that emotions are visible through facial expressions, which is the case when people interact with each other.

Davis et al. (1995) compared facial electromyography with standard self-report of affect. He obtained a good correlation between activity of facial muscles and self-report of affect. The pattern of muscular activation could be used to indicate categories of affect, such as happy and sad, and the amplitude of electromyographic signals gave information on degree of emotions. In other words, Davis et al. (1995) was able to categorize as well as quantify affective states using facial electromyography.

**Vocal Measures of Emotion**    Most of the emotions conveyed in speech are from the verbal content. Additionally, the style of the voice, such as pitch, loudness, tone, and timing, can convey information about the speaker's emotional state. This is to be expected because vocalization is "a bodily process sensitive to emotion-related changes" (Larsen and Fredrickson, 1999). A simple and maybe also the best way to analyze the emotional content would be to listen to recordings of voice messages. Scherer (1986) noted that judges seem to be rather accurate in decoding emotional meaning from vocal cues. Some emotions are easier to recognize than others. Sadness and anger are easiest to recognize, whereas joy, disgust, and contempt are difficult to recognize and distinguish from one another.

Maffiolo and Chateau (2003) investigated the emotional quality of speech messages used by France Telecom Orange. Each year, vocal servers were used to respond to hundreds of millions of phone calls. The audio messages can be help messages, navigation messages, and information messages. The purpose of the study was to create a set of voice messages that were perceived as friendly, sincere, and helpful. Later, Maffiolo et al. (2007) improved the automatic detection and characterization of emotion-related expressions in human voice by an approach based on human auditory perception. A listening test was set up with 72 listeners. The corpus was constituted of 18 voice messages extracted from a real-life application. Message segments of different temporal length were listened to by listeners who were asked to verbalize their perception. Fourteen metacategories were obtained related to age, gender, regional accent, timbre, personality, emotion, sound quality, expression style, and so on. The temporal windows of listening necessary for listeners to perceive and verbalize these categories underlie the building of submodels relevant to the automatic recognition of emotion-related expressions.

A more high-tech method is to digitize voice recordings and analyze the voice by decomposing the speech sound waves into a set of acoustic parameters and then analyze the psychoacoustics and psychophonetics content (Larsen and Fredrickson, 1999). This includes analysis of pitch, small deviations in pitch, speaking rate, use of pauses, and intensity.

*Emotive Alert*, a voicemail system designed by Inanoglu and Caneel of the Media Lab at the Massachusetts Institute of Technology (Biever, 2005), labels messages according to the caller's tone of voice. It can be installed in a telephone exchange or in an intelligent answering machine. It will analyze incoming messages and send the recipient a text message along with an emoticon indicating whether the message is urgent, happy, excited, or formal. In tests on real-life messages, the software was able to tell the difference between excited and calm and between happy and sad but found it harder to distinguish between formal and informal and urgent and nonurgent. This is because excitement and happiness are often conveyed through speech rate and volume, which are easy to measure, while formality and urgency are normally expressed through the choice of words and not easy to measure (Biever, 2005). At the present time the first method, listening to speech, is probably the more reliable.

Regardless of the method used, vocal measures of emotion are sometimes difficult to use since (1) voice is not a continuous variable as people do not speak continuously and thus vocal indicators of emotion are not always present; (2) positive and negative emotions are sometimes difficult to distinguish; and (3) the voice can reflect both emotional/physiological and sociocultural habits, which are difficult to distinguish (Scherer, 1998).

### 4.2.3  Psychophysiological Measures

Emotions often affect the activity of the autonomic nervous system (ANS) and thereby the activation level and arousal. At the same time, there are increases and decreases in bodily functions, such as in heart functions, electrodermal activity, and respiratory activity (Picard, 1997). Thus, there is a variety of physiological responses that can be measured, including blood pressure, skin conductivity, pupil size, brain waves, and heart rate frequency and variability. For example, in situations of surprise and startle, the electrical conductivity of certain sweat glands is momentarily increased. This is referred to as a galvanic skin response (GSR). These sweat glands are primarily found on the inside of the hands and on the soles of the feet. Electrodes are then attached to measure the electrical conductivity (Helander, 1978). The nerve signals take about 1.5 s to travel from the

brain to the hand, and therefore the response is a bit delayed.

Researchers in the field of affective computing are actively developing "ANS instruments," such as IBM's emotion mouse (Ark et al., 1999) and a variety of wearable sensors (e.g., Picard, 2000). With these instruments, computers can gather a multitude of psychophysiological information while a person is experiencing an emotion and learn which pattern is most indicative of which emotion.

ANS responses can be investigated in experiments, for example, by using film clips to induce the type of emotions investigated (e.g., amusement, anger, contentment, disgust, fear, sadness), while electrodermal activity, blood pressure, and an electrocardiogram (ECG) are recorded. Therefore, it is possible to associate a variety of emotions with specific physiological reactions.

While autonomic measures are fruitful, it is important to note that:

1. Autonomic measures vary widely in how invasive they are. The less invasive measures include pulse rate and skin conductance, while measures of blood pressure are often invasive since they use pressure cuffs which are deflated. This may distract a person, so that the emotion is lost.

2. The temporal resolution of various autonomic measures varies widely. Some measures are instantaneous, such as GSR, while impedance cardiography, for example, requires longer duration for reliable measurement (Larsen and Fredrickson, 1999).

3. Different measures have different sensitivity. Depending upon the emotion which is recorded, it is best to first validate the particular physiological measures so as to understand if it is sensitive enough to record differences in the intensity of the emotion.

Infrared thermography (IRT) offers human factors researchers a highly accurate, noncontact and objective measurement tool for exploring the dynamics of a user's affective state during user–product interaction. It provides useful visual and statistical data for designers to understand the nature and quality of user experience. Jenkins et al. (2009) compared thermographic, EEG, and subjective measures of affective appearance during simulated product interactions. The results showed the utility of IRT in the measurement of cognitive work and affective state changes but the causal relationships between facial temperature dynamics, cognitive demand, and affective experiences need to be explored further.

### 4.2.4 Performance Measures

These measures typically indicate the effect of emotions on decision making. Emotion-sensitive performance measures may be obtained through judgment tasks. One popular task is to have participants make probability estimates of the likelihood of various good and bad events. It has been shown that persons in unpleasant emotional states tend to overestimate the probability of bad events (Johnson and Tversky, 1983).

Ketelaar (1989) showed that people in a good mood also overestimated the probability of pleasant events. Another useful performance task is to ask participants to generate associations to positive, neutral, and negative stimuli. Mayer and Bremer (1985) showed that performance on this task correlated with the naturally occurring mood.

A second category of performance measures involves information-processing parameters. Reaction times in lexical decision tasks have been shown to be sensitive to affective states (Challis and Krane, 1988). The task involves judging if a string of letters presented on the computer screen represents a word or nonword. Participants in positive affective states are quicker and sometimes more accurate at judging positive words as words compared to participants in neutral states, and vice versa for unpleasant moods (Niedenthal and Setterlund, 1994).

## 5 CONCLUSION

The expectations of users in terms of customer needs are changing: Functionality, attractiveness, ease of use, affordability, and safety are taken for granted. The new trends are for objects or artifacts that inspire users, enhance their lives, and evoke emotions and dreams in their minds. In product design, there are two evaluations— the first one based on cognition (knowledge and functionality) and the second one on affect.

Professionals in human factors and HCI have come to realize that usability is not enough. The main goal is to please the user rather than maximize transactions and productivity. This is particularly critical for e-commerce and other Web-based activities, where there are many alternatives. Human–computer interaction may hence be designed to induce affective and memorable experiences.

There is no such thing as a neutral interface; any design will elicit emotions from the user and the designer. The designer should aim to enhance the user experience through a deliberate design effort, thus bridging the gap between the affective user and the designer's environment, as outlined in Figure 5 of the affective user-designer model. The user, on the other hand, will gain pleasurable experience from a more fun interface. This will promote productivity and enjoyment.

Affective engineering, which is concerned with measuring people's affective responses to products and identifying the properties of the products to which they are responding, is now expanding as a method for use in both research and the industry. The most commonly used approach is the self-report whereby adjectives that people use to describe the product are identified and embodied into a semantic differential scale or questionnaire with rating scale. The use of such subjective methods has drawbacks. They rely heavily on the use of words and adjectives. The subject's vocabulary should be taken into account, because some people may have little comprehension of some of the words used in the methods mentioned above. The subject should be allowed to use his or her principal

language, or else some feelings might be misinterpreted. The words or adjectives must be concise and easy to understand and take into account cultural as well as contextual factors (Larsen and Fredrickson, 1999).

Physiological methods (or nonverbal instruments) have a better advantage as they are language independent and can be used in different cultures. A second advantage is that they are unobtrusive and do not disturb participants during the measurement. There are however limitations to the physiological measures. They can only reliably assess a limited set of "basic" emotions and cannot assess mixed emotions. For pleasures of the mind, it is doubtful if any of the psychophysiological methods will be sensitive enough to capture the subtleness of emotions.

Objective methods such as vocal content and facial expressions can be used to measure mixed emotions, but they are difficult to apply between cultures. It would be important to make cultural comparisons between vocal and facial expressions. For this purpose a multimedia database can be developed and shared by the research community. The database could contain images of faces (still and motion), vocalizations and speech, psychophysiological correlates of specific facial actions, and interpretations of facial scores in terms of emotional state, cognitive process, and other internal processes. This would facilitate an integration of research efforts by highlighting contradictions and consistencies and suggesting fruitful avenues for new research.

## REFERENCES

Aboulafia, A., and Bannon, L. J. (2004), "Understanding Affect in Design: An Outline Conceptual Framework," *Theoretical Issues on Ergonomics Science,* Vol. 5, No. 1, pp. 4–15.

Alcántara, E., Artacho, M. A., González, J. C., and García, A. C. (2005), "Application of Product Semantics to Footwear Design. Part I—Identification of Footwear Semantic Space Applying Differential Semantics," *International Journal of Industrial Ergonomics*, Vol. 35, No. 8, pp. 713–725.

Apter, M. J. (1989), *Reversal Theory: Motivation, Emotion and Personality*, Routledge, London.

Ark, W., Dryer, D. C., and Lu, D. J. (1999), "The Emotion Mouse," in Human-computer interaction: Ergonomics and user interfaces, *Proceedings of HCI International 99: 8th International Conference on Human-Computer Interaction*, H.-J. Bullinger and J. Ziegler, Eds., Lawrence Erlbaum Associates, Mahwah NJ, pp.818–823.

Ashkanasy, N. M, and Cooper, C. L. (2008), *Research Companion to Emotions in Organizations,* Edward Elgar Publishing, Northampton, MA.

Bahn, S., Lee, C., Nam, C. S., and Yun, M. H. (2009), "Incorporating Affective Customer Needs for Luxriousness into Product Design Attributes," *Human Factors and Ergonomics in Manufacturing*, Vol. 19, No. 2, pp. 105–127.

Bannon, L. (1993), "CSCW: An Initial Exploration," *Scandinavian Journal of Information Systems*, Vol. 5, pp. 3–24.

Bannon, L., and Bødker, S. (1991), "Beyond the Interface: Encountering Artifacts in Use," in *Designing Interaction: Psychology at the~Human-Computer Interface,* J. M. Carroll, Ed., Cambridge University Press, New York, pp. 227–253.

Bardzell, J. (2009), "Interaction Criticism and Aesthetics," in *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI'09),* ACM, New York, pp. 2357–2366.

Bergman, E. (2000), *Information Appliances and Beyond*, Academic, San Diego, CA.

Biever, C. (2005), "Voicemail Software Recognises Callers' Emotions," *New Scientist*, Vol. 2481, p. 21.

Bloch, H. (1995), "Seeking the Ideal Form: Product Design and Consumer Response," *Journal of Marketing*, Vol. 59, No. 3, pp. 16–29.

Bloch, P., Brunel, F. F., and Arnold, T. J. (2003), "Individual Differences in the Centrality of Visual Product Aesthetics: Concept and Measurement," *Journal of Consumer Research,* Vol. 29, pp. 551–565.

Blythe, M. A., Overbeeke, K., Monk, A. F., and Wright, P.C., Eds. (2004). *Funology: From Usability to Enjoyment,* Kluwer Academic Press, Dordrecht, pp. 43–53.

Bouchard, C., Mantelet, F., Aoussat, A., Solves, C., Gonzalez, J. C., Pearce, K., Van Lottum, C., and Coleman, S. (2009), "A European Emotional Investigation in the Field of Shoe Design," *International Journal of Product Development,* Vol. 7, No. 1/2, pp. 3–27.

Brave, S., and Nass, C. (2003), "Emotion in Human-Computer Interaction," in *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, J. Jacko and A. Sears, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 81–96.

Brown, S. L., and Schwartz, G. E. (1980), "Relationships between Facial Electromyography and Subjective Experience during Affective Imagery," *Biological Psychology,* Vol. 11, No. 1, pp. 49–62.

Buck, R. (1999), "The Biological Affects: A Typology," *Psychological Review,* Vol. 106, No. 2, pp. 301–336.

Cabanac, M. (1992), "Pleasure: The Common Currency," *Journal of Theoretical Biology,* Vol. 155, pp. 173–200.

Camerer, C., Loewenstein, G., and Prelec, D. (2005), "Neuroeconomics: How Neuroscience Can Inform Economics," *Journal of Economic Literature*, Vol. 43, No. 1, pp. 9–64.

Carroll, J. M. (2004), "Beyond Fun," *Interactions*, Vol. 11, No. 4, pp. 38–40.

Challis, B. H., and Krane, R. V. (1988), "Mood Induction and the Priming of Semantic Memory in a Lexical Decision Task: Assymetric Effects of Elation and Depression," *Bulletin of the Psychonomic Society,* Vol. 26, pp. 309–312.

Chang, H. C., Lai, H. H., and Chang, Y. M. (2006), "Expression Modes Used by Consumers in Conveying Desire for Product Form: A Case Study of Car," *International Journal of Industrial Ergonomics*, Vol. 36, pp. 3–10.

Chen, L. L., and Liang, J. (2001), "Image Interpolation for Synthesizing Affective Product Shapes," in Proceedings of the International Conference on Affective Human Factors Design, M. G. Helander, H. M. Khalid, and M. P. Tham, eds., Asean Academic Press, London, pp. 531–537.

Chen, L. L., Wang, G. F., Hsiao, K-A, and Liang, J. (2003), "Affective Product Shapes through Image Morphing," in *Proceedings of the 2003 International Conference on Designing Pleasurable Products and Interfaces*, ACM, New York, pp. 11–16.

Chen, X., Shao, F., Barnes, C., Childs, T., and Henson, B. (2009), "Exploring Relationships between Touch Perception and Surface Physical Properties," *International Journal of Design*, Vol. 3, No. 2, pp. 67–76.

Choi, K., and Jun, C. (2007), "A Systematic Approach to the Kansei Factors of Tactile Sense Regarding the Surface Roughness," *Applied Ergonomics*, Vol. 38, No. 1, pp. 53–63.

Chuang, Y., and Chen, L-L. (2008), "How to Rate 100 Visual Stimuli Efficiently," *International Journal of Design*, Vol. 2, No. 1, pp. 31–43.

Coelho, D. A., and Dahlman, S. (2000), "Comfort and Pleasure," in *Pleasure in Product Use*, P. W. Jordan and B. Green, Eds., Taylor and Francis, London, pp. 321–331.

Consolvo, S., Harrison, B., Smith, I., Chen, M., Everitt, K., Froehlich, J., and Landay, J. (2007), "Conducting in Situ Evaluations for and with Ubiquitous Computing Technologies," *International Journal of Human-Computer Interaction*, Vol. 22, Nos. 1–2, pp. 103–118.

Consumer Reports (2007), "Annual Car Reliability Survey", available: http://www.baileycar.com/CReliability.html.

Crandall, B., Klein, G., and Hoffman, R. (2006), *Working Minds. A Practitioners' Guide to Cognitive Task Analysis*, MIT Press, Cambridge, MA.

Crilly, N., Moultrie, J., and Clarkson, P. J. (2009), "Shaping Things: Intended Consumer Response and the Other Determinants of Product Form," *Design Studies,* Vol. 30, pp. 224–254.

Cross, N. (2000), *Engineering Design Methods: Strategies for Product Design*, 3rd ed., Wiley, Chichester.

Csikszentmihalyi, M. (1997), *Finding Flow*, Basic Books, New York.

Csikszentmihalyi, M. (1990), *Flow: The Psychology of Optimal Experience*, Harper and Row, New York.

Damasio, A. R. (1994), *Descartes' Error: Emotion, Reason, and the Human Brain*, Grosset/Putnam, New York.

Davis, W. J., Rahman, M. A., Smith, L. J., Burns, A., Senecal, L., McArthur, D., Halpern, J. A., Perlmutter, A., Sickels, W., and Wagner, W. (1995), "Properties of Human Affect Induced by Static Color Slides (IAPS): Dimensional, Categorical, and Electromyographic Analysis," *Biological Psychology*, Vol. 41, No. 3, pp. 229–253.

Demir, E., Desmet, P. M. A., and Hekkert, P. (2009), "Appraisal Patterns of Emotions in Human-Product Interaction," *International Journal of Design*, Vol. 3, No. 2, pp. 41–51.

Demirbilek, O., and Sener, B. (2003), "Product Design, Semantics, and Emotional Response," *Ergonomics,* Vol. 46, Nos. 13–14, pp. 1346–1360.

Desmet, P. M. A. (2003), "A Multilayered Model of Product Emotions," *The Design Journal*, Vol. 6, No. 2, pp. 4–13.

Desmet, P. M. A., and Schifferstein, H. N. J. (2008), "Sources of Positive and Negative Emotions in Food Experience," *Appetite*, Vol. 50, pp. 290–301.

Diener, E. (1994), "Assessing Subjective Well-Being: Progress and Opportunities," *Social Indicators Research,* Vol. 31, pp. 103–157.

Dong, A., Kleinsmann, M., and Valkenburg, R. (2009), "Affect-in-Cognition through the Language of Appraisals," *Design Studies*, Vol. 30, pp. 138–153.

Ekman, P. (1982), *Emotion in the Human Face*, Second Edition, Cambridge University Press, New York.

Ekman, P. (1992), "An Argument for Basic Emotions," *Cognition and Emotion*, Vol. 6, pp. 169–200.

Ekman, P. (1994), "Strong Evidence for Universals in Facial Expressions: A Reply to Russell's Mistaken Critique," *Psychological Bulletin,* Vol. 115, No. 2, pp. 268–287.

Ekman, P., and Friesen, W. V. (1976), "Measuring Facial Movement," *Environmental Psychology and Nonverbal Behavior*, Vol. 1, No. 1, pp. 56–75.

Epstein, S. (1994), "Integration of the Cognitive and Psychodynamic Unconscious," *American Psychologist*, Vol. 49, No. 8, pp. 709–724.

Fishbein, M., and Ajzen, I. (1975), *Belief, Attitude, Intention and Behaviour—An Introduction to Theory and Research*, Addison-Wesley, Boston.

Goel, V., and Pirolli, P. (1992), "The Structure of Design Problem Spaces," *Cognitive Science*, Vol. 16, pp. 395–429.

Goh, J. C. (2008), "Expressing Preferences for Affective Design Features in Automobiles," Final Year Project, Nanyang Technological University, Singapore.

Goleman, D. (1995), *Emotional Intelligence: Why It Can Matter More Than IQ*, Bantam Books, New York.

Golightly, D., Lohse, N., Opperud, A., and Khalid, H.M. (2010), "Citarasa Ontology for the Capture and Interpretation of Affective Needs and Design Parameters," *Theoretical Issues in Ergonomics Science*, iFirst. DOI: 10.1080/1463922X.2010.520097.

Hancock, P. A. (2000), personal email communication with Martin Helander.

Hancock, P. A., Pepe, A. A., and Murphy, L. L. (2005), "Hedonomics: The Power of Positive and Pleasurable Ergonomics," *Ergonomics in Design*, Vol. 13, No. 1, pp. 8–14.

Hancock, P. A., Muller, M., Kim, Y., and Mandryk, R. (2008), "Measuring Affect in HCI: Going Beyond the Individual," in *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI 2008),* ACM Press, New York.

Hartono, M., and Tan, K. C. (2009), "A Proposed Framework of Kansei Engineering Application in Dealing with Customer Emotional Needs in Services", in *Proceedings of the 17th Triennial Congress of the International Ergonomics Association*, IEA 2009, Beijing, China (CD-ROM).

Hassenzahl, M. (2004), "The Interplay of Beauty, Goodness, and Usability in Interactive Products," *Human-Computer Interaction*, Vol. 19, No. 4, pp. 319–349.

Hassenzahl, M. (2010), *Experience Design. Technology for the Right Reasons*, Morgan and Claypool Publishers, San Rafael, CA.

Hekkert, P., and Schifferstein, H. N. J. (2008), "Introducing Product Experience," in *Product Experience*, H. N. J. Schifferstein and P. Hekkert, Eds., Elsevier, Amsterdam, pp. 1–8.

Hektner, J. M., Schmidt, J. A., and Csikszentmihalyi, M. (2007), "Experience Sampling," *International Journal of HCI*, Vol. 22, No. 1, pp. 107–122.

Helander, M. G. (1978), "Applicability of Drivers' Electrodermal Response to the Design of the Traffic Environment," *Journal of Applied Psychology*, Vol. 9, pp. 481–488.

Helander, M. G., and Khalid, H. M. (2006), "Affective and Pleasurable Design," in *Handbook on Human Factors and Ergonomics*, G. Salvendy, Ed., Wiley, Hoboken, NJ, pp. 543–572.

Helander, M. G., and Khalid, H. M. (2009), "Citarasa Engineering for Identifying and Evaluating Affective Product Design," in *Proceedings of the 17th Triennial Congress of the International Ergonomics Association, IEA 2009*, Beijing, China (CD-ROM).

Helander, M. G., and Tham, M. P. (2003), "Hedonomics—Affective Human Factors Design," *Ergonomics*, Vol. 46, Nos. 13–14, pp. 1269–1272.

Helander, M. G., and Zhang L. (1997), "Field Studies of Comfort and Discomfort in Sitting," *Ergonomics*, Vol. 40, No. 9, pp. 895–915.

Helander, M. G., Khalid, H. M., and Tham, M. P. (2001), *Proceedings of the International Conference on Affective Human Factors Design*, ASEAN Academic, London.

Helander, M. G., Khalid, H. M., and Peng, H. (2007), "Citarasa—A Framework for Affective Design and Evaluation," in *Proceedings of the 8th International Conference of Work with Computing Systems (WWCS 2007)*, KTH, Stockholm (CD-ROM).

Helander, M. G., Khalid, H. M., Lim, T. Y., Peng, H., and Yang, X. (in press), "Emotional Needs in Car Buyers and Emotional Intent in Car Designers," *Theoretical Issues in Ergonomics Science*.

Hodgetts, H. M., and Jones, D. M. (2007), "Reminders, Alerts, and Pop Ups: The Cost of Computer-Initiated Interruptions," *Human-Computer Interaction*, in J. Jacko, Ed., Part I, HCII 2007, LNCS 4550, Springer-Verlag, Berlin, Heidelberg, pp. 818–826.

Hoffman, D. L., and Novak, T. P. (1996), "Marketing in Hypermedia Computer-Mediated Environments: Conceptual Foundations," *Journal of Marketing*, Vol. 60, No. 3, pp. 50–68.

Housen, A. (1992), "Validating a Measure of Aesthetic Development for Museums and Schools," *ILVS Review*, p. 2.

Isen, A. M., and Erez, A. (2006), "Some Measurement Issues in the Study of Affect," in *Oxford Handbook of Methods in Positive Psychology*, A. Ong and M. vanDulman, Eds., Oxford University Press, New York.

Ishihara, S., Ishihara, K., and Nagamachi, M. (2009), "Statistical Visualization Methods for Tabular Data in Kansei Engineering," in *Proceedings of the 17th Triennial Congress of the International Ergonomics Association, IEA 2009*, Beijing, China (CD-ROM).

Izard, C. E. (1977), *Human Emotions,* Plenum Press, New York.

Izard, C. E. (1979), *The Maximally Discriminative Facial Movement Coding System (MAX)*, Instructional Resources Centre, University of Delaware, Newark.

Izard, C. E. (1992), "Basic Emotions, Relationships between Emotions, and Emotion-Cognition Relations," *Psychological Review,* Vol. 99, pp. 561–567.

Jenkins, S., Brown, R., and Rutterford, N. (2009), "Comparing Thermographic, EEG, and Subjective Measures of Affective Experience during Simulated Product Interactions," *International Journal of Design*, Vol. 3, No. 2, pp. 53–66.

Jiao, J., Zhang, Y., and Helander, M.G. (2006), "A Kansei Mining System for Affective Design," *Expert Systems with Applications*, Vol. 30, No. 4, pp. 658–673.

Johnson, E. J., and Tversky, A. (1983), "Affect, Generalization, and the Perception of Risk," *Journal of Personality and Social Psychology,* Vol. 45, pp. 21–31.

Johnson, D., and Wiles, J. (2003), "Effective Affective User Interface Design in Games," *Ergonomics,* Vol. 46, Nos. 13–14, pp. 1332–1345.

Jordan, P. W. (1998), "Human Factors for Pleasure in Product Use," *Applied Ergonomics*, Vol. 29, No. 1, pp. 25–33.

Jordan, P. W. (2000), "The Four Pleasures—A Framework for Pleasures in Design," in *Proceedings of Conference on Pleasure Based Human Factors Design,* P. W. Jordan, Ed., Philips Design, Groningen, The Netherlands.

Jordan, P. W. (2002), *How to Make Brilliant Stuff That People Love and Make Big Money Out of It*, Wiley, Chichester.

Kahneman, D., Diener, E., and Schwarz, N., Eds. (1999), *Well-Being: Foundations of Hedonic Psychology,* Russell Sage Foundation, New York.

Kapoor, A., and Horvitz, E. (2008), "Experience Sampling for Building Predictive User Models: A Comparative Study," in *Proceedings of the 26th Annual CHI Conference on Human Factors in Computing Systems (CHI '08)*, ACM, New York, pp. 657–666.

Karlsson, B. S. A., Aronsson, N., and Svensson, K. A. (2003), "Using Semantic Environment Description as a Tool to Evaluate Car Interiors," *Ergonomics*, Vol. 46, Nos. 13–14, pp. 1408–1422.

Kato, T. (2006), "Kansei Robotics, Measurement and Modeling of Kansei from Robotics Aspect," *Journal of Japanese Society for Artificial Intelligence*, Vol. 21, No. 2, pp. 183–188.

Ketelaar, T. (1989), "Examining the Circumplex Model of Affect in the Domain of Mood-Sensitive Tasks," Master's Thesis, Purdue University, West Lafayette, IN.

Khalid, H. M. (2004), "Conceptualizing Affective Human Factors Design," *Theoretical Issues in Ergonomics Science*, Vol. 5, No. 1, pp. 1–3.

Khalid, H. M. (2006), "Embracing Diversity in User Needs for Affective Design," *Applied Ergonomics*, Vol. 37, pp. 409–418.

Khalid, H. M. (2008), "Affective Design as Ergonomics Lifestyle," in *Proceedings of the 40th Annual Conference of the Nordic Ergonomics Society (NES 2008)*, Icelandic Ergonomics Society, Reyjkavik (CD-ROM).

Khalid, H. M., and Helander, M. G. (2004), "A Framework for Affective Customer Needs in Product Design," *Theoretical Issues in Ergonomics Science,* Vol. 5, No. 1, pp. 27–42.

Khalid, H. M., Dangelmaier, M., and Lim, T. Y. (2007a), "The CATER Approach to Vehicle Mass Customization," in *Proceedings of IEEE International Conference on Industrial Engineering and Engineering Management (IEEM2007),* pp. 1273–1276.

Khalid, H. M., Opperud, A., Helander, M. G., Seccomaria, G., Gemou, M., Cobb, S., and Genoux, F. (2007b), "D2.1 Customer Needs Data," Technical Report for CATER Project FP6 IST-5-035030, European Commission, Brussels.

Khalid, H. M., Abd Karim, H. R., and Radha, J. K. (2009), "Citarasa System for Elicitation and Analysis of Affective Needs," in *Proceedings of IEA Congress 2009*, Beijing, August 9–14, 2009 (CD-ROM).

Khalid, H. M., Radha, J. K., Yang, X., and Helander, M. G. (2010), "Analyzing ABC of Disaster Experience Using Text Mining," in *Advances in Cross-Cultural Decision*, D. Schmorrow and D. Nicholson, Eds., CRC Press/Taylor and Francis, London.

Khalid, H. M., Opperud, A., Radha, J. K., Xu, Q., and Helander, M.G. (2011), "Elicitation and Analysis of Affective Needs in Vehicle Design," *Theoretical Issues in Ergonomics Science*, iFIRST, 1–17. DOI (10.1080/1463922X.2010.506559)

Khangura, K. (2009), "User Centric Design and Kansei Engineering," *CIRP Journal of Manufacturing Science and Technology*, Vol. 1, No. 3, pp. 172–178.

Khoo, K. L. U. (2007), "Using Cognitive Task Analysis to Document Customer Expertise in Buying a Car," Final Year Project, School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore.

Killgore, W. D. (1998), "The Affect Grid: A Moderately Valid, Nonspecific Measure of Pleasure and Arousal," *Psychological Reports*, Vol. 83, No. 2, pp. 639–642.

King, S. C., and Meiselman, H. L. (2009), "Development of a Method to Measure Consumer Emotions Associated with Foods," *Food Quality and Preference*, Vol. 21, pp. 168–177.

Krippendorff, K. (2006), *The Semantic Turn: A New Foundation for Design*, Taylor and Francis/CRC Press, Boca Raton, FL.

Kuang, J., and Jiang, P. (2009), "Product Platform Design for a Product Family Based on Kansei Engineering," *Journal of Engineering Design*, Vol. 20, No. 6, pp. 589–607.

Kubovy, M. (1999), "On the Pleasures of the Mind," in *Well-Being: The Foundations of Hedonic Psychology,* D. Kahneman, E. Diener, and N. Schwarz, Eds., Russell Sage Foundation, New York, pp. 134–154.

Küller, R. (1975), *Semantisk Milöbeskriving (SMB)* (in Swedish) [Semantic Descriptions of Environments (Impression of Colors and Colored Environments)], Byggforskningsrådet, Stockholm, Sweden.

Larsen, R. J., and Fredrickson B. (1999), "Measurement Issues in Emotion Research," in *Well-Being: The Foundations of Hedonic Psychology,* D. Kahneman, E. Diener, and N. Schwarz, Eds., Russell Sage Foundation, New York, pp. 40–60.

Larson, R., and Csikszentmihalyi, M. (1983), "The Experience Sampling Method," *New Directions for Methodology of Social and Behavioral Science,* Vol. 15, pp. 41–56.

LeDoux, J.E. (2003). *Synaptic Self: How Our Brains Become Who We Are*, Penguin Putnam, New York.

LeDoux, J. E. (1995), "Emotion: Clues from the Brain," *Annual Review of Psychology,* Vol. 46, pp. 209–235.

Leontjev, A. N. (1978), *Activity, Consciousness and Personality,* Prentice-Hall, London.

Lerner, J. S., Small, D. A., and Loewenstein, G. (2004), "Heart Strings and Purse Strings: Carry Over Effects of Emotions on Economic Decisions," *Psychological Science,* Vol. 15, No. 5, pp. 337–341.

Lew, G. S. (2009), "What Do Users Really Do? Experience Sampling in the 21st Century," Lecture Notes In Computer Science, Vol. 5610, in *Proceedings of the 13th International Conference on Human-Computer Interaction. Part I: New Trends*, Springer Verlag, San Diego, CA.

Lin, Y., Lai, H, and Yeh, C. (2007), "Consumer-Oriented Product Form Design Based on Fuzzy Logic: A Case Study of Mobile Phones," *International Journal of Industrial Ergonomics,* Vol. 37, No. 6, pp. 531–543.

Lottridge, D. (2008), "Emotional Response as a Measure of Human Performance," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08),* ACM Press, New York, pp. 2617–2620.

Maffiolo, V., and Chateau, N. (2003), "The Emotional Quality of Speech in Voice Services," *Ergonomics,* Vol. 46, Nos. 13–14, pp. 1375–1385.

Maffiolo, V., Chateau, N., and Le Chenadec, G. (2007), "Temporal hierarchical organization in listeners' perception of the speakers' emotions and characteristics: a way to improve the automatic recognition of emotion-related states in human voice," in *Proceedings of the HUMAINE International Conferences on Affective Computing & Intelligent Interaction (ACII'2007. Available at http://cordis.europa.eu/*.

Maslow, A. (1968), *Towards a Psychology of Being*, Van Nostrand, Princeton, NJ.

Mayer, J. D., and Bremer, D. (1985), "Assessing Mood with Affect Sensitive Tasks," *Journal of Personality Assessment,* Vol. 49, pp. 95–99.

McNair, D. M., Lorr, M., and Droppleman, L. F. (1971), *Profile of Mood States*, Educational and Industrial Testing Service, San Diego.

Minsky, M. (2007), *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*, Simon and Shuster, New York.

Monk, A. F., Hassenzahl, M., Blythe, M., and Reed, D. J. (2002), "Funology: Designing Enjoyment," *Extended Abstracts, of the ACM Conference on Human Factors in Computing Systems (CHI 2002),* ACM Press, New York, pp. 924–925.

Nagamachi, M. (2001), *Research on Kansei Engineering: Selected Papers on Kansei Engineering*, Nakamoto Printing, Hiroshima, Japan.

Nagamachi, M. (2008), "Perspectives and the New Trend of Kansei/Affective Engineering", *The TQM Journal*, Vol. 20, No. 4, pp. 290–298.

Nagamachi, M. (2010), Kansei/Affective Engineering-Industrial Innovation v.18. CRC Press, London.

Nagamachi, M., and Lokman, M. A. (2010), *Innovations in Kansei Engineering*, CRC Press, London.

Niedenthal, P. M., and Setterlund, M. B. (1994), "Emotion Congruence in Perception," *Personality and Social Psychology Bulletin,* Vol. 20, pp. 401–411.

Norman, D. (1991), "Cognitive Artifacts," in *Designing Interaction: Psychology at the Human-Computer Interface*, J. Carroll, Ed., Cambridge University Press, Cambridge, pp. 17–38.

Norman, D. A. (2005), *Emotional Design: Why Do We Love (or Hate) Everyday Things*, Basic Books, New York.

Nowlis, V., and Green, R. (1957), "The Experimental Analysis of Mood," Technical Report, Nonr-668(12), Office of Naval Research, Washington, DC.

Osgood, C., Suci, G., and Tannenbaum, P. (1967), *The Measurement of Meaning*, University of Illinois Press, Urbana-Champaign, IL.

Peter, J. P., and Olson, J. C. (1996), *Consumer Behavior and Marketing Strategy*, Richard D. Irwin, Chicago, IL.

Phelps, E. A. (2006), "Emotion and Cognition: Insights from Studies of the Human Amygdala," *Annual Review of Psychology*, Vol. 57, pp. 27–53.

Picard, R. W. (1997), *Affective Computing*, MIT Press, Cambridge, MA.

Picard, R. W. (2000), "Towards Computer That Recognize and Respond to User Emotion," *IBM Systems Journal*, Vol. 39, pp. 3–4.

Rapaille, C. (2006), *The Culture Code: An Ingenious Way to Understand Why People Around the World Buy and Live as They Do*, Random House, New York.

Rasmussen, J., Pejtersen, M. A., and Goodstein, L. P. (1994), *Cognitive Systems Engineering,* Wiley, New York.

Rosenberg, M. J., and Hovland, C. I. (1960), "Cognitive, Affective and Behavioral Components of Attitudes," in *Attitude Organization and Change*, C. I. Hovland and M. J. Rosenberg, Eds., Yale University Press, New Haven, CT, pp. 1–14.

Russell, J. A., Weiss, A., and Mendelsohn, G. A. (1989), "The Affect Grid: A Single-Item Scale of Pleasure and Arousal," *Journal of Personality and Social Psychology,* Vol. 57, pp. 493–502.

Saerbeck, M., and Bartneck, C. (2010), "Attribution of Affect to Robot Motion," in *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, *HRI2010,* ACM Press, New York, pp. 53–60.

Sandvik, E., Diener, E., and Seidlitz, L. (1993), "Subjective Well-Being: The Convergence and Stability of Self-Report and Non Self-Report Measures," *Journal of Personality,* Vol. 61, pp. 317–342.

Scherer, K. R. (1986), "Vocal Affect Expression: A Review and a Model for Future Research," *Psychological Bulletin*, Vol. 99, pp. 143–165.

Scherer, K. R. (1998), "Analyzing Emotion Blends," in *Proceedings of the 10th Conference of the International Society for Research on Emotions*, A. Fischer, Ed., ISRE Publications, Storrs, CT, pp. 142–148.

Schimmack, U. (2003), "Affect Measurement in Experience Sampling Research," *Journal of Happiness Studies,* Vol. 4, pp. 79–106.

Schneider, W., and Shiffrin, R. M. (1997), "Controlled and Automatic Human Information Processing: I. Detection, Search and Attention," *Psychological Review,* Vol. 84, No. 1, pp. 1–66.

Schutte, S., and Eklund, J. (2005), "Design of Rocker Switches for Work Vehicles—An Application of Kansei Engineering," *Applied Ergonomics*, Vol. 236, No. 5, pp. 557–567.

Schütte, S., Eklund, J., Ishihara, S., and Nagamachi, M. (2008), "Affective Meaning: The Kansei Engineering Approach," in *Product Experience,* 1st ed., H. N. J. Schifferstein and P. Hekkert, Eds., Elsevier, Oxford, pp. 477–496.

Seligman, M. E. P., and Csikszentmihalyi, M. (2000), "Positive Psychology: An Introduction," *American Psychologist,* Vol. 55, pp. 5–14.

Seva, R. S., and Helander, M. G (2009), "The Influence of Cellular Phone Attributes on Users' Affective Experiences: A Cultural Comparison," *International Journal of Industrial Ergonomics*, Vol. 39, pp. 341–346.

Shami, N. S., Hancock, J. T., Peter, C., Muller, M., Kim, Y., and Mandryk, R. (2008), "Measuring Affect in HCI: Going Beyond the Individual," in *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI 2008),* ACM Press, New York.

Simon, H. A. (1969), *The Science of the Artificial*, MIT Press, Cambridge, MA.

Thaler, R. H. (1980), "Toward a Positive Theory of Consumer Choice," *Journal of Economic Behavior and Organizations,* Vol. 1, pp. 39–60.

Tiger, L. (1992), *The Pursuit of Pleasure,* Little Brown, Boston.

Titchener, E. B. (1910), "The Past Decade in Experimental Psychology," *American Journal of Psychology,* Vol. 21, pp. 404–421.

Triesman, A., and Souther, J. (1985), "Search Asymmetry: A Diagnostic for Preattentive Processing of Separable Features," *Journal of Experimental Psychology: General*, Vol. 114, pp. 285–310.

Turner, M., Love, S., and Howell, M. (2008), "Understanding Emotions Experienced When Using a Mobile Phone in Public: The Social Usability of Mobile (Cellular) Telephones," *Telematics and Informatics*, Vol. 25, pp. 201–215.

Tversky, A. (1972), "Elimination by Aspect: A Theory of Choice," *Psychological Review,* Vol. 61, pp. 281–289.

Ullman, D. G. (2009), *The Mechanical Design Process*, 4th ed., Mc-Graw Hill, New York.

Vastenburg, M. H., and Herrera, N. R. (2010), "Adaptive Experience Sampling: Addressing the Dynamic Nature of In-Situ User Studies," *ISAmI International Symposium on Ambient Intelligence*, Guimaraes, Portugal, Springer Advances in Soft Computing, Vol. 72/2010, pp. 197–200.

Velásquez, J. D. (1998), "When Robots Weep: Emotional Memories and Decision-Making," in *Proceedings of the 15th National Conference on Artificial Intelligence*, AAAI Press, Menlo Park, CA, pp. 70–75.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009), "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition," *Perspectives on Psychological Science,* Vol. 4, pp. 274–290.

Vygotsky, L. (1962), *Thought and Language*, MIT Press, Cambridge, MA.

Warr, P. (1999), "Well-Being and the Workplace," in *Well-Being: Foundations of Hedonic Psychology,* D. Kahneman, E. Diener, and N. Schwarz, Eds., Russell Sage Foundation, New York, pp. 392–412.

Watson, D., Lee, A. C., and Tellegen, A. (1988), "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales," *Journal of Personality and Social Psychology,* Vol. 54, No. 6, pp. 1063–1070.

Webster, J., Trevino, L. K., and Ryan, L. (1993), "The Dimensionality and Correlates of Flow in Human-Computer Interactions," *Computers in Human Behavior,* Vol. 9, pp. 411–426.

Wellings, T., Williams, M., and Tennant, C. (2010), "Understanding Customers' Holistic Perception of Switches in Automotive Human-Machine Interfaces," *Applied Ergonomics*, Vol. 41, pp. 8–17.

Xu, Q., Jiao, R. J., Yang, X., Helander, M., Khalid, H. M., and Opperud, A. (2009), "An Analytical Kano Model for Customer Need Analysis," *Design Studies*, Vol. 30, No. 1, pp. 87–110.

Yalch, R., and Brunel, F. (1996). "Need Hierarchies in Consumer Judgments of Product Designs: Is It Time to Reconsider Maslow's Theory?" *Advances in Consumer Research*, Vol. 23, pp. 405–410.

Zajonc, R. B. (1998), "Emotions," in *The Handbook of Social Psychology,* D. Gilbert, S. Fiske, and G. Lindzey, Eds., Oxford University Press, New York, pp. 591–632.

Zuckerman, M., and Lubin B. (1965), *The Multiple Affect Adjective Check List,* Educational and Industrial Testing Service, San Diego, CA.

# EQUIPMENT, WORKPLACE, AND ENVIRONMENTAL DESIGN

# CHAPTER 21

# WORKPLACE DESIGN

**Nicolas Marmaras and Dimitris Nathanael**
**National Technical University of Athens**
**Athens, Greece**

## 1 INTRODUCTION

Workplace design deals with the shape, the dimensions, and the layout (i.e., the placement and orientation) of the different material elements that surround one or more working persons. Examples of such elements are the seat, working surfaces, desk, equipment, tools, controls, and displays used during the work as well as the passages, windows, and heating/cooling equipment.

The ergonomic workplace design aims at improving work performance (both in quantity and quality) as well as ensuring occupational safety and health through:

- Minimizing the physical workload and the associated strain on the working person
- Facilitating task execution, that is, ensuring effortless information exchange with the environment, minimization of the physical constraints, and so on
- Achieving ease of use of the various workplace elements

Putting together a workplace which meets ergonomics requirements while at the same time satisfies task demands is not a trivial problem. In fact, to achieve this one should consider an important number of interacting and variable elements, and try to meet many requirements, some of which may be contradictory. As shown in Figure 1, in any work setting there is a continuous mutual adjustment between the workplace components, the task demands, and the working person.

This mutual adjustment is also subject to broader environmental conditions. Therefore, regardless of how well each individual component of the workplace is designed, the habitual body movement and postures in everyday work emerge by an exploration of the constraints and affordances of the workplace as a whole.

Consider, for example, a person working in a computerized office (task demand: work with a computer). If the desk (workplace component 1) is too low and the seat (workplace component 2) is too high for the anthropometric characteristics of the worker (characteristic of the working person), the worker will lean forward (awkward posture), with negative effects on his or her physical workload, health (particularly if he or she should work for a long period in this workplace), and finally overall performance. Furthermore, if behind the worker there is a window causing glare on the computer's screen (characteristic of the environment), he or she will probably bend sideways (awkward posture) in order to be able to see what is presented on the screen (task demand), causing similar effects. Consequently, when designing a workplace, one has to adopt a systemic view, considering at least the characteristics of the working person, the task demands, and the environment in which the task will be performed.

Furthermore, the elements of the work system are variable. The task demands may be multiple and variable. For example, at a secretarial workstation, the task may require exclusive use of the computer for a period of time, then data entry from paper forms, and then face-to-face contact with visitors. At the same time, the secretary should be able to monitor both the
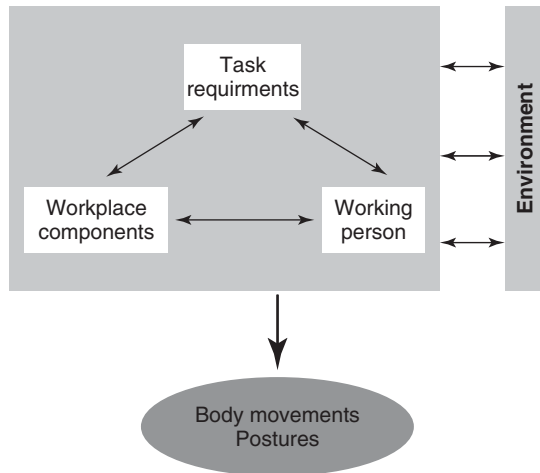
**Figure 1** There is interdependence between the working person, task demands, workplace elements, environment, and body movements and postures.

entry and the director's doors. Finally, the workplace environment may be noisy or quiet, warm or cool, with annoying air streams, illuminated by natural or artificial light, and all the above may change during the course of a working day.

If to the complexity of the work system and the multiplicity of ergonomics criteria one adds the financial and aesthetic issues, successful design of a workplace becomes extremely complex. Hence, some people maintain that designing a good workplace is more an "art" than a "discipline" as there is no standard theory or method that ensures a successful result, the output depending heavily on the designer's "inspiration". Although this is true to a certain extent, good knowledge of the characteristics of the working persons who will occupy the workplace, of the tasks' demands, and of the broader environment, combined with an effort for discipline during the design process, contribute decisively to a successful design.

### 1.1 Importance of Satisfying Task Demands

Despite the multiple external determinants, the workplace still leaves many degrees of freedom to the working person, who can exploit them in more than one ways. As stated above, the habitual body movement and

postures in everyday work emerge by an exploration of the constraints and affordances of the workplace. This exploration and adaptation process can be considered as a control task: The working person, exploring the constraints and affordances of the workplace, tries to achieve an optimal balance between multiple demands related to the task, his or her physical abilities, and perceived comfort.

This control task can be approximated by a cybernetic model such as the one depicted in Figure 2 (Marmaras et al., 2008). According to this model, the postures adopted by the working persons are under the influence of two nested feedback loops: (a) a positive-feedback loop regarding the satisfaction of task demands (e.g., easy reading and writing for an office worker) and (b) a negative-feedback loop regarding the attenuation of the perceived physical strain and pain due to the body postures and the eventual disorders built because of them. These two loops work toward different objectives (i.e., meeting task demands vs. comfort satisfaction). The simultaneous satisfaction of both objectives may be conflicting. In such situations their resolution involves a trade-off which will be moderated by the feedback power and the pace of incoming information for each of the two loops. However, the two loops operate on different time scales. The positive one regarding the satisfaction of task demands is immediate and constantly perceived, easily linked to the particular arrangement of the various workplace components, and equally easily interpretable (e.g., if an office worker cannot access the keyboard because it is probably placed too far, he or she will move either the keyboard or the chair or extend his or her upper limbs). The negative loop, regarding the attenuation of the perceived physical strain and pain, takes time to be perceived as it has a cumulative character, requiring prolonged exposure (e.g., it takes months or years for musculoskeletal disorders to be built). Furthermore, such feedback is not easily interpretable and attributable to either postures or workplace settings by a nonexpert (e.g., even if back pain is felt, it is difficult for an individual to attribute it to a specific posture or workplace setting).

We can argue therefore, that workers' postures are more readily affected by the positive-feedback loop, which as a constant attractor forces the system to be self-organized in a way that favors the satisfaction of task demands. Such an argument has already been validated by Dainoff (1994) in a research conducted in laboratory settings. This research indicated that
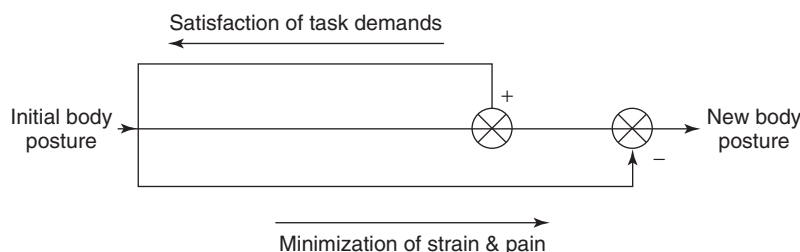


**Figure 2** Cybernetic model depicting working person's control task related to body posture modifications.

participants performing a high-speed data entry task found it effective to sit in the forward-tilt posture, while participants who performed screen-based editing tasks (with very low keying requirements) found the backward-leaning posture more effective.

The model presented above stresses both the existence of regulation mechanisms operating continuously as part of the working person's activities and the need to put the task demands and resulting work activities at the center of the design process. It is the latter that makes the ergonomic design synonymous to the user-centered design.

The present chapter is mainly methodological; it presents and discusses a number of methods, techniques, guidelines, and typical design solutions which aim to support the decisions to be taken during the workplace design process. The next section discusses the problem of working postures and stresses the fact that there is no *one best posture* which can be assumed for long periods of time. Consequently, the effort should be put on designing the components of the workplace in such a way as to form a "malleable envelope" that permits the working persons to adopt various healthy postures. The two remaining sections deal with the design of individual workstations and with the layout of groups of workstations in a given space.

## 2   PROBLEM OF WORKING POSTURES

A central issue of the ergonomic workplace design is the postures the working person will adopt. In fact, the decisions made during the workplace design will affect to a great extent the postures that the working person will be able to adopt or not. The two most common working postures are sitting and standing. Between the two, the sitting posture is of course more comfortable. However, there is research evidence that sitting adopted for prolonged periods of time results in discomfort, aches, or even irreversible injuries. For example, Figure 3 shows the most common musculoskeletal disorders encountered at office workstations.

Studying the effects of "postural fixity" while sitting, Griego (1986) found that it causes, among others, (i) reduction of nutritional exchanges at the spine disks and in the long term may promote their degeneration, (ii) static loading of the back and shoulder muscles, which can result in aches and cramping, and (iii) restriction in blood flow to the legs, which can cause swelling (edema) and discomfort. Consequently, the following conclusion can be drawn: The workplace should permit the alteration between various postures because there is no "ideal" posture which can be adopted for a long period of time.

Based on this conclusion, the standing–sitting workstation has been proposed, especially for cases where the task requires long periods of continuous work (e.g., bank tellers or assembly workstations). This workstation permits to perform a job alternating the standing with the sitting posture (see Figure 4 for an example).

Despite the absence of an ideal posture, there are however postures which are more comfortable and healthy than others. The ergonomic research aims at identifying these postures and formulating requirements and principles which should be considered during the design of the components of a workplace. In this way the resulting design will promote healthy work postures and constrain the prolonged adoption of unhealthy postures.
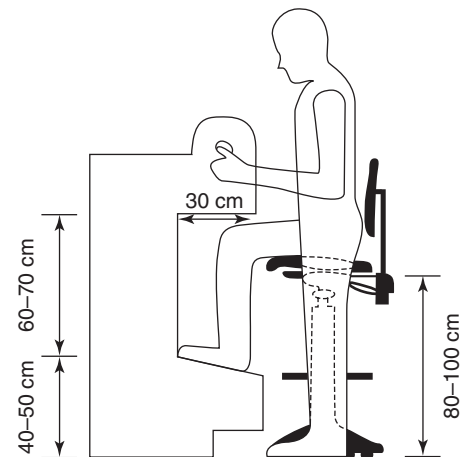


**Figure 4**   Example of standing–sitting workstation.
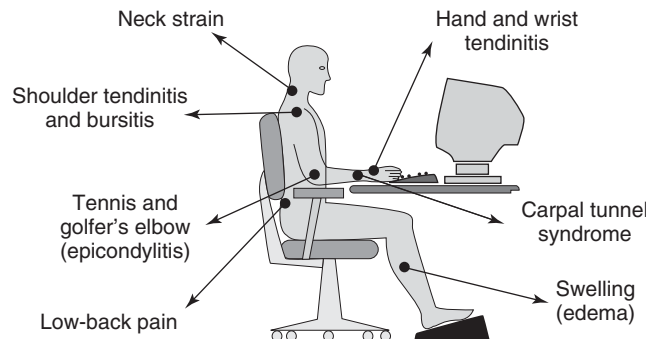


**Figure 3**   Common musculoskeletal disorders encountered at office workstations.

## 2.1  Sitting Posture and Seats

The problem of designing seats that are appropriate for work is far from solved. In recent decades the sitting posture and the design of seats have attracted the interest of researchers, designers, and manufacturers due to the ever-increasing number of office workers and the importance of musculoskeletal problems encountered by them. This has resulted in the emergence of a proper research domain and subsequently to a plethora of publications and design solutions (see, e.g., Lueder and Noro, 1994; Mandal, 1985; Marras, 2005; Corlett, 2009).

As already stated, sitting posture poses a number of problems at a musculoskeletal level. One of the more important of them is *lumbar kyphosis*. When one is sitting, the lumbar region of the back flattens out and may even assume an outward bend. This shape of the spine is called *kyphotic* and is somewhat the opposite to the *lordotic* shape of the spine when someone is standing erect (Figure 5). The more the angle between the thighs and the body is smaller, the greater the kyphosis. This occurs because of the restrained rotation of the hip joint, which forces the pelvis to rotate backward. Kyphosis provokes increased pressure on the spine disks at the lumbar portion. Nachemson and Elfstrom (1970), for example, found that unsupported sitting in upright posture resulted in a 40% increase in the disks' pressure compared to the pressure when standing. There are three complementary ways to minimize lumbar kyphosis: (i) by using a thick lumbar support; (ii) by reclining the backrest; and (iii) by providing a forward-tilting seat. Andersson et al. (1979) found that the use of a 4-cm-thick lumbar support combined with a backrest recline of 110° resulted in a lumbar curve resembling closely the lumbar curve of a standing person. Another finding of Andersson et al. (1979) was that the exact location of the support within the lumbar region did not significantly influence any of the angles measured in the lumbar region. The studies
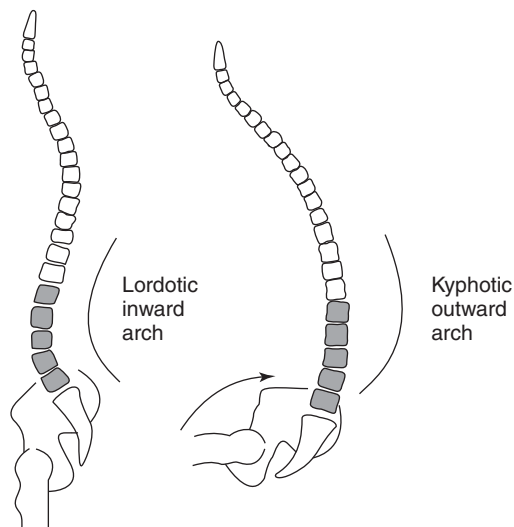
of Bendix (1986) and Bridger (1988) support the proposition of Mandal (1985) for the forward-tilting seat.

Considering the above, the following ergonomics requirements should be met:

1. The seats should dispose a backrest which can recline.
2. The backrest should provide a lumbar support.
3. The seat should provide a forward-tilting seat.

However, as Dainoff (1994) observes, when tasks require close attention to the objects on the working surface or the computer screen, people usually bend forward, and the backrest support becomes useless.

A design solution which aims to minimize lumbar kyphosis is the *kneeling* or *balance chair* (Figure 6), where the seat is inclined more than 20° from the horizontal plane. Besides the somewhat unusual way of sitting, this chair has also the drawbacks of loading the area of knees as they receive a great part of the body's load and of constraining the legs' movements. On the other hand, it enforces a lumbar lordosis very close to the one adopted while standing and does not constrain the torso to move freely forward, backward, or sideways.

There are quite a lot of detailed ergonomic requirements concerning the design of seats used at work. For example:

- The seat should be adjustable in order to fit to the various anthropometric characteristics of their users as well as to different working heights.
- The seat should offer stability to the user.
- The seat should offer freedom of movement to the user.
- The seat should be equipped with armrests.
- The seat lining material should be water absorbent to absorb body perspiration.

The detailed requirements will not be presented extensively here, as the interested reader can find them



**Figure 5**  Lordotic and kyphotic postures of the spine (Grandjean, 1987).
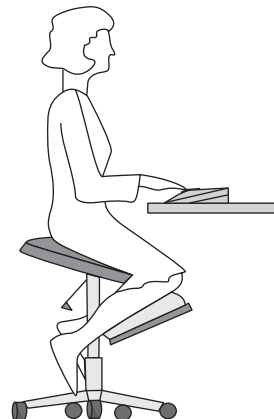


**Figure 6**  Example of a kneeling chair (source: www.comcare.gov.au/officewise.html).

easily in any specialized handbook. Furthermore, these requirements became "classical" and have been adopted by regulatory documents such as health and safety or design standards, legislation, and so on [e.g., European standard EN 1335, International Organization for Standardization (ISO) 9241, American National Standards Institure (ANSI)/HFS 100-1988, and the German standard DIN 4543 for office work or EN 1729 for chairs and tables for educational institutions and ISO/DIS 16121 for the driver's workplace in line-service buses].

Although most of the modern seats for office work meet the basic ergonomics requirements, the design of their controls does not meet the usability principles. This fact, combined with the poor users' knowledge on healthy sitting, results in the nonuse of the adjustment possibilities offered by the seats (Vitalis et al., 2000). Lueder (1986) provides the following guidelines for increasing the usability of controls:

- Controls should be easy to find and interpret.
- Controls should be easily reached and adjusted from the standard seated work position.
- Controls should provide immediate feedback (e.g., seats that adjust in height by rotating pan delay feedback because user must get up and down repeatedly to determine the correct position).
- The direction of operation of controls should be logical and consistent with their effect.
- Few motions should be required to use the controls.
- Adjustments should require the use of only one hand.
- Special tools should not be necessary for the adjustment.
- Labels and instructions on the furniture should be easy to understand.

However, modern office chairs are still far from meeting satisfactorily the above guidelines (Groenesteijna et al., 2009).

## 2.2 Sitting Posture and Work Surface Height

Besides the problem of lumbar kyphosis, sitting working posture may also provoke excessive muscle strain at the level of the back and the shoulders. For example, if the working surface is too low, the person will bend forward too far; if it is too high, he or she will be forced to raise the shoulders.

To minimize these problems, appropriate design of the workplace is required. More specifically, the working surface should be at a height that permits a person to work with the shoulders at the relaxed posture. It should be noticed here that the working height does not always equate to the work surface height. The former depends on what one is working on (e.g., the keyboard of a computer), while the later is the height of the upper surface of the table, desk, bench, and so on. Furthermore, to define the appropriate work surface height, one should consider the angles between the upper arms and the

elbows and the angle between the elbows and the wrists. To increase comfort and minimize the occupational risks, the first of the two angles should be about 90° if no force is required and a little bit broader if application of force is required. The wrists should be straight as far as possible in order to avoid carpal tunnel syndrome.

Two other common problems encountered by people working in the sitting posture are neck aches and dry-eye syndrome. These problems are related to the prolonged gazing at objects placed too high, for example, when the visual display terminal of a computer workstation is placed too high (Ankrum, 1997). The research which aims at determining the optimal placement of such objects, considering the mechanisms of both the visual and musculoskeletal systems, is still active (for a review see Ankrum and Nemeth, 2000). However, most research findings agree that (i) neck flexion is more comfortable than extension, with the zero point (dividing flexion from extension) described as the posture of the head/neck when standing erect and looking at a visual target 15° below eye level, and (ii) the visual system prefers downward gaze angles. Furthermore, there is evidence that when assuming an erect posture, people prefer to tilt their head, with the ear–eye line (i.e., the line which crosses by the cartilaginous protrusion in front of the ear hole and the outer slit in the eyelid) being about 15° below the horizontal plane (Grey et al., 1966; Jampel and Shi, 1992). Based on these findings many authors propose the following rule of thumb for the placement of the monitor: The center of the monitor should be placed at a minimum of 15° below the eye level, with the top and the bottom at an equal distance from the eyes. (i.e., the screen plane should be facing slightly upward).

Sanders and McCormick (1992) propose in addition the following general ergonomics recommendations for work surfaces:

- If at all possible the work surface height should be adjustable to fit individual physical dimensions and preferences.
- The work surface should be at a level that places the working height at elbow height, with shoulders at relaxed posture.
- The work surface should provide adequate clearance for a person's thighs under the work surface.

## 2.3 Spatial Arrangement of Work Artifacts

While working one uses a number of artifacts, for example, the controls and displays on a control panel, the different parts of an assembled object at an assembly workstation, or the keyboard, the mouse, the visual display terminal, the hard-copy documents, and the telephone at an office workstation. Application of the following ergonomic recommendations for the arrangement of these artifacts helps to decrease workload, facilitate the work flow, and improve overall performance (adapted from Sanders and McCormick, 1992):

- *Frequency of Use and Criticality*. Artifacts that are frequently used or are of special importance

should be placed in prominent positions, for example, in the center of the work surface or near the right hand for right-hand people and vice versa for left-hand people.

- *Sequential Consistency*. When a particular procedure is always executed in a sequential order, the artifacts involved should be arranged according to this order.
- *Topological Consistency*. Where the physical location of controlled elements is important for the work, the layout of the controlling artifacts should reflect the geographical arrangement of the former.
- *Functional Grouping*. Artifacts (e.g., dials, controls, visual displays) that are related to a particular function should be grouped together.

Application of the above recommendations requires detailed knowledge of the task demands. Task analysis provides enough data to appropriately apply these recommendations as well as solve eventual contradictions between them by deciding which arrangement best fits the situation at hand.

## 3  DESIGNING INDIVIDUAL WORKSTATIONS

Figure 7 presents a generic process for the design of individual workstations, with the various phases, the data or sources of data to be considered at each phase, and methods that could be applied. It has to be noted that certain phases of the process may be carried out concurrently or in a different order depending on the particularities of the workstation to design or the preferences and experience of the designers.
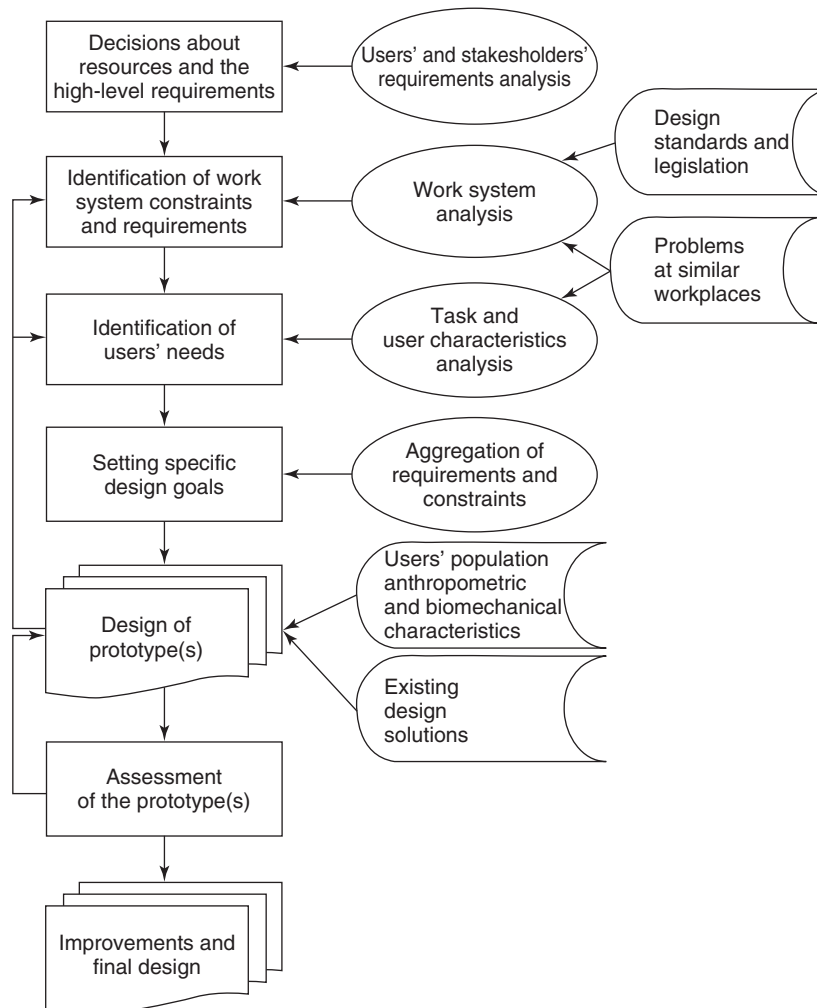


**Figure 7**  Generic process for the ergonomic design of individual.

### 3.1 Phase 1: Decisions about Resources and High-Level Requirements

The aim of the first phase of the design process is to decide the time to spend and the people who will participate in the design team. These decisions depend on the high-level requirements of the stakeholders (e.g., improvement of working conditions, increase of productivity, innovation, occupational safety and health protection) as well as the money they are ready to spend and the importance of the project (e.g., number of identical workstations, significance of the tasks carried out, special characteristics of the working persons). An additional issue that has to be dealt with in this phase is to ensure participation in the design team of representatives of the people who will occupy the future workstations. The access to workstations where similar jobs are being performed is also advisable.

The rest of the design process will be significantly influenced by the decisions made in this phase.

### 3.2 Phase 2: Identification of Work System Constraints and Requirements

The aim of this phase is to identify the different constraints and requirements imposed by the work system in which the workstation will be installed. More specifically, during this phase the design team has to collect data about:

- Types of tasks to be carried out at the workstation
- Work organization, for example, working hours, interdependencies between the tasks to be carried out at the workstation, and other tasks or organizational entities in the proximal environment
- Various technological equipment and tools that will be used, their functions and manipulation, their shape and dimensions, and user interfaces
- Environmental conditions of the broader area in which the workstation will be installed (e.g., illumination and sources of light, level of noise and noise sources, thermal conditions, and sources of warm or cold draughts)
- Normal as well exceptional situations in which the working persons could be found (e.g., electricity breakdowns, fire)
- Any other element or situation of the work system that may directly or indirectly interfere with the workstation

These data can be collected by questioning the appropriate people as well as observation and analysis of similar work situations. Specific design standards (e.g., ANSI, EC, DIN, or ISO) as well as legislation related to the type of the workstation designed should also be collected and studied in this phase.

### 3.3 Phase 3: Identification of Users' Needs

The needs of the future workstation users are identified during this phase, considering their task demands as well as their specific characteristics. Consequently, task analysis (see Chapter 13) and users' characteristics analysis should be carried out in this phase.

The task analysis aims at identifying mainly:

- Work processes that will take place and the workstation elements implicated in them
- Physical actions that will be carried out, for example, fine manipulations, whole-body movement, and force exertion
- Required information exchange (visual, auditory, kinesthetic, etc.) and the information sources providing them
- Required privacy
- Required proximity with other workstations, equipment, or elements of the proximal working environment

The observation and analysis of existing work situations with similar workstations may provide valuable information about the users' needs. In fact, as Leplat (2006) points out, work activity is a complex process which comprises essential dynamic and temporal aspects and which integrates the effect of multiple constraints and demands. It should be distinguished from behavior that only constitutes its observable facet: Activity includes behavior and its regulating mechanisms (Leplat, 2006). Although work activity can operationally be described from many views using diverse models, its most fundamental characteristic is that it should be studied intrinsically as an original construction by the workers (Daniellou and Rabardel, 2005; Nathanael and Marmaras, 2009). Therefore, users' needs cannot be fully identified by a simple task analysis.

The specific characteristics of the users' population may include their gender, age, particular disabilities, previous experiences and work practices, and cultural or religious obligations (e.g., in certain countries women are obliged to wear particular costumes).

At this phase, data about performance and health problems of persons working in similar work situations should also be collected. Literature related to ergonomics and to occupational safety and health may be used as the main source for the collection of such data [see the websites of the U.S. Occupational Safety and Health Administration (http://www.osha.gov) and the European Organization of Occupational Safety and Health (http://osha.europa.eu)].

Finally, as in the previous phase, the users' needs should be identified not only for normal but also for exceptional situations in which the workstation occupants may be found (e.g., working under stress, electricity blackout, fire).

### 3.4 Phase 4: Setting Specific Design Goals

Considering the outputs of the previous phases, the design team can now transform the generic ergonomics requirements of workstation design into a set of specific goals. These specific design goals will guide the choices and the decisions to be made in the next phase. Furthermore, they will be used as criteria for assessing the designed prototype and will guide its improvement.

The specific goals are an aggregation of *shoulds* and consist of:

- Requirements of the stakeholders (e.g., the workstation should be convenient for the 95% of the user population, should cost a maximum of $X$ dollars, should increase productivity at least 10%)
- Constraints and requirements imposed by the work system in which the designed workstation(s) will be installed (e.g., the workstation should not exceed $X$ centimeters of length and $Y$ centimeters of width, should offer working conditions not exceeding $X$ decibels of noise and $Y$ degrees of wet bulb globe temperature)
- Users' needs (e.g., the workstation should accommodate elderly people, should be appropriate for prolonged computer work, should facilitate cooperation with the neighboring workstations, should permit the alteration of sitting and standing postures)
- Requirements to avoid common health problems associated with similar situations (e.g., the workstation should minimize upper limb musculoskeletal problems)
- Design standards and related legislation (e.g., the workstation should ensure absence of glare or cold draughts)

The systematic record of all the specific design goals is very helpful for the next phases. It is important to note that agreement on these specific goals between the design team, the management, and user representatives is indispensable.

### 3.5   Phase 5: Design of Prototype

This phase is the most demanding of the design process. In fact, the design team has to generate design solutions meeting all the specific design goals identified in the previous phase. Due to the large number of design goals, as well as the fact that some of them may be conflicting, the design team has to make appropriate compromises, considering some goals as more important than others and eventually passing by some of them. As already stated, good knowledge of the task demands and users' needs, as well as the specific users' characteristics, is the only way to set the right priorities and avoid serious mistakes. Furthermore, the use of data related to (i) the size of the body parts (anthropometry, see Chapter 11) and (ii) the ability and limits of their movements (biomechanics, see Chapter 12) of the users' population should be considered in this phase.

A first decision to make is the working posture(s) that will assume the users of the workstation. Table 1 provides some recommendations for this.

Once the working posture has been decided, the design may continue to define the shape, the dimensions, and the arrangement of the various elements of the workstation. To do so, one has to consider the anthropometric and biomechanical characteristics of the users'

**Table 1  Recommendations for Choosing Working Posture**

| Working posture | Task requirements |
|---|---|
| Working person's choice | It is preferable to arrange for both sitting and standing (see Figure 4) |
| Sitting | Where a stable body is needed:<br>• For accurate control, fine manipulation<br>• For light manipulation work (continuous)<br>• For close visual work — with prolonged attention<br>• For limited headroom, low work heights<br>Where foot controls are necessary (unless of infrequent or short duration)<br>Where a large proportion of the working day requires standing |
| Standing | For heavy, bulky loads<br>Where there are frequent moves from the workplace<br>Where there is no knee room under the equipment<br>Where there is limited front–rear space<br>Where there is a large number of controls and displays<br>Where a large proportion of the working day requires sitting |
| Support seat (see Figure 8) | Where there is no room for a normal seat but a support is desirable |

*Source*: (Corlett and Clark, 1995)



Footrest

**Figure 8**   Example of support seat (Helander, 1995).

population as well as the working actions that will be performed. Besides the ergonomics recommendations presented in previous sections, some additional recommendations for the design of the workstation are the following:

- To define the *clearance*, that is, the minimum required free space for placement of the various body parts, one has to consider the largest user (usually the anthropometric dimensions corresponding to the 97.5 percentile). In fact, providing free space for these users, all shorter users will also have enough space to place their body. For example, if the vertical, lateral, and forward clearances below the working desk are designed considering the height of the thigh upper surface for a sitting person, the hip width and the thigh length corresponding to the 97.5 percentile of the users' population (plus 1 or 3 cm for allowance), 97.5% of the users of this desk will be able to easily approach the desk while sitting.

- To position the different elements of the workplace that must be *reached* by the users, consider the smaller user (usually the anthropometric dimensions corresponding to the 2.5 percentile). In fact, if the smaller users easily reach the various workstation elements, that is, without leaning forward or bending sideways, all larger users will also easily reach them.

- Draw the common kinetospheres or comfort zones for the larger and smaller users and include the various elements of the workstation that have to be manipulated (e.g., controls) (Figure 9).

- When necessary, provide the various elements of the workstation with appropriate adjustability in order to fit in the anthropometric characteristics of the users' population. In this case, it is important to ensure the usability of the corresponding controls.

- While envisioning design solutions continuously check to ensure that the workstation elements do not obstruct the users' courses of action (e.g., perception of necessary visual information, manipulation of controls).

It should be stressed that at least some iterations between phases 2, 3, and the present phase of the design process are unavoidable. In fact, it is almost impossible to identify from the start all the constraints and requirements of the work system, the users' characteristics, or the task requirements that intertwine with the elements of the anticipated workstation.

Another issue to deal with in this phase is designing for protection of the working person from possible annoying or hazardous environmental factors. If the workstation has to be installed in a harsh environment (noisy, cold, or warm, in a hazardous atmosphere, etc.), one has to provide appropriate protection. Again, attention should be paid to the design of such protective elements. These should take into consideration the anthropometric characteristics of the users' population and the task demands in order not to obstruct the processes involved in both normal and degraded operation (e.g., maintenance, breakdowns).

Other important issues that have to be resolved in this phase are the workstation maintainability, its unrestricted evacuation, its stability, and robustness as well as other safety issues such as rough corners.

The search for already existing design ideas and solutions is quite useful. However, they should be carefully examined before their adoption. In fact, such design ideas, although valuable for anticipation, may not be readily applicable for the specific users' population, the specific task demands, or the environment in which the workplace will be installed. Furthermore, many existing design solutions may disregard important ergonomics issues. Finally, although the adoption of already existing design solutions exploits the design community's experience and saves time, it deprives the design team of generating innovative solutions.

The use of computer-aided design (CAD) software with human models is very helpful in this phase. If such software is not available, appropriate drawings and
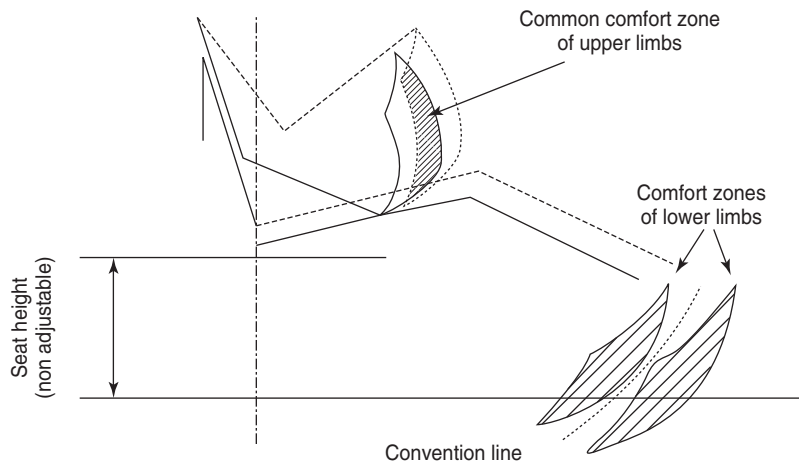


**Figure 9** Drawing the common comfort zones of hands and legs for the large and small users of a driving workplace with nonadjustable chair.

mock-ups should be developed for the generation of design solutions as well as for their assessment (see next phase).

Given the complexity of generating good design solutions, the search for alternatives is useful. The members of the design team should not be anchored at the first design solution that comes to their minds. They should try to generate as many alternative ideas as possible, gradually converging on the ones that better satisfy the design goals.

### 3.6 Phase 6: Assessment of Prototype

Assessment of the designed prototype(s) is required in order to check how well the specific design goals, set in phase 4, have been met, as well as to uncover possible omissions during the identification of the work system constraints and requirements and the users' needs analysis (phases 2 and 3).

The assessment can be performed analytically or/and experimentally, depending on the importance of the project. In the analytical assessment the design team assesses the designed workplace considering exhaustively the specific design goals using the drawings and mock-ups as support. Applying a multi criteria method, the design team may rank the degree to which the design goals have been met. This ranking may be used as a basis for the next phase of the design process (improvement of the prototype) as well as a means to choose among alternative design solutions.

The experimental assessment (or user testing) is performed with the participation of a sample of future users, simulating the work with a full-scale mock-up of the designed workstation prototype(s). The assessment should be made in conditions as close as possible to the real work. Development of use scenarios of both normal and exceptional work situations is useful for this reason. Experimental assessment is indispensable for the identification of problematic aspects that are difficult, if not impossible, to realize before having a real workplace with real users. Furthermore, this type of assessment provides valuable insights for eventual needs during implementation (e.g., the training needed, the eventual need for a users' manual).

### 3.7 Phase 7: Improvements and Final Design

In this phase, the design team proceeds with the required modifications of the designed prototype, considering the outputs of the assessment. The opinions of other specialists such as architects and decorators which have more to do with the aesthetics or production engineers and industrial designers which have more to do with production or materials and robustness matters should be considered in this phase (if such specialists are not already part of the design team).

The final design should be complemented with:

- Drawings for production and appropriate documentation, including the rationale behind the adopted solutions
- Cost estimation for the production of the workstation(s) designed

- Implementation requirements such as the training needed and the users' manual, if required

### 3.8 Final Remark

The reason for conducting the users' needs and requirements analysis is to anticipate the future work situation in order to design a workstation that fits its users, their tasks, and the surrounding environment. However, it is impossible to completely anticipate a future work situation in all its aspects, as work situations are complex, dynamic, and evolving. Furthermore, if the workstation is destined to form part of an already existing work system, it might affect the overall work ecology, something which is also very difficult to anticipate. Therefore, a number of modifications will eventually be needed some time after installation and use. Thus it is strongly suggested to conduct a new assessment of the designed workstation once the users have been familiarized with the new work situation.

## 4 LAYOUT OF WORKSTATIONS

Layout deals with the placement and orientation of individual workstations in a given space (building). The main ergonomics requirements concern the tasks performed, the work organization, and the environmental factors:

- The layout of the workstations should facilitate the work flow.
- The layout of the workstations should facilitate cooperation (of both personnel and external persons, e.g., customers).
- The layout of the workstations should conform to the organizational structure.
- The layout should ensure the required privacy.
- There should be appropriate lighting, conforming to the task's and working person's needs.
- The lighting should be uniform throughout the working person's visual field.
- There should be no annoying reflections or glare in the working area.
- There should be no annoying hot or cold draughts in the workplace.
- Access to the workstations should be unobstructed and safe.

In this section we will focus on the layout of workplaces for office work for the following reasons: First, office layout is an exemplar case for the arrangement of a number of individual workstations in a given space, encompassing all the main ergonomics requirements found in most types of workplaces (with the exception of workplaces where the technology involved determines to a large extent the layout, e.g., workstations in front of machinery). Second, office workplaces concern a growing percentage of the working population worldwide. For example, during the twentieth century the percentage of office workers increased from 17% to over 50% of the workforce in the United States, the rest

working in agriculture, sales, industrial production, and transportation (Czaja, 1987). With the spread of information technologies, the proportion of office workers is expected to further increase; in fact, Brounen and Eichholtz (2004) and Veitch et al. (2007) estimate that at least 50% of the world's population currently works in some form of office. Third, there are still a significant number of office workers which suffer from musculoskeletal disorders or other work-related problems (Corlett, 2006; Griffiths et al., 2007; Luttmann et al., 2010). Finally, current health problems encountered by office workers are to a great extent related to the inappropriate layout of their workplaces (Marmaras and Papadopoulos, 2002).

## 4.1 Generic Types of Office Layouts

There is a number of generic types of office layouts (Shoshkes, 1976; Zelinsky, 1998). The two extremes are the "private office," where each worker has his or her personal closed space/room, and the "open plan," where all the workstations are placed in a common space. In between are a multitude of combinations of private offices with open plans. Workstation arrangements in open plans can be either orthogonal, with single, double, or fourfold desks forming parallel rows, or with the workstations arranged in groups, matching the organizational or functional structure of the work. A recent layout philosophy is the "flexible office," where the furniture and the equipment are designed to be easily movable in order to be able to modify the workstation arrangement depending on the number of people present at the office as well as the number of running projects or work schemes (Brunnberg, 2000). Finally, in order to respond to the current needs for flexibility in the organization and structure of enterprises as well as reduce costs, a new trend in office management is the "free address office" or "nonterritorial office," where workers do not have their own workstation but use the workstation they find free whenever at the office.

Each type of layout has its strengths and weaknesses. Private offices offer increased privacy and better control of environmental conditions, fitting to the particular preferences and needs of their users. However, they are more expensive both in construction and maintenance, not easily modifiable to match changing organizational needs, and render cooperation and supervision difficult. Open-plan offices offer flexibility in changing organizational needs and facilitate cooperation between co-workers but tend to suffer from environmental annoyances such as noise and suboptimum climatic conditions as well as lack of privacy [see De Croon et al. (2005) for a review]. To minimize the noise level and to create some sense of privacy in the open plans, movable barriers may be used. To be effective, the barriers have to be at least 1.5 m high and 2.5 m wide. Furthermore, Wichman (1984) proposes the following specific design recommendations to enhance the working conditions in an open-plan office:

- Use sound-absorbing materials on all major surfaces wherever possible. Noise is often more of a problem than expected.

- Equip the workstations with technological devices of low noise (printers, photocopy machines, telephones, etc.). For example, provide telephones that flash a light for the first two "rings" before emitting an auditory signal.

- Leave some elements of design for the workstation user. People need to have control over their environments; leave some opportunities for changing or rearranging things.

- Provide both vertical and horizontal surfaces for the display of personal belongings. People like to personalize their workstations.

- Provide several easily accessible islands of privacy. This would include small rooms with full walls and doors that can be used for conferences and private or long-distance telephone calls.

- Provide all private work areas with a way to signal willingness of the occupant to be disturbed.

- Have clearly marked flow paths for visitors. For example, hang signs from the ceiling showing where secretaries and department boundaries are located.

- Design workstations so it is easy for drop-in visitors to sit down while speaking. This will tend to reduce disturbances to other workers.

- Plan for ventilation air flow. Most traditional offices have ventilation ducting. This is usually not the case with open-plan cubicles, so they become dead-air cul-de-sacs that are extremely resistant to post hoc resolution.

- Overplan for storage space. Open-plan systems with their emphasis on tidiness seem to chronically underestimate the storage needs of people.

The decision about the generic type of layout should be taken by the stakeholders. The role of the ergonomist here is to indicate the strengths and weaknesses of each alternative in order to facilitate the adoption of the most appropriate type of layout for the specific situation. After this decision has been made, the design team should proceed to the detailed layout of the workstations. The next section describes a systematic method for this purpose.

## 4.2 Systematic Method for Office Layout

This method proposes a systematic way to design workplaces for office work. The method aims at alleviating the design process for arranging the workstations by decomposing the whole problem to a number of stages during which only a limited number of ergonomic requirements are considered. Another characteristic of the method is that the ergonomics requirements to be considered have been converted to design guidelines (Margaritis and Marmaras, 2003). Figure 10 presents the main stages of the method.

Before starting the layout design, the design team should collect data concerning the activities that will be performed in the workplace and the needs of the

Inputs
(data and requirements)                                    Ergonomic requirements

- Walls
- Columns
- Doors                    Determination of              - Easy access to:
- Windows                  space available                 Doors
- Radiators                                                 Windows
- ....                                                      Radiators
                                                            ...

- Desk
- Chairs                   Design of workstation         - Fitting the workstation
- Computer                 modules                         to its users and tasks
- Lockers                                                - Easy access to
- ....                                                     workstations

- Cooperation
- Available surface                                      - Facilitation of cooperation
- Required surface                                         (internal and external)
- Organizational units     Placement of                 - Ensure working conditions
- Already existing         organizational units           meeting task requirements
- closed spaces                                          - Utilization of already existing
- ...                                                      closed spaces

- Organizational units                                   - Facilitation of cooperation
- Cooperation                                              (internal and external)
- Lighting needs                                         - Ensure working conditions
- Privacy needs            Placement of the               meeting task requirements
- External visitors        workstation modules           - Easy access to
- Structural elements                                      the workstations
- ....                                                   - Ensure required privacy
                                                         - ...

- Cooperation                                            - Facilitation of cooperation
- Doors                                                    (internal and external)
- Windows                  Orientation of               - Ensure appropriate
- Supervision              workstation modules            lighting conditions
- ....                                                   - Easy access to
                                                           workstations
                                                         - Facilitation of supervision
                                                         - ...

**Figure 10**   Main stages of a method for office layout meeting the ergonomic requirements.

workers. More specifically, the following information should be gathered:

- The number of people that will work permanently or occasionally.
- The organizational structure and the organizational units it comprises.
- The activities carried out by each organizational unit. Of particular interest are the needs

for cooperation between the different units (and consequently the desired relative proximity between them), the need for reception of external visitors (and consequently the need to provide easy access to them), and any other need related to the particularities of the unit (e.g., security requirements).

- The tasks carried out by each worker. Of particular interest are the needs for cooperation with other workers, the privacy needs, the

reception of external visitors, and the specific needs for lighting.

- The equipment required for each task (e.g., computer, printer, storage).

At this stage the design team should also get the detailed ground plan drawings of the space concerned, including all elements which should be considered as fixed (e.g., structural walls, heating systems).

### 4.2.1 Stage 1: Determination of Available Space

The aim of this stage is to determine the space where no furniture should be placed in order to ensure free passage by the doors and to allow the necessary room for elements such as windows and radiators for manipulation and maintenance purposes.

To determine the free-of-furniture spaces the following suggestions can be used (Figure 11). Allow for

- An area of 50 cm in front of any window
- An area of 3 m in front and 1 m on both sides of the main entrance door
- An area of 1.50 m in front and 50 cm on both sides of any other door
- An area of 50 cm around any radiator

### 4.2.2 Stage 2: Design of Workstation Modules

The aim of this stage is to design workstation modules that meet the needs of the workers. Each module is composed of the appropriate elements for the working activities, that is, desk, seat, storage cabinets, visitors' seats, and any other equipment required for the work. A free space should be provided around the furniture for passages between the workstations as well as for unobstructed sitting and getting up from the seat. This free space may be delimited in the following way (minimum areas).

Allow for:

- An area of 55 cm along the front side of the desk or the outer edge of the visitor's seat
- An area of 50 cm along the entry side of the workstation
- An area of 75 cm along the back side of the desk (seat side)
- An area of 100 cm along the back side of the desk if there are storage cabinets behind the desk

A number of different modules will result from this stage, depending on the particular work requirements (e.g., secretarial module, head of unit module, client service module) (Figure 12).



**Figure 12** Examples of workstation modules.



**Figure 11** Determining the available space.

Laying out workstation modules instead of individual elements such as desks, seats, and so on, permits the designer to focus on the requirements related to the overall layout of the workplace, at the same time ensuring compliance with the requirements related to the individual workstations.

### 4.2.3 Stage 3: Placement of Organizational Units

The aim of this stage is to decide the placement of the different organizational units (i.e., departments, working teams, etc.) within the various free spaces of the building. There are five main issues to be considered here: (i) the shape of each space, (ii) the exploitable area of each space, that is, the area where workstations can be placed, (iii) the required area for each unit, (iv) the desired proximity between the different units, and (v) eventual particular requirements of each unit which may determine their absolute placement within the building (e.g., the reception should be placed right next to the main entrance).

The exploitable area of each space is an approximation of the "free-of-furniture spaces" defined in the first stage, considering also narrow shapes where modules cannot fit. Specifically, this area can be calculated as follows:

$$A_{\text{exploitable}} = A_{\text{total}} - A_{\text{where no modules can be placed}}$$

where

$$A_{\text{total}} = \text{total area of each space}$$
$$A_{\text{where no modules can be placed}} = \text{nonexploitable area, where workstation modules should not or cannot be placed}$$

The required area for each organizational unit can be estimated considering the number of workstation modules needed and the area required for each module. Specifically, in order to estimate the required area for each organizational unit, $A_{\text{required}}$, one has to calculate the sum of the areas of the different workstation modules of the unit.

Comparing the exploitable area of the different spaces with the required area for each unit, the candidate spaces for placing the different units can be defined. Specifically, the candidate spaces for the placement of a particular unit are the spaces where

$$A_{\text{exploitable}} \geq A_{\text{required}}$$

Once the candidate spaces for each unit have been defined, the final decisions about the placement of organizational units can be made. This is done in two steps. In the first step the designer designates spaces for eventual units which present particular placement requirements (e.g., reception). In the second step he or she positions the remaining units considering their desirable relative proximity plus additional criteria such as the need for natural lighting or the reception of external visitors. To facilitate the placement of the organizational units according to their proximity requirements, a *proximity table* as well as *proximity diagrams* may be used.

The *proximity table* represents the desired proximity of each unit with any other one, rated by using the following scale:

**9**: The two units cooperate firmly and should be placed close together.

**3**: The two units cooperate from time to time, and it would be desirable to place them in proximity.

**1**: The two units do not cooperate frequently, and it is indifferent if they will be placed in proximity.

Figure 13 presents the proximity table of a hypothetical firm consisting of nine organizational units. At the right bottom of the table, the total proximity rate (TPR) has been calculated for each unit as the sum of its individual proximity rates. The TPR is an indication of the cooperation needs of each unit with all the others. Consequently, the designer should try to place the units with high TPRs at a central position.

*Proximity diagrams* are a graphical method for the relative placement of organizational units. They facilitate the heuristic search for configurations which minimize the distance between units with close cooperation. Proximity diagrams are drawn on a sheet of paper with equidistant points, like the one shown at Figure 14. The different units are alternated at the different points, trying to find arrangements where the units with close cooperation will be as close as possible to each other.



**Figure 13**  Proximity table of a hypothetical firm.

**Figure 14**   Example of a proximity diagram.

The following rules may be applied to obtain a first configuration:

- Place the unit with the highest TPR at the central point.
- If more than one unit has the same TPR, place the unit with the closest proximity rates (9's) first.
- Continue placing the units having the higher proximity rates with the ones that have already been positioned.
- If more than one unit has proximity rates equal to the one already positioned, place the unit with the higher TPR first.
- Continue in the same manner until all the units have been positioned.

More than one alternative arrangement may be obtained in this way. It should be noted that the proximity diagrams are drawn without taking into account the required area for each unit and the exploitable area of the spaces where the units may be placed. Consequently, the arrangements drawn cannot be directly transposed to the ground plan of the building without modifications. Drawing the proximity diagrams is a means of facilitating the decision concerning the relative positions between organizational units. This method becomes useful when the number of units is high.

### 4.2.4   Stage 4: Placement of Workstation Modules

Considering the outputs of the previous stage, placement of the workstation modules of each unit can start. The following guidelines provide help in meeting the ergonomic requirements:

1.  Place the workstations in a way that facilitates cooperation between co-workers. In other words, workers who cooperate tightly should be placed near each other.
2.  Place the workstations which receive external visitors near the entrance doors.
3.  Place as many workstations as possible near the windows. Windows may provide benefits besides variety in lighting and a view (Hall, 1966). They permit fine adjustment of light through curtains or venetian blinds and provide distant points of visual focus, which can relieve eye fatigue. Furthermore, related research has found that people strongly prefer the workstations placed near windows (Manning, 1965; Sanders and McCormick, 1992).
4.  Avoid placing the working persons in airstreams created by air conditioners, open windows, and doors.
5.  Place the workstation modules in a way that forms straight corridors leading to the doors. Corridors widths for one-person passage should be at least 60 cm and for two-person passage at least 120 cm (Alder, 1999).
6.  Leave the required space in front and to the sides of electric switches and wall plugs.
7.  Leave the required space for waiting visitors. In cases where waiting queues are expected, provide at least a free space of 120 cm width and $n \times 45$ cm length, where $n$ is the maximum expected number of waiting people. Add to this length another 50 cm in front of the queue.

### 4.2.5   Stage 5: Orientation of Workstation Modules

The aim of this stage is to define the direction of the workstation modules of each unit to meet the
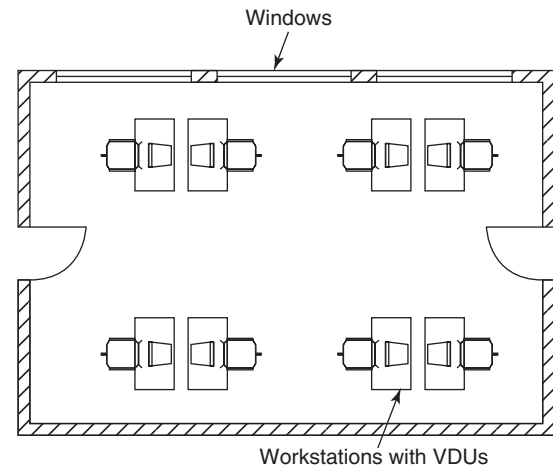


**Figure 15**   Workstations with VDT ideally should be placed at right angles to the windows.
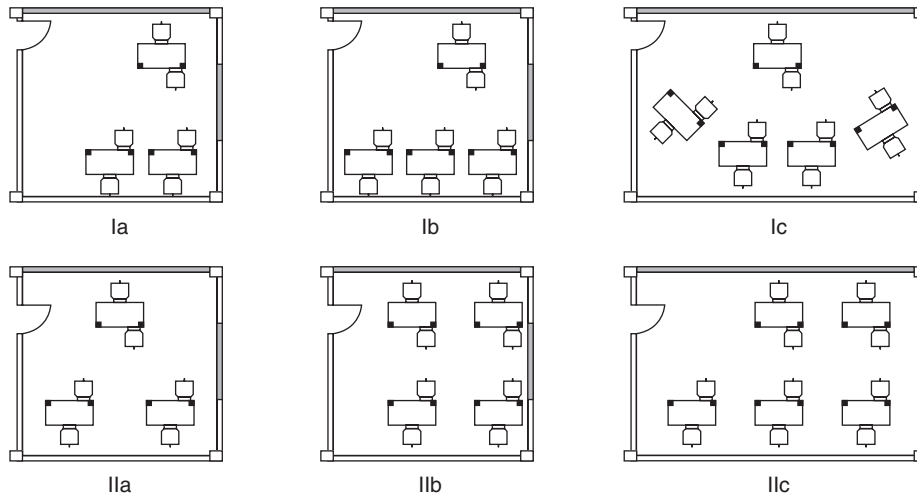
**Figure 16**  Alternative orientations of workstations, depending on the number of team members and the presence or not of a leader. Ia, Ib, and Ic: arrangements with leader; IIa, IIb, and IIc: arrangements without leader.

ergonomics requirements. This stage can be carried out either concurrently with or after the previous stage. The following guidelines may be applied, making appropriate trade-offs if all of them cannot be satisfied:

1. Orient the workstations in such a way that there are no windows directly in front or behind the workers when they are looking toward a visual display terminal (VDT). In offices, windows play a role similar to lights: A window right in front of a worker disturbs through direct glare while directly behind produces reflected glare. For this reason VDT workstations ideally should be placed at right angles to the windows (Grandjean, 1987). (Figure 15).

2. Orient the workstations in such a way that there are no direct lighting sources within ±40° in the vertical and horizontal directions from the line of sight in order to avoid direct glare (Kroemer et al., 1994).

3. Orient the workstations in a way that allows workers to observe entrance doors.

4. Orient the workstations so as to facilitate cooperation between members of work teams. Figure 16 shows alternative orientations of workstations, depending on the number of team members and the presence or not of a leader (Cummings et al., 1974).

## 4.3  Concluding Remarks

Given the complexity of workplace layout design, the design team, trying to apply the various ergonomics guidelines in the different phases, will almost definitely encounter contradictions. To resolve them the design team should be able to focus on the ones considered more important for the case at hand and pay less attention or eventually neglect others. Good knowledge of the generic human abilities and limitations, the

specific characteristics of the people who will work in the designed workplace, and the specificities of the work which will be carried out by them is a prerequisite for successful decisions. Furthermore, the members of the design team should have open and innovative minds and try as many solutions as possible. A systematic assessment of these alternative solutions is advisable to decide on the most satisfactory solution. The participation of the different stakeholders in this process is strongly recommended.

The use of specialized CAD tools may prove very helpful for the application of the presented method, greatly facilitating the generation and assessment of alternative design solutions.

## REFERENCES

Alder, D. (1999), *Metric Handbook Planning and Design Data*, 2nd ed., Architectural Press, New York.

Andersson, G., Murphy, R., Ortengren, R. and Nachemson, A. (1979), "The Influence of Backrest Inclination and Lumbar Support on the Lumbar Lordosis in Sitting", *Spine*, Vol. 4, pp. 52–58.

Ankrum, D. R. (1997), "Integrating Neck Posture and Vision at VDT Workstations", in *Proceedings of the Fifth International Scientific Conference on Work with Display Units*, H. Miyamoto, S. Saito, M. Kajiyama and N. Koizumi, Eds., Waseda University, Tokyo, pp. 63–64.

Ankrum, D. R., and Nemeth, K. J. (2000), \$"\$Head And Neck Posture at Computer Workstations_What's Neutral?" *Proceedings of the 14th Triennial Congress of the International Ergonomics Association*, Vol. 5, pp. 565–568.

Bendix, T. (1986), "Seated Trunk Posture at Various Seat Inclinations, Seats Heights, and Table Heights", *Human Factors*, Vol. 26, pp. 695–703.

Bridger, R. (1988), "Postural Adaptations to a Sloping Chair and Work-Surface", *Human Factors*, Vol. 30, pp. 237–247.

Brounen, D., and Eichholtz, P. (2004), "Demographics and the Global Office Market—Consequences for Property

Portfolios", *Journal of Real Estate Portfolio Management*, Vol. 10, No. 3, pp. 231–242.

Brunnberg, H. (2000), "Evaluation of Flexible Offices", in *Proceedings of the IEA 2000/HFES 2000 Congress*, Vol. 1, HFES, San Diego, pp. 667–670.

Corlett, E. N. (2006), "Background to Sitting at Work: Research-Based Requirements for the Design of Work Seats", *Ergonomics*, Vol. 49, No. 14, pp. 1538–1546.

Corlett, E. N. (2009), "Ergonomics and Sitting at Work", *Work*, Vol. 34, No. 2, pp. 235–238.

Corlett, E. N., and Clark, T. S. (1995), *The Ergonomics of Workspaces and Machines*, Taylor & Francis, London.

Cummings, L., Huber, G. P., and Arendt, E. (1974), "Effects of Size and Spatial Arrangements on Group Decision-Making", *Academy of Management Journal*, Vol. 17, No. 3, pp. 460–475.

Czaja, S. J. (1987), "Human Factors in Office Automation", in *Handbook of Human Factors*, G. Salvendy, Ed., Wiley, New York.

Dainoff, M. (1994), "Three Myths of Ergonomic Seating", in *Hard Facts about Soft Machines: The Ergonomics of Seating*, R. Lueder and K. Noro, Eds., Taylor and Francis, London.

Daniellou, F., and Rabardel, P. (2005), "Activity-Oriented Approaches to Ergonomics: Some Traditions and Communities", *Theoretical Issues in Ergonomics Science*, Vol. 6, No. 5, pp. 353–357.

De Croon, E., Sluiter, J., Kuijer, P. P. and Frings-Dresen. M. (2005), "The Effect of Office Concepts on Worker Health and Performance: A Systematic Review of the Literature", *Ergonomics*, Vol. 48, No. 2, pp. 119–134.

Grandjean, E. (1987), *Ergonomics in Computerized Offices*, Taylor and Francis, London.

Grieco, A. (1986), "Sitting posture: an old problem and a new one", *Ergonomics*, 29, pp. 345–362.

Grey, F. E., Hanson, J. A., and Jones, F. P. (1966), "Postural Aspects of Neck Muscle Tension", *Ergonomics*, Vol. 9, No. 3, pp. 245–256.

Griffiths, K. L., Mackey, M. G., and Adamson, B. J. (2007), "The Impact of a Computerized Work Environment on Professional Occupational Groups and Behavioural and Physiological Risk Factors for Musculoskeletal Symptoms: A Literature Review", *Journal of Occupational Rehabilitation*, Vol. 17, No. 4, pp. 743–765.

Groenesteijna, L., Vinka, P., de Loozea, M. and Krause F. (2009), "Effects of Differences in Office Chair Controls, Seat and Backrest Angle Design in Relation to Tasks", *Applied Ergonomics*, Vol. 40, No. 3, pp. 362–370.

Hall, E. T. (1966), *The Hidden Dimension*, Doubleday, New York.

Helander, M. (1995), *A Guide to the Ergonomics of Manufacturing*, Taylor and Francis, London.

Jampel, R. S. and Shi, D. X. (1992), "The Primary Position of the Eyes, the Resetting Saccade, and the Transverse Visual Head Plane", *Investigative Ophthalmology and Visual Science*, Vol. 33, pp. 2501–2510.

Kroemer, K., and Grandjean, E. (1997), *Fitting the Task to the Human: A Textbook of Occupational Ergonomics*, 5th ed., Taylor & Francis, London.

Kroemer, K., Kroemer, H., and Kroemer-Elbert, K. (1994), *How to Design for Ease and Efficiency*, Prentice-Hall, Englewood Cliffs, NJ.

Leplat, J. (2006), "Activity," in *International Encyclopedia of Ergonomics and Human Factors*, W. Karwowski, Ed., Taylor & Francis, London, pp. 567–571.

Lueder, R. (1986), "Work Station Design", in *The Ergonomics Payoff: Designing the Electronic Office*, R. Lueder, Ed., Holt, Rinehart & Winston, Toronto.

Lueder, R., and Noro, K. (1994), *Hard Facts about Soft Machines: The Ergonomics of Seating*, Taylor & Francis, London.

Luttmann, A., Schmidt, K. H., and Jägera M. (2010), "Working Conditions, Muscular Activity and Complaints of Office Workers", *International Journal of Industrial Ergonomics*, Vol. 40, No. 5, pp. 549–559.

Mandal, A. (1985), *The Seated Man*, Dafnia Publications, Denmark.

Manning, P. (1965), *Office Design: A Study of Environment by the Pilkington Research Unit*, University of Liverpool Press, Liverpool.

Margaritis, S., and Marmaras, N. (2003), "Making the Ergonomic Requirements Functional: The Case of Computerized Office Layout", in *Proceedings of the XVth Triennial Congress of the International Ergonomics Association and The 7th Conference of Ergonomics Society of Korea /Japan Ergonomics Society*, Vol. 2, The Ergonomics Society of Korea, Seoul.

Marmaras, N. and Papadopoulos, St. (2002), A study of computerized offices in Greece: Are ergonomic design requirements met? *International Journal of Human-Computer Interaction*, 16 (2), pp. 261–281.

Marmaras, N., Nathanael, D. and Zarboutis, N. (2008), "The Transition from CRT To LCD Monitors: Effects on Monitor Placement and Possible Consequences in Viewing Distance and Body Postures", *International Journal of Industrial Ergonomics*, Vol. 38, Nos. 7–8, pp. 584–592.

Marras, W. S. (2005), "The Future of Research in Understanding and Controlling Work-Related Low Back Disorders", *Ergonomics*, Vol. 48, pp. 464–477.

Nachemson, A., and Elfstrom, G. (1970), "Intravital Dynamic Pressure Measurements in Lumbar Disks", *Scandinavian Journal of Rehabilitation Medicine*, Suppl., 1, pp. 1–40.

Nathanael, D. and Marmaras, N. (2009), "The Redesign of a Tram Drivers Workstation: An Activity Approach", in *Proceedings of the XVIIth World Congress of the IEA*. Sauter, S., Dainoff, M. & Smith, M., Eds., IEA, Peking, China.

Sanders, S. M., and McCormick, J. E. (1992), *Human Factors in Engineering and Design*, 7th ed., McGraw-Hill, New York.

Shoshkes, L. (1976), *Space Planning: Designing the Office Environment*, Architectural Record Books, New York.

Veitch, J. A., Charles, K. E., Farley, K.M. J., and Newsham, G. R. (2007), "A Model of Satisfaction with Open-Plan Office Conditions: COPE Field Findings", *Journal of Environmental Psychology*, Vol. 27, No. 3, pp. 177–189.

Vitalis, A., Marmaras, N., Legg, St., and Poulakakis, G. (2000), "Please Be Seated", in *Proceedings of the 14th Triennial Congress of the International Ergonomics Association*, HFES, San Diego, pp. 43–45.

Wichman, H. (1984), "Shifting from Traditional to Open Offices: Problems and Suggested Design Principles", in *Human Factors in Organizational Design and Management*, H. Hendrick and O. Brown, Jr., Eds., Elsevier, Amsterdam.

Zelinsky, M. (1998), *New Workplaces for New Work Styles*, McGraw-Hill, New York.

# CHAPTER 22

# VIBRATION AND MOTION

**Michael J. Griffin**
**Human Factors Research Unit, Institute of Sound and Vibration Research**
**University of Southampton**
**Southampton, England**

## 1 INTRODUCTION

In work and leisure activities the human body experiences movement. The motion may be voluntary (as in some sports) or involuntary (as for passengers in vehicles). Movements may occur simultaneously in six different directions: three translational directions (fore-and-aft, lateral, and vertical) and three rotational directions (roll, pitch, and yaw). Translational movements at constant velocity (i.e., with no change of speed or direction) are mostly imperceptible, except where exteroceptors (e.g., the eyes or ears) detect a change of position relative to other objects. Translational motion can also be detected when the velocity changes, causing acceleration or deceleration of the body that can be perceived via interoceptors (e.g., the vestibular, cutaneous, kinesthetic, or visceral sensory systems). Rotation of the body at constant velocity may be detected because it gives rise to translational acceleration in the body, because it re-orientates the body relative to the gravitational force of Earth, or because the changing orientation relative to other objects is perceptible through exteroceptors. Vibration is oscillatory motion: the velocity is changing and so the movement is detectable by interoceptors and exteroceptors.

Vibration of the body may be desirable or undesirable. It can be described as pleasant or unpleasant, it can interfere with the performance of various tasks and cause injury and disease. Low-frequency oscillations of the body and movements of visual displays can cause motion sickness. It is convenient to consider human exposure to oscillatory motion in three categories:

1. *Whole-body vibration* occurs when the body is supported on a surface that is vibrating (e.g., sitting on a seat that vibrates, standing on a vibrating floor or lying on a vibrating surface). Whole-body vibration occurs in transport (e.g., road, off-road, rail, air and marine transport) and when near some machinery.

2. *Motion sickness* can occur when real or illusory movements of the body or the environment lead to ambiguous inferences as to the movement or orientation of the human body. The movements associated with motion sickness are always of very low frequency, usually below 1 Hz.

3. *Hand-transmitted vibration* is caused by various processes in industry, agriculture, mining, construction, and transport where vibrating tools or workpieces are grasped or pushed by the hands or fingers.

There are many different effects of oscillatory motion on the body and many variables influencing each effect. The variables may be categorized as extrinsic variables (those occurring outside the human body) and intrinsic variables (the variability that occurs between

**Table 1 Variables Influencing Human Responses to Oscillatory Motion**

| Extrinsic variables | Intrinsic variables |
|---|---|
| Vibration variables: | Intrasubject variability |
|   Vibration magnitude |   Body posture |
|   Vibration frequency |   Body position |
|   Vibration direction |   Body orientation (sitting, |
|   Vibration input positions |     standing, recumbent) |
|   Vibration duration | Intersubject variability |
| Other variables: |   Body size and weight |
|   Other stressors (noise, |   Body dynamic response |
|     temperature, etc.) |   Age |
|   Seat dynamics |   Gender |
| |   Experience, expectation, |
| |     attitude, and personality |
| |   Fitness |



**Figure 1** Displacement, velocity, and acceleration waveforms for a sinusoidal vibration. If the vibration has frequency $f$ (in hertz) and peak displacement $D$ (in meters), the peak velocity is $V = 2\pi f D$ (in meters per second) and the peak acceleration is $A = (2\pi f)^2 D$ (in meters per second per second).

and within people), as in Table 1. Some variables, especially intersubject variability, have large effects but are not easily measured. Consequently, it is often not practicable to make highly accurate predictions of the discomfort, interference with activities, or health effects for an individual. However, the average effect, or the probability of an effect, can be predicted for groups of people.

This chapter introduces human responses to oscillatory motion, summarizes current methods of evaluating exposures to oscillatory motion, and identifies some methods of minimizing unwanted effects of vibration.

## 2 MEASUREMENT OF VIBRATION AND MOTION

### 2.1 Vibration Magnitude

When vibrating, an object has alternately a velocity in one direction and then a velocity in the opposite direction. This change in velocity means that the object is constantly accelerating, first in one direction and then in the opposite direction. Figure 1 shows the displacement waveform, the velocity waveform, and acceleration waveform for a movement occurring at a single frequency (i.e., a sinusoidal oscillation). The magnitude of a vibration can be quantified by its displacement, its velocity, or its acceleration. For practical convenience, the magnitude of vibration is now usually expressed in terms of the acceleration and measured using accelerometers. The units of acceleration are meters per second per second (i.e., m s$^{-2}$, or m/s$^2$). The acceleration due to gravity on Earth is approximately 9.81 m s$^{-2}$.

The magnitude of an oscillation can be expressed as the difference between the maximum and minimum values of the motion (e.g., the peak-to-peak acceleration) or the maximum deviation from some central point (e.g., the peak acceleration). Most often, magnitudes of vibration are expressed in terms of an average measure of the oscillatory motion, usually the root-mean-square (r.m.s.) value of the acceleration (i.e., m s$^{-2}$ r.m.s. for
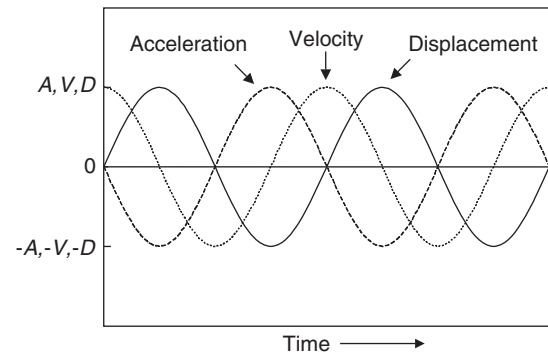
translational acceleration, rad s$^{-2}$ r.m.s. for rotational acceleration). For a sinusoidal motion, the r.m.s. value is the peak value divided by $\sqrt{2}$ (i.e., the peak value divided by approximately 1.4).

When observing vibration, it is sometimes possible to estimate the displacement caused by the motion. For a sinusoidal motion, the acceleration $a$ can be calculated from the frequency $f$ in hertz and the displacement $d$:

$$a = (2\pi f)^2 d$$

For example, a sinusoidal motion with a frequency of 1 Hz and a peak-to-peak displacement of 0.1 m will have an acceleration of 3.95 m s$^{-2}$ peak to peak, 1.97 m s$^{-2}$ peak, and 1.40 m s$^{-2}$ r.m.s. Although this expression can be used to convert acceleration measurements to corresponding displacements, it is only accurate when the motion occurs at a single frequency (i.e., it has a sinusoidal waveform as shown in Figure 1).

Logarithmic scales for quantifying vibration magnitudes in decibels are sometimes used. When using the reference level in International Standard 1683 [International Organization for Standardization (ISO), 2008], the acceleration level $L_a$ is expressed by $L_a = 20 \log_{10}(a/a_0)$, where $a$ is the measured acceleration (in m s$^{-2}$) and $a_0$ is the reference level of $10^{-6}$ m s$^{-2}$. With this reference, an acceleration of 1 m s$^{-2}$ corresponds to 120 dB, and an acceleration of 10 m s$^{-2}$ corresponds to 140 dB.

### 2.2 Vibration Frequency

The frequency of vibration is expressed in cycles per second using the SI unit hertz (Hz). The frequency of vibration influences the extent to which vibration is transmitted to the surface of the body (e.g., through seating), the extent to which it is transmitted through the body (e.g., from seat to head), and the responses to vibration within the body. From Section 2.1 it will be seen that the relation between the displacement and the

acceleration of a motion depends on the frequency of oscillation: A displacement of 1 mm corresponds to a low acceleration at low frequencies (e.g., 0.039 m s$^{-2}$ at 1 Hz) but a very high acceleration at high frequencies (e.g., 394 m s$^{-2}$ at 100 Hz).

### 2.3   Vibration Direction

The responses of the body to motion differ according to the direction of the motion. Vibration is often measured at the interfaces between the body and the vibrating surfaces in three orthogonal directions. Figure 2 shows a coordinate system used when measuring vibration of a hand holding a tool.

The three principal directions of whole-body vibration for seated and standing persons are $x$ axis (fore-and-aft), $y$ axis (lateral), and $z$ axis (vertical). The vibration is measured at the interface between the body and the surface supporting the body (e.g., on the seat beneath the ischial tuberosities for a seated person, beneath the feet for a standing person). Figure 3 illustrates the translational and rotational axes for an origin at the ischial tuberosities on a seat and the translational axes at a backrest and the feet of a seated person.
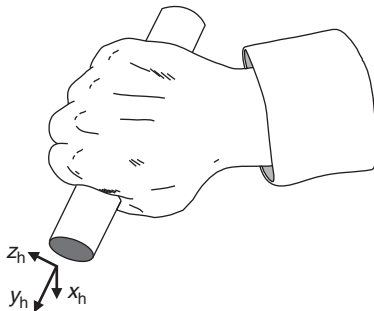


**Figure 2**   Axes of vibration used to measure exposures to hand-transmitted vibration.
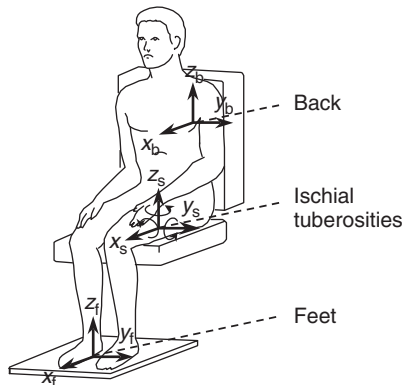


**Figure 3**   Axes of vibration used to measure exposures to whole-body vibration.

### 2.4   Vibration Duration

Some human responses to vibration depend on the duration of exposure. Additionally, the duration of measurement may affect the measured magnitude of the vibration. The root-mean-square (i.e., r.m.s.) acceleration may not provide a good indication of vibration severity if the vibration is intermittent, contains shocks, or otherwise varies in magnitude from time to time (see, e.g., Section 3.3).

## 3   WHOLE-BODY VIBRATION

Whole-body vibration may affect health, comfort, and the performance of activities. The comments of persons exposed to vibration mostly derive from the sensations produced by vibration rather than certain knowledge that the vibration is causing harm or reducing their performance. Vibration of the whole body is produced by various types of industrial machinery and by all forms of transport (including road, off-road, rail, sea, and air transport).

### 3.1   Vibration Discomfort

The relative discomfort caused by different oscillatory motions can be predicted from measurements of the vibration. For very low magnitude motions it is possible to estimate the percentage of persons who will be able to feel vibration and the percentage who will not be able to feel the vibration. For higher vibration magnitudes, an approximate indication of the extent of subjective reactions is available in a semantic scale of discomfort.

Limits appropriate to the prevention of vibration discomfort vary between different environments (e.g., between buildings and transport) and between different types of transport (e.g., between cars and trucks) and within types of vehicle (e.g., between sports cars and limousines). The design limit depends on external factors (e.g., cost and speed) and the comfort in alternative environments (e.g., competitive vehicles).

#### 3.1.1   Effects of Vibration Magnitude

The absolute threshold for the perception of vertical whole-body vibration in the frequency range 1 to 100 Hz is, very approximately, 0.01 m s$^{-2}$ r.m.s.; a magnitude of 0.1 m s$^{-2}$ will be easily noticeable; magnitudes around 1 m s$^{-2}$ r.m.s. are usually considered uncomfortable; magnitudes of 10 m s$^{-2}$ r.m.s. are usually dangerous. The precise values depend on vibration frequency and the exposure duration and they are different for other axes of vibration (Morioka and Griffin, 2006a,b).

A doubling of vibration magnitude (expressed in m s$^{-2}$) produces, very approximately, a doubling of the sensation of discomfort; the precise increase depends on the frequency and direction of vibration. For many motions, a halving of the vibration magnitude therefore greatly reduces discomfort.

#### 3.1.2   Effects of Vibration Frequency and Direction

The dynamic responses of the body and the relevant physiological and psychological processes dictate that

subjective reactions to vibration depend on the frequency and the direction of vibration. The extent to which a given acceleration will cause a greater or lesser effect on the body at different frequencies is reflected in *frequency weightings*: frequencies capable of causing the greatest effect are given the greatest 'weight' and others are attenuated according to their relative importance.

Frequency weightings for human response to vibration have been derived from laboratory experiments in which volunteer subjects have been exposed to a set of motions having different frequencies. The subjects' responses are used to determine *equivalent comfort contours* (Morioka and Griffin, 2006a). The reciprocal of such a curve forms the shape of the frequency weighting. Figure 4 shows frequency weightings $W_b$ to $W_f$ as defined in British standard 6841 [British Standards Institution (BSI), 1987]. International standard 2631 (ISO, 1997) allows the same weightings for evaluating vibration with respect to comfort but requires the use of $W_k$ in place of the almost identical weighting $W_b$ when evaluating vertical vibration at the seat with respect to health effects. Table 2 defines simple asymptotic (i.e., straight-line) approximations to these weightings and Table 3 shows how the weightings should be applied to the 12 axes of vibration illustrated in Figure 3. [The weightings $W_g$ and $W_f$ are not required to predict vibration discomfort: $W_g$ has been used for assessing interference with activities and is similar to the weighting for vertical vibration in an outdated International Standard (ISO 2631, 1974, 1985); $W_f$ is used to predict motion sickness caused by vertical oscillation; see Section 4.]

In order to minimize the number of frequency weightings, some are used for more than one axis of vibration, with different *axis-multiplying factors* allowing for overall differences in sensitivity between

**Table 2 Asymptotic Approximations to Frequency Weightings, $W(f)$, in British Standard 6841 (BSI, 1987) for Comfort, Health, Activities, and Motion Sickness**

| Weighting Name | Weighting Definition | |
|---|---|---|
| $W_b$ | $0.5 < f < 2.0$ | $W(f) = 0.4$ |
| | $2.0 < f < 5.0$ | $W(f) = f/5.0$ |
| | $5.0 < f < 16.0$ | $W(f) = 1.00$ |
| | $16.0 < f < 80.0$ | $W(f) = 16.0/f$ |
| $W_c$ | $0.5 < f < 8.0$ | $W(f) = 1.0$ |
| | $8.0 < f < 80.0$ | $W(f) = 8.0/f$ |
| $W_d$ | $0.5 < f < 2.0$ | $W(f) = 1.00$ |
| | $2.0 < f < 80.0$ | $W(f) = 2.0/f$ |
| $W_e$ | $0.5 < f < 1.0$ | $W(f) = 1.00$ |
| | $1.0 < f < 80.0$ | $W(f) = 1.00/f$ |
| $W_f$ | $0.100 < f < 0.125$ | $W(f) = f/0.125$ |
| | $0.125 < f < 0.250$ | $W(f) = 1.0$ |
| | $0.250 < f < 0.500$ | $W(f) = (0.25/f)^2$ |
| $W_g$ | $1.0 < f < 4.0$ | $W(f) = (f/4)^{1/2}$ |
| | $4.0 < f < 8.0$ | $W(f) = 1.00$ |
| | $8.0 < f < 80.0$ | $W(f) = 8.0/f$ |

*Note:* $f$ = frequency, Hz; $W(f) = 0$ where not defined.

axes (see Table 3). The frequency-weighted acceleration should be multiplied by the axis-multiplying factor before the component is compared with components in other axes or included in any summation over axes. The r.m.s. value of this acceleration (i.e., after frequency weighting and after being multiplied by the axis-multiplying factor) is sometimes called a *component ride value* (Griffin, 1990).
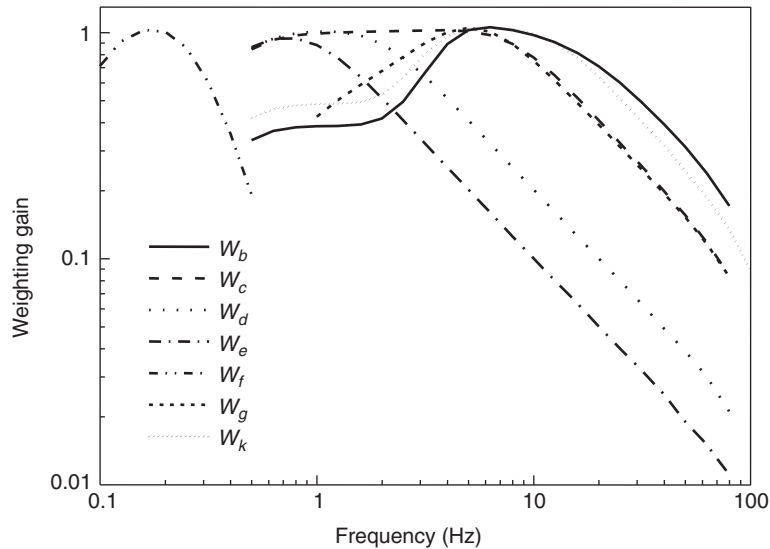


**Figure 4** Acceleration frequency weightings for whole-body vibration and motion sickness as defined in standards BS 6841 (BSI, 1987) and ISO 2631-1 (ISO, 1997).

**Table 3 Application of Frequency Weightings for Evaluation of Vibration with Respect to Discomfort**

| Input Position | Axis | Weighting | Axis-Multiplying Factor |
|---|---|---|---|
| Seat | $x$ | $W_d$ | 1.0 |
| | $y$ | $W_d$ | 1.0 |
| | $z$ | $W_b$ | 1.0 |
| | $r_x$ (roll) | $W_e$ | 0.63 |
| | $r_y$ (pitch) | $W_e$ | 0.40 |
| | $r_z$ (yaw) | $W_e$ | 0.20 |
| Seat back | $x$ | $W_c$ | 0.80 |
| | $y$ | $W_d$ | 0.50 |
| | $z$ | $W_d$ | 0.40 |
| Feet | $x$ | $W_b$ | 0.25 |
| | $y$ | $W_b$ | 0.25 |
| | $z$ | $W_b$ | 0.40 |

*Note:* $f$ = frequency, Hz; $W(f) = 0$ where not defined.

Vibration occurring in several axes is more uncomfortable than vibration occurring in a single axis. To obtain an *overall ride value*, the '*root-sums-of-squares*' of the component ride values is calculated:

$$\text{Overall ride value} = \left[ \sum (\text{component ride values})^2 \right]^{1/2}$$

Overall ride values from different environments can be compared: a vehicle having the highest overall ride value would be expected to be the most uncomfortable with respect to vibration. The overall ride values can also be compared with the discomfort scale shown in Table 4. This scale indicates the approximate range of vibration magnitudes that are significant in relation to the range of vibration discomfort that might be experienced in vehicles.

**Table 4 Scale of Vibration Discomfort from British Standard 6841 (BSI, 1987) and International Standard 2631 (ISO, 1997)**

| | r.m.s Weighted Acceleration ($ms^{-2}$) | |
|---|---|---|
| | 3.15 | |
| Extremely uncomfortable | 2.5 | |
| | 2.0 | Very uncomfortable |
| | 1.6 | |
| | 1.25 | |
| Uncomfortable | 1.0 | |
| | 0.8 | Fairly uncomfortable |
| | 0.63 | |
| | 0.5 | |
| A little uncomfortable | 0.4 | |
| | 0.315 | Not uncomfortable |
| | 0.25 | |

*Source*: BSI (1987a) and ISO (1997).

### 3.1.3 Effects of Vibration Duration

Vibration discomfort tends to increase with increasing duration of exposure to vibration. The rate of increase may depend on many factors, but a simple *fourth-power* time dependency is used to approximate how discomfort varies with duration of exposure from the shortest possible shock to a full day of vibration exposure [i.e., $(\text{acceleration})^4 \times \text{duration} = \text{constant}$; see Section 3.3].

## 3.2 Interference with Activities

Vibration and motion can interfere with the acquisition of information (e.g., by the eyes), the output of information (e.g., by hand or foot movements), or the complex central processes that relate input to output (e.g., learning, memory, decision making). Effects of oscillatory motion on human performance may impair safety.

There is most evidence of whole-body vibration affecting performance for input processes (mainly vision) and output processes (mainly continuous hand control). In both cases there may be a disturbance occurring entirely outside the body (e.g., vibration of a viewed display or vibration of a hand-held control), a disturbance at the input or output (e.g., movement of the eye or hand), and a disturbance within the body affecting the peripheral nervous system (i.e., afferent or efferent nervous system). Central processes may also be affected by vibration, but understanding is currently too limited to make confident generalized statements (see Figure 5).

The effects of vibration on vision and manual control are most usually caused by the movement of the affected part of the body (i.e., eye or hand). The effects may be decreased by reducing the transmission of vibration to the eye or to the hand or by making the task less susceptible to disturbance (e.g., increasing the size of a display or reducing the sensitivity of a control). Often, the effects of vibration on vision and manual control can be much reduced by redesign of the task.

### 3.2.1 Vision

Reading a newspaper in a moving vehicle may be difficult because the paper is moving, the eye is moving, or both the paper and the eye are moving. There are many variables which affect visual performance in these conditions: it is not possible to represent adequately the effects of vibration on vision without considering the effects of these variables.

***Stationary Observer*** When a stationary observer views a moving display, the eye may be able to track the position of the display using *pursuit eye movements*. This closed-loop reflex will give smooth pursuit movements of the eye and clear vision if the display is moving at frequencies less than about 1 Hz and with a low velocity. At slightly higher frequencies of oscillation, the precise value depending on the predictability of the motion waveform, the eye will make *saccadic* eye movements to redirect the eye with small jumps. At frequencies greater than about 3 Hz, the eye will best be directed to one extreme of the oscillation and attempt to view the image as it is temporarily stationary
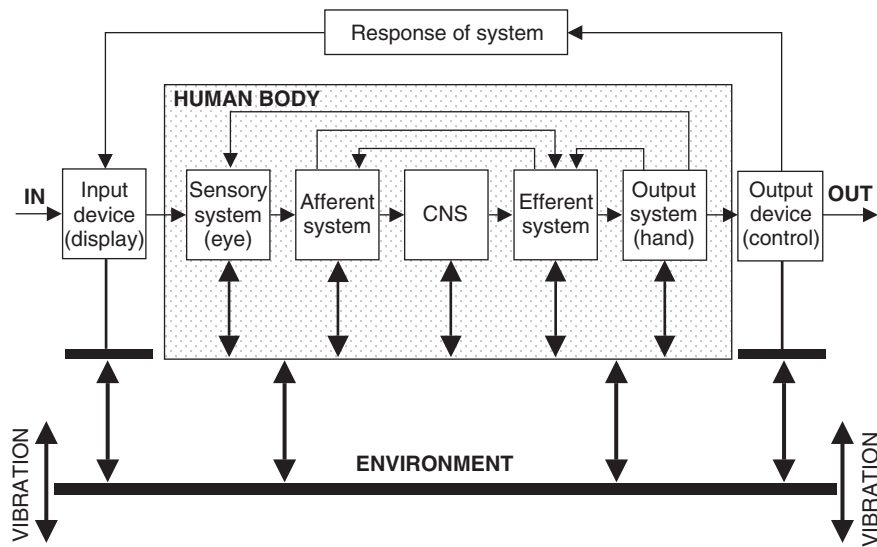
**Figure 5**  Information flow in a simple system and the areas where vibration may affect human activities.

while reversing the direction of movement (i.e., at the 'nodes' of the motion).

In some conditions, the absolute threshold for the visual detection of the vibration of an object occurs when the peak-to-peak oscillatory motion gives an angular displacement at the eye of approximately 1 min of arc. The acceleration required to achieve this threshold is very low at low frequencies but increases in proportion to the square of the frequency to become very high at high frequencies. When the vibration displacement is greater than the visual detection threshold, there will be perceptible blur if the vibration frequency is greater than about 3 Hz. The effects of vibration on visual performance (e.g., effects on reading speed and reading accuracy) may then be estimated from the maximum time that the image spends over some small area of the retina (e.g., the period of time spent near the nodes of the motion with sinusoidal vibration). For sinusoidal vibration this time decreases (and so reading errors increase) in linear proportion to the frequency of vibration and in proportion to the square root of the displacement of vibration (O'Hanlon and Griffin, 1971). With dual-axis vibration (e.g., combined vertical and lateral vibration of a display) this time is greatly reduced and reading performance drops greatly (Meddick and Griffin, 1976). With narrow-band random vibration there is a greater probability of low image velocity than with sinusoidal vibration of the same magnitude and predominant frequency, so reading performance tends to be less affected by random vibration than sinusoidal vibration (Moseley et al., 1982). Display vibration reduces the ability to see fine detail in displays while having little effect on the clarity of larger forms.

***Vibrating Observer***  If an observer is sitting or standing on a vibrating surface, the effects of vibration depend on the extent to which the vibration is transmitted to the eye. The motion of the head is highly dependent on body posture but is likely to occur in translational axes (i.e., in the *x*-, *y*-, and *z*-axes) and in rotational axes (i.e., in the roll, pitch, and yaw axes). Often, the predominant head motions affecting vision are in the vertical and pitch axes of the head. The dynamic response of the body may result in greatest head acceleration in these axes at frequencies around 5 Hz, but vibration at higher and lower frequencies can also have large effects on vision.

The pitch motion of the head is well compensated by the *vestibulo-ocular reflex*, which serves to help stabilize the line of sight of the eyes at frequencies less than about 10 Hz (e.g., Benson and Barnes, 1978). So, although there is often pitch oscillation of the head at 5 Hz, there is less pitch oscillation of the eyes at this frequency. Pitch oscillation of the head, therefore, has a less than expected effect on vision—unless the display is attached to the head, as with a helmet-mounted display (see Wells and Griffin, 1984).

The effects on vision of translational oscillation of the head depend on viewing distance: the effects are greatest when close to a display. As the viewing distance increases, the retinal image motions produced by translational displacements of the head decrease until, when viewing an object at infinite distance, there is no retinal image motion produced by translational head displacement (Griffin, 1976).

For a vibrating observer there may be little difficulty with low-frequency pitch head motions when viewing a fixed display and no difficulty with translational head motions when viewing a distant display. The greatest problems for a vibrating observer occur with pitch head motion when the display is attached to the head and with translational head motion when viewing near displays. Additionally, there may be resonances of the eye within the head, but these are highly variable between individuals and often occur at high frequencies

(e.g., 30 Hz and greater) and it is often possible to attenuate the vibration entering the body at these high frequencies.

***Observer and Display Vibrating*** When an observer and a display oscillate together, in phase, at low frequencies, the retinal image motions (and decrements in visual performance) are less than when either the observer or the display oscillates separately (Moseley and Griffin, 1986b). However, the advantage is lost as the vibration frequency increases because there is increasing phase difference between the oscillation of the head and the oscillation of the display. At frequencies around 5 Hz the phase lags between seat motion and head motion may be $90°$ or more (depending on seating conditions) and sufficient to eliminate any advantage of moving the seat and the display together. Figure 6 shows an example of how the time taken to read information on a screen is affected for the three viewing conditions (display vibration with a stationary observer, vibrating observer with stationary display, both observer and display vibrating) with sinusoidal vibration in the frequency range 0.5 to 5 Hz.

***Other Variables*** Some common situations in which vibration affects vision do not fall into one of the three categories in Figure 6. For example, when reading a newspaper on a train the motion of the arms may result in the motion of the paper being different in magnitude and phase from the motions of both the seat and the head of the observer. The dominant axis of motion of the newspaper may be different from the dominant axis of motion of the person (Griffin and Hayward, 1994).

Increasing the size of detail in a display will often greatly reduce adverse effects of vibration on vision

(Lewis and Griffin, 1979). In one experiment a 75% reduction in reading errors was achieved with only a 25% increase in the size of Landolt C targets (O'Hanlon and Griffin, 1971). Increasing the spacing between rows of letters and choosing appropriate character fonts can also be beneficial. The contrast of the display or other reading material also has an effect, but maximum performance may not occur with maximum contrast. The influence of such factors is summarized in a design guide for visual displays to be used in vibration environments (Moseley and Griffin, 1986a)

Optical devices may increase or decrease the effects of vibration on vision. Simple optical magnification of a vibrating object will increase both the apparent size of the object and the apparent magnitude of the vibration. Sometimes this will be beneficial if the benefits of increasing the size of the detail more than offset the effects of increased magnitude of vibration. The effect is similar to reducing the viewing distance, which can be beneficial for stationary observers viewing vibrating displays. If the observer is vibrating, the use of binoculars (and other magnifying devices) can be detrimental if the vibration of the device (e.g., rotation in the hand holding the binoculars) causes such an increase in the image movement that it is not sufficiently compensated by the increase in image size. The use of binoculars and telescopes in moving vehicles becomes difficult for these reasons.

### 3.2.2 Manual Control

Simple and complex manual control tasks can also be impeded by vibration. Studies of the effects of whole-body vibration on the performance of hand tracking tasks have been reviewed elsewhere (e.g., McLeod and Griffin, 1989). The characteristics of the task and the characteristics of the vibration combine to determine effects of vibration on activities: a given vibration may greatly affect the performance of one task but have little effect on the performance of another task.

***Effects Produced by Vibration*** The most obvious consequence of vibration on a continuous manual control task is the direct mechanical jostling of the hand causing unwanted movement of the control. This is sometimes called *breakthrough* or *feedthrough* or *vibration-correlated error*. The inadvertent movement of a pencil caused by "jostling" while writing in a vehicle is a form of *vibration-correlated error*. In a simple tracking task, where the operator is required to follow movements of a target, some of the error will also be correlated with the target movements. This is called *input-correlated error* and often mainly reflects the inability of an operator to follow the target without delays inherent in visual, cognitive, and motor activity. The part of the tracking error which is not correlated with either the vibration or the tracking task is called the 'remnant'. This includes operator-generated *noise* and any source of non-linearity: drawing a freehand straight line does not result in a perfect straight line even in the absence of environmental vibration. The effects of vibration on vision can result in increased remnant with some tracking tasks and some studies show



**Figure 6** Average times taken to read information on a display for (i) stationary observers reading from a vibrating display, (ii) vibrating observers reading from a stationary display, and (iii) vibration observers reading from a vibrating display with observer and display vibrating in phase. Data obtained with sinusoidal vertical vibration at 2.0m s$^{-2}$ r.m.s. (from Moseley and Griffin, 1986b).

**Figure 7** Linear model of a pursuit manual control system showing how tracking errors may be caused by the vibration (vibration-correlated error), the task (input-correlated error), or some other cause (remnant).

that vibration, usually at frequencies greater than about 20 Hz, interferes with neuromuscular processes, which may be expected to result in increased remnant. The causes of the three components of the tracking error are shown in the model presented as Figure 7.

***Effects of Task Variables*** The gain (i.e., sensitivity) of a control determines the control output corresponding to a given force, or displacement, applied to the control by the operator. The optimum gain in static conditions (high enough to not cause fatigue but low enough to prevent inadvertent movement) is likely to be greater than the optimum gain during exposure to vibration where inadvertent movement is more likely (Lewis and Griffin, 1977). First-order and second-order control tasks (i.e., rate and acceleration control tasks) are more difficult than zero-order tasks (i.e., displacement control tasks) and so tend to give more errors. However, there may sometimes be advantages with such controls that are less affected by vibration breakthrough at higher vibration frequencies.

In static conditions, isometric controls (that respond to force without movement) tend to result in better tracking performance than isotonic controls (that respond to movement but require the application of no force). However, several studies show that isometric controls may suffer more from the effects of vibration (e.g., Allen et al., 1973; Levison and Harrah, 1977). The relative merits of the two types of control and the optimum characteristics of a spring-centered control will depend on control gain and control order.

The results of studies investigating the influence of the position of a control appear consistent with differences being dependent on the transmission of vibration to the hand in different positions (e.g., Shoenberger and Wilburn, 1973). Torle (1965) showed that the provision of an armrest could substantially reduce the effects of vibration on the performance of a task with a side-arm controller. The shape and orientation of controls may also be expected to affect performance—either by modifying the amount of

vibration breakthrough or by altering the proprioceptive feedback to the operator.

Vibration may affect the performance of tracking tasks by reducing the visual performance of the operator. Wilson (1974) and McLeod and Griffin (1990) have shown that collimating a display by means of a lens so that the display appears to be at infinity can reduce, or even eliminate, errors with some tasks. It is possible that visual disruption has played a significant part in the performance decrements reported in other experimental studies of the effects of vibration on manual control.

Some simple tasks can be so easy that they are immune to disruption by vibration. At the other extreme, a task may be so difficult that any additional difficulty caused by vibration may be insignificant. Tasks with moderate ranges of difficulty in static conditions tend to be most disrupted by whole-body vibration.

***Effects of Vibration Variables*** The vibration transmissibility of the body is approximately linear (i.e., doubling the magnitude of vibration at the seat may be expected to approximately double the magnitude of vibration at the head or at the hand). Vibration-correlated error may therefore increase in approximately linear proportion to vibration magnitude.

There is no simple relation between the frequency of vibration and its effects on control performance. The effects of frequency depend on the control order (that varies between tasks) and the biodynamic responses of the body (that varies with posture and between operators). With zero-order tasks and the same magnitude of acceleration at each frequency, the effects of vertical seat vibration may be greatest in the range 3 to 8 Hz since transmissibility to the shoulders is greatest in this range (see McLeod and Griffin, 1989). With horizontal whole-body vibration (i.e., in the $x$- and $y$-axes of the seated body) the greatest effects appear to occur at lower frequencies: around 2 Hz or below. Again, this corresponds to the frequencies at which there is greatest transmission of vibration to the shoulders. The axis of the control task most affected by vibration may not be the same axis as that in which most vibration occurs at

the seat. Often, fore-and-aft movements of the control (which generally correspond to vertical movements on a display) are most affected by vertical whole-body vibration. Few controls are sensitive to vertical hand movements and these have rarely been studied.

Multiple-frequency vibration causes more disruption to manual control performance than the presentation of any one of the constituent single frequencies alone. Similarly, the effects of multiple-axis vibration are greater than the effects of any single axis vibration.

The impression that prolonged exposure to vibration causes fatigue gave rise to the *fatigue-decreased proficiency* boundary in International Standard 2631, first published in 1974 (ISO, 1974, 1985, p. 4). This standard proposed a complex time-dependent magnitude of vibration which is said to be "a limit beyond which exposure to vibration can be regarded as carrying a significant risk of impaired working efficiency in many kinds of tasks, particularly those in which time-dependent effects ("fatigue") are known to worsen performance as, for example, in vehicle driving." Reviews of experimental studies show time-dependent effects of performance in only a few cases, with performance sometimes improving with time. It can be concluded that the evidence supporting the old ISO 2631 fatigue-decreased proficiency boundary is weak or nonexistent. There are no data justifying a time-dependent limit for the effects of vibration on performance with the complexity included in International Standard 2631 (ISO, 1974, 1985). Any duration-dependent effects of vibration may be influenced by complex central factors including motivation, arousal, and similar concepts that depend on the form of the task: they may not lend themselves to satisfactory representation by a single time-dependent limit in an international standard. The most common and most easily understood 'direct' effects of vibration on vision and manual control are not intrinsically dependent on the duration of vibration exposure.

***Other Variables***   Repeated exposure to vibration may allow subjects to develop techniques for minimizing vibration effects by, for example, adjusting body posture to reduce the transmission of vibration to the head or the hand or by learning how to recognize images blurred by vibration. Results of experiments performed in one experimental session of vibration exposure may not necessarily apply to situations where operators have an opportunity to learn techniques to ameliorate the effects of vibration.

There have been few investigations of the effects of vibration on common everyday tasks. Corbridge and Griffin (1991) found that the effects of vertical whole-body vibration on spilling liquid from a hand-held cup were greatest close to 4 Hz. They also found that the effects of vibration on writing speed and subjective estimates of writing difficulty were most affected by vertical vibration in the range 4 to 8 Hz. Although 4 Hz was a sensitive frequency for both the drinking task and the writing task, the dependence on frequency of the effects of vibration were different for the two activities.

Whole-body vibration can cause a warbling of speech due to fluctuations in the airflow through the larynx.

Greatest effects may occur with vertical vibration in the range 5–20 Hz, but they are not usually sufficient to reduce greatly the intelligibility of speech (e.g., Nixon and Sommer, 1963). Some studies suggest that exposure to vibration may contribute to noise-induced hearing loss, but further study is required to allow a full interpretation of these data.

### 3.2.3   Cognitive Tasks

To be useful, studies of cognitive effects of vibration must be able to show that any changes associated with exposure to vibration were not caused by vibration affecting input processes (e.g., vision) or output processes (e.g., hand control). Only a few investigators have addressed possible cognitive effects of vibration with care and considered such problems. For example, Shoenberger (1974) found that with the Sternberg memory-reaction-time task the time taken for subjects to recall letters presented on a display depended on the angular size of the letters. He was able to conclude that performance was degraded by visual effects of vibration and not by cognitive effects of vibration. In most other studies there has been little attempt to develop hypotheses to explain any significant effects of vibration in terms of the component processes involved in cognitive processing.

Simple cognitive tasks (e.g., simple reaction time) appear to be unaffected by vibration, other than by changes in arousal or motivation or by direct effects on input and output processes. This may also be true for some complex cognitive tasks. However, the scarcity and diversity of experimental studies allow the possibility of real and significant cognitive effects of vibration (see Sherwood and Griffin, 1990, 1992). Vibration may influence 'fatigue', but there is no scientific foundation for the *fatigue-decreased proficiency limit* offered in the former International Standard 2631 (ISO, 1974, 1985).

### 3.3   Health Effects

Epidemiological studies have reported disorders among persons exposed to vibration from occupational, sport, and leisure activities [see Dupuis and Zerlett, 1986; Griffin, 1990; National Institute for Occupational Safety and Health (NIOSH), 1997; Bovenzi and Hulshof, 1998; Bovenzi, 2009]. The studies do not all agree on either the type or the extent of disorders and rarely have the findings been related to measurements of the vibration exposures. However, the incidence of some disorders of the back (back pain, displacement of intervertebral discs, degeneration of spinal vertebrae, osteoarthritis, etc.) appears to be greater in some groups of vehicle operators, and it is thought that this is sometimes associated with their vibration exposure. There may be several alternative causes of an increase in disorders of the back among persons exposed to vibration (e.g., poor sitting postures, heavy lifting). It is not always possible to conclude confidently that a back disorder in a person occupationally exposed to whole-body vibration is solely, or primarily, caused by vibration.

Other disorders that have been claimed to be due to occupational exposures to whole-body vibration include abdominal pain, digestive disorders, urinary

frequency, prostatitis, hemorrhoids, balance and visual disorders, headaches, and sleeplessness. Further research is required to confirm whether these signs and symptoms are causally related to exposure to vibration.

### 3.3.1 Vibration Evaluation

Epidemiological data alone are not sufficient to define how to evaluate whole-body vibration so as to predict the relative risks to health from the different types of vibration exposure. A consideration of such data in combination with an understanding of biodynamic responses and subjective responses is used to provide current guidance. The manner in which the health effects of oscillatory motions depend upon the frequency, direction, and duration of motion is currently assumed to be similar to that for vibration discomfort (see Section 3.1). However, it is assumed that the *total exposure*, rather than the *average exposure*, is important and so a *dose measure* is used.

British Standard 6841 (BSI, 1987) and International Standard 2631-1 (ISO, 1997) can be interpreted as providing similar guidance, but there is more than one method within ISO 2631-1 and the alternative methods can yield different conclusions (Griffin, 1998).

### 3.3.2 British Standard 6841

British Standard 6841 (BSI, 1987) defines an *action level* for vertical vibration based on *vibration dose values*. The vibration dose value uses a 'fourth-power' time-dependency to accumulate vibration severity over the exposure period from the shortest possible shock to a full day of vibration:

$$\text{Vibration dose value} = \left[ \int_{t=0}^{t=T} a^4(t) dt \right]^{1/4} \qquad (1)$$

where $a(t)$ is the frequency-weighted acceleration. If the exposure duration ($t$, seconds) and the frequency-weighted r.m.s. acceleration ($a_{rms}$, m s$^{-2}$ r.m.s.) are known for conditions in which the vibration characteristics are statistically stationary, it can be useful to calculate the *estimated vibration dose value* (eVDV):

$$\text{eVDV} = 1.4 a_{rms} t^{1/4} \qquad (2)$$

The eVDV is not applicable to transients, shocks, or repeated shock motions in which the crest factor (peak value divided by the r.m.s. value) is high.

No precise limit can be offered to prevent disorders caused by whole-body vibration, but British Standard 6841 (BSI, 1987, p. 18) offers the following guidance: "High vibration dose values will cause severe discomfort, pain and injury. Vibration dose values also indicate, in a general way, the severity of the vibration exposures which caused them. However there is currently no consensus of opinion on the precise relation between vibration dose values and the risk of injury. It is known that vibration magnitudes and durations which produce vibration dose values in the region of 15 m s$^{-1.75}$ will usually cause severe discomfort. It is reasonable to assume that increased exposure to vibration will be accompanied by increased risk of injury." An action level might be set higher or lower than 15 m s$^{-1.75}$. Figure 8 shows this action level for exposure durations from 1 s to one day.

### 3.3.3 International Standard 2631

International Standard 2631 (ISO, 1997, p. 22) offers two different methods of evaluating vibration severity with respect to health effects, and for both methods there are two boundaries. When evaluating vibration using the vibration dose value, it is suggested that below a boundary corresponding to vibration dose value of 8.5 m s$^{-1.75}$ "health risks have not been objectively observed," between 8.5 and 17 m s$^{-1.75}$ "caution with respect to health risks is indicated," and above 17 m s$^{-1.75}$ "health risks are likely." The two boundaries define a *VDV health guidance caution zone*. The alternative method of evaluation in ISO 2631 (ISO, 1997) uses a time dependency in which the acceptable vibration does not vary with duration between 1 and 10 min and then decreases in inverse proportion to the square root of duration from 10 min to 24 h. This method suggests an *r.m.s. health guidance caution zone*, but the method is not fully defined in the text, it allows very high accelerations at some durations, it conflicts with the vibration dose value method, and it may not be applicable to exposure durations less than 1 min (Figure 8).

When the possibility of severe exposures to vibration or shock can be foreseen, it is appropriate to consider the fitness of the exposed persons, warn of the risks and train on ways of minimizing risks, minimize the duration of exposure to vibration, and minimize the magnitude of exposure (by suitable selection and maintenance of machinery or driving routes and the design of antivibration devices). Suitable health surveillance and monitoring of vibration-exposed persons may also be appropriate.

### 3.3.4 EU Machinery Safety Directive

The Machinery Safety Directive of the European Community (2006/42/EC, paragraph 1.5.9) states: "Machinery must be designed and constructed in such a way that risks resulting from vibrations produced by the machinery are reduced to the lowest level, taking account of technical progress and the availability of means of reducing vibration, in particular at source" (European Parliament and the Council of the European Union, 2006). The instruction handbooks for machinery causing whole-body vibration must specify the frequency-weighted acceleration if this exceeds a frequency-weighted acceleration of 0.5 m s$^{-2}$ r.m.s. The relevance of any such value will depend on the test conditions to be specified in other standards. Many work vehicles exceed this value at some stage during an operation or journey. Standardized procedures for testing work vehicles are being prepared; the values currently quoted by manufacturers may not always be representative of the operating conditions in the work for which the machinery is used. The Machinery Safety Directive
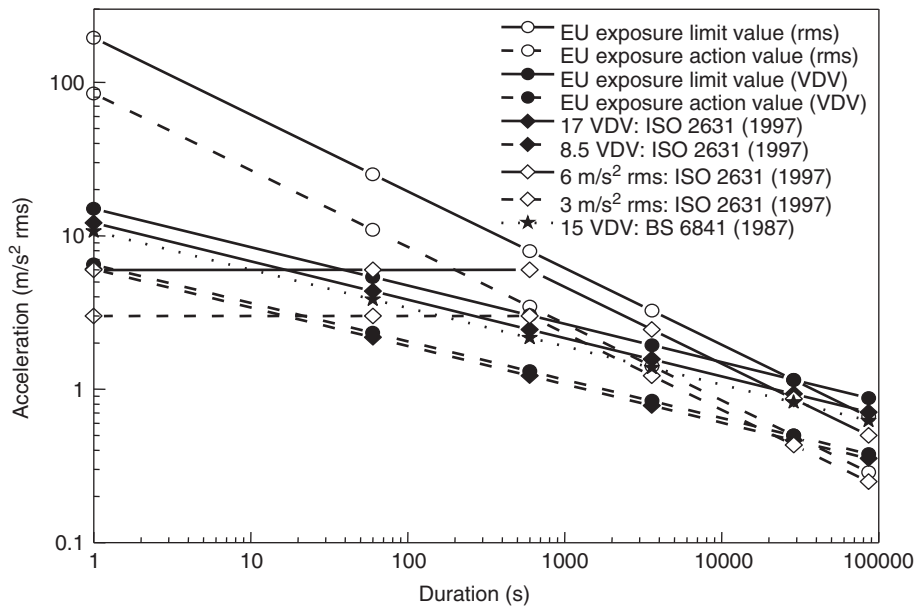
**Figure 8** Comparison between health guidance caution zones for whole-body vibration in ISO 2631-1 (ISO, 1997) (3 to 6 ms$^{-2}$ r.m.s.; 8.5 to 17m s$^{-1.75}$), 15m s$^{-1.75}$ action level implied in BS 6841 (BSI, 1987), the r.m.s. and eVDV exposure limit values and exposure action values for whole-body vibration in the EU Physical Agents (Vibration) Directive.

affects all manufacturers wishing to sell machines in the EU and therefore has a worldwide impact.

### 3.3.5 EU Physical Agents Directive (2002)

The Parliament and Commission of the European Community have defined *minimum health and safety requirements* for the exposure of workers to the risks arising from vibration (European Parliament and the Council of the European Union, 2002). For whole-body vibration, the Directive defines an 8-h equivalent *exposure action value* of 0.5 m s$^{-2}$ r.m.s. (or a vibration dose value of 9.1 m s$^{-1.75}$) and an 8-h equivalent *exposure limit value* of 1.15 ms$^{-2}$ r.m.s. (or a vibration dose value of 21 m s$^{-1.75}$).

The Directive says that workers shall not be exposed above the 'exposure limit value'. If the 'exposure action value' is exceeded, the employer shall establish and implement a program of technical and/or organizational measures intended to reduce to a minimum exposure to mechanical vibration and the attendant risks. The Directive says workers exposed to vibration in excess of the exposure action values shall be entitled to appropriate health surveillance. Health surveillance is also required if there is any reason to suspect that workers may be injured by the vibration even if the exposure action value is not exceeded.

The probability of injury arising from occupational exposures to whole-body vibration at the exposure action value and the exposure limit value cannot be estimated because epidemiological studies have not yet produced dose–response relationships. However, it seems clear that the Directive does not define *safe exposures* to whole-body vibration since the r.m.s.

values are associated with extraordinarily high magnitudes of vibration (and shock) when the exposures are short: these exposures may be assumed to be hazardous (see Figure 8; Griffin, 2004). The vibration dose value procedure suggests more reasonable vibration magnitudes for short-duration exposures.

### 3.4 Disturbance in Buildings

Acceptable magnitudes of vibration in buildings are generally close to, or below, vibration perception thresholds (Morioka and Griffin, 2008). The effects of vibration in buildings are assumed to depend on the use of the building in addition to the vibration frequency, the vibration direction and the vibration duration. International Standard 2631-2 (ISO, 2003) provides some information on the measurement and evaluation of building vibration, but limited practical guidance. British Standard 6472-1 (BSI, 2008) offers guidance on the measurement, the evaluation, and the assessment of vibration in buildings, and BS 6472-2 (BSI, 2008) defines a method used for assessing the vibration of buildings caused by blasting. Using the guidance contained in BS 6472-1 (BSI, 2008) it is possible to predict the acceptability of vibration in different types of building by reference to a simple table of vibration dose values [see Table 5 and British Standard 6472-1 (BSI, 2008)]. The vibration dose values in Table 5 are applicable irrespective of whether the vibration occurs as a continuous vibration, intermittent vibration, or repeated shocks.

### 3.5 Biodynamics

The human body is a complex mechanical system that does not, in general, respond to vibration in the same

**Table 5 Vibration Dose Value Ranges Expected to Result in Various Degrees of Adverse Comment in Residential Buildings**

| Place | Low Probability Of Adverse Comment (ms$^{-1.75}$) | Adverse Comment Possible (ms$^{-1.75}$) | Adverse Comment Probable (ms$^{-1.75}$) |
|---|---|---|---|
| Residential buildings 16-h day | 0.2–0.4 | 0.4–0.8 | 0.8–1.6 |
| Residential buildings 8-h night | 0.1–0.2 | 0.2–0.4 | 0.4–0.8 |

*Note:* Based on British Standard 6472-1 (BSI, 2008). For offices and workshops, multiplying factors of 2 and 4, respectively, can be applied to the vibration dose value ranges for a 16-h day.

manner as a rigid mass: relative motions between the body parts vary with the frequency and the direction of the applied vibration. Although there are resonances in the body, it is over-simplistic to summarize the dynamic responses of the body by merely mentioning one or two resonance frequencies. The biodynamics of the body affect human responses to vibration, but the discomfort, the interference with activities, and the health effects of vibration cannot be well predicted solely by considering the body as a mechanical system.

### 3.5.1 Transmissibility of Human Body

The extent to which the vibration at the input to the body (e.g., the vertical vibration at a seat) is transmitted to a part of the body (e.g., vertical vibration at the head or the hand) is described by the transmissibility. At low frequencies of oscillation (e.g., less than about 1 Hz), the oscillations of the seat and the body are very similar and so the transmissibility is approximately 1.0. With increasing frequency of oscillation the motions of the body increase above that measured at the seat; the ratio of the motion of the body to the motion of the seat will reach a peak at one or more frequencies (i.e., resonance frequencies). At high frequencies the body motion will be less than that at the seat.

The resonance frequencies and the transmissibilities at resonance vary according to where the vibration is measured on the body and the posture of the body. For seated persons, there may be resonances to the head and the hand at frequencies in the range 4 to 12 Hz when exposed to vertical vibration, at frequencies less than 4 Hz when exposed to $x$-axis (i.e., fore-and-aft) vibration, and less than 2 Hz when exposed to $y$-axis (i.e., lateral) vibration (see Paddan and Griffin, 1988a,b). A seat back can greatly increase the transmission of fore-and-aft seat vibration to the head and upper-body of seated people, and bending of the legs can greatly affect the transmission of vertical vibration to the heads of standing persons.

### 3.5.2 Mechanical Impedance of Human Body

Mechanical impedance reflects the relation between the driving force at the input to the body and the resultant movement of the body. If the human body were rigid, the ratio of force to acceleration applied to the body would be constant and indicate the mass of the body. Because the human body is not rigid, the ratio of force to acceleration is only close to the body mass at very low frequencies (less than about 2 Hz with vertical vibration; less than about 1 Hz with horizontal vibration).

Measures of mechanical impedance usually show a principal resonance for the vertical vibration of seated people at about 5 Hz, and sometimes a second resonance in the range 7 to 12 Hz (Fairley and Griffin, 1989; Matsumoto and Griffin, 2000, 2001; Nawayseh and Griffin, 2003). Unlike some of the resonances affecting the transmissibility of the body, these resonances are only influenced by movement of large masses close to the input of vibration to the body. The large difference in impedance between that of a rigid mass and that of the human body means that the body cannot usually be represented by a rigid mass when measuring the vibration transmitted through seats.

### 3.5.3 Biodynamic Models

Various mathematical models of the responses of the body to vibration have been developed. A simple model with one or two degrees-of-freedom can provide an adequate representation of the vertical mechanical impedance of the body (e.g., Fairley and Griffin, 1989; Wei and Griffin, 1998a) and be used to predict the transmissibility of seats (Wei and Griffin, 1998b) or construct an anthropodynamic dummy for seat testing (Lewis and Griffin, 2002). Compared with mechanical impedance, the transmissibility of the body is affected by many more variables and so requires a more complex model reflecting the posture of the body and the translation and rotation associated with the various modes of vibration (Matsumoto and Griffin, 2001).

### 3.6 Protection from Whole-Body Vibration

Wherever possible, vibration should be reduced at the source. This may involve reducing the undulations of the terrain, or reducing the speed of travel of vehicles, or improving the balance of rotating parts. Methods of reducing the transmission of vibration to operators require an understanding of the characteristics of the vibration environment and the route for the transmission of vibration to the body. For example, the magnitude of vibration often varies with location: lower magnitudes will be experienced in some areas adjacent to machinery or in different parts of vehicles.

### 3.6.1 Seating Dynamics

Most seats exhibit a resonance at low frequencies that results in higher magnitudes of vertical vibration occurring on the seat than on the floor! At high frequencies there is usually attenuation of vibration. The resonance frequencies of common seats are usually in the region of 4 Hz (see Figure 9). The amplification at resonance is partially determined by the 'damping' in the seat. Increases in the damping of a seat cushion tend to reduce the amplification at resonance but increase the transmission of vibration at higher frequencies. The
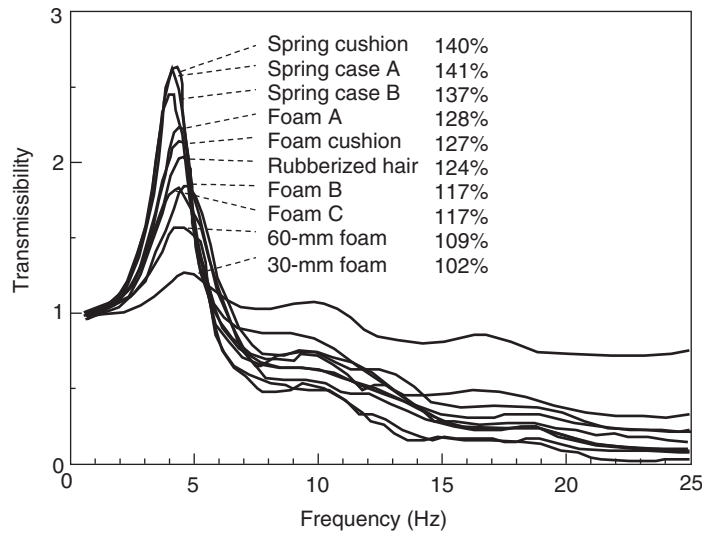
**Figure 9** Comparison of vertical transmissibilities and SEAT values for 10 alternative cushions of passenger railway seats (data from Corbridge et al., 1989).

variations in transmissibility between seats are sufficient to result in significant differences in the vibration experienced by people supported by different seats.

A simple numerical indication of the isolation efficiency of a seat for a specific application is provided by the *seat effective amplitude transmissibility* (SEAT) (Griffin, 1990). A SEAT value greater than 100% indicates that, overall, the vibration on the seat is 'worse' than the vibration on the floor beneath the seat:

$$SEAT(\%) = \frac{\text{ride comfort seat}}{\text{ride comfort floor}} \times 100$$

Values below 100% indicate that the seat has provided some useful attenuation. Seats should be designed to have the lowest SEAT value compatible with other constraints.

In practice, the SEAT value is a mathematical procedure for predicting the effect of a seat on ride comfort. The *ride comfort* that would result from sitting on the seat or on the floor can be predicted using the frequency weightings in the appropriate standard. The SEAT value may be calculated from the r.m.s. values or the vibration dose values of the frequency-weighted acceleration on the seat and the floor:

$$SEAT(\%) = \frac{\text{vibration dose value on seat}}{\text{vibration dose value on floor}} \times 100$$

The SEAT value is a characteristic of the vibration input and not merely a description of the dynamics of the seat: different values are obtained with the same seat in different vehicles. The SEAT value indicates the suitability of a seat for a particular type of vibration.

A separate suspension mechanism is provided beneath the seat pan in *suspension seats*. These seats,

used in some off-road vehicles, trucks, and coaches, have low resonance frequencies (often less than about 2 Hz) and so can attenuate vibration at frequencies much greater than 2 Hz. The transmissibilities of these seats are usually determined by the seat manufacturer, but their isolation efficiencies vary with operating conditions.

## 4 MOTION SICKNESS

Motion sickness is not an illness but a normal response to motion that is experienced by many fit and healthy people. A variety of different motions can cause sickness and reduce the comfort, impede the activities, and degrade the well-being of both those directly affected and those associated with the motion sick. Although vomiting can be the most inconvenient consequence, other effects (e.g., yawning, cold sweating, nausea, stomach awareness, dry mouth, increased salivation, headaches, bodily warmth, dizziness, and drowsiness) can also be unpleasant. In some cases the symptoms can be so severe as to result in reduced motivation to survive difficult situations.

### 4.1 Causes of Motion Sickness

Motion sickness can be caused by many different movements of the body (e.g., translational and rotational oscillation, constant speed rotation about an off-vertical axis, Coriolis stimulation), movements of the visual scene, and various other stimuli producing sensations associated with movement of the body (see Table 6 and Griffin, 1991). Motion sickness is neither explained nor predicted solely by the physical characteristics of the motion, although some motions can reliably be predicted as more nauseogenic than others.

**Table 6 Examples of Environments, Activities, and Devices that Can Cause Symptoms of Motion Sickness**

| | |
|---|---|
| Boats | Camel rides |
| Ships | Elephant rides |
| Submarines | |
| Hydrofoils | Vehicle simulators |
| Hovercraft | |
| Swimming | Fairground devices |
| Fixed-wing aircraft | Cinerama |
| Helicopters | Inverting or distorting spectacles |
| Spacecraft | Microfiche readers |
| | Head-coupled visual displays |
| Cars | |
| Coaches | Rotation about off-vertical axis |
| Buses | Coriolis stimulation |
| Trains | Low-frequency translational oscillation |
| Tanks | |

Motions of the body may be detected by three basic sensory systems: the vestibular system, the visual system, and the somatosensory system. The vestibular system is located in the inner ear and comprises the semicircular canals, which respond to the rotation of the head, and the otoliths, which respond to translational forces (either translational acceleration or rotation of the head relative to an acceleration field, such as the force of gravity). The eyes may detect relative motion between the head and the environment, caused by either head movements (in translation or rotation) or movements of the environment or a combination of the movements of the head and the environment. The somatosensory systems respond to force and displacement of parts of the body and give rise to sensations of body movement, or force.

It is assumed that in 'normal' environments the movements of the body are detected by all three sensory systems and that this leads to an unambiguous indication of the movements of the body in space. In some other environments the three sensory systems may give signals corresponding to different motions (or motions that are not realistic) and lead to some form of conflict. This leads to the idea of a *sensory conflict theory* of motion sickness in which sickness occurs when the sensory systems disagree on the motions which are occurring. However, this implies some absolute significance to sensory information, whereas the 'meaning' of the information is probably learned. This led to the *sensory rearrangement theory* of motion sickness that states: all situations which provoke motion sickness are characterized by a condition of sensory rearrangement in which the motion signals transmitted by the eyes, the vestibular system and the non-vestibular proprioceptors are at variance either with one another or with what is expected from previous experience (Reason, 1970, 1978). Reason and Brand (1975) suggest that the conflict may be sufficiently

considered in two categories: intermodality (between vision and the vestibular receptors) and intramodality (between the semicircular canals and the otoliths within the vestibular system). For both categories it is possible to identify three types of situations in which conflict can occur (see Table 7). The theory implies that all situations which provoke motion sickness can be fitted into one of the six conditions shown in Table 7 (see Griffin, 1990).

There is evidence that the average susceptibility to sickness among males is less than that among females, and susceptibility decreases with increased age among both males and females (Lawther and Griffin, 1988a; Turner and Griffin, 1999). However, there are larger individual differences within any group of either gender at any age: some people are easily made ill by motions that can be endured indefinitely by others. The reasons for these differences are not properly understood.

## 4.2 Sickness Caused by Oscillatory Motion

Motion sickness is not caused by oscillation (however violent) at frequencies much greater than about 1 Hz: the phenomenon arises from motions at the low frequencies associated with normal postural control of the body. Various experimental investigations have explored the extent to which vertical oscillation causes sickness at different frequencies. These studies have allowed the formulation of a frequency weighting, $W_f$ (see Figure 4), and the definition of a *motion sickness dose value*, MSDV. The frequency weighting $W_f$ reflects greatest sensitivity to acceleration in the range 0.125 to 0.25 Hz, with a rapid reduction in sensitivity at higher frequencies. The motion sickness dose value predicts the probability of sickness from knowledge of the frequency and magnitude of vertical oscillation (see Lawther and Griffin, 1987; British Standard 6841, 1987; International Standard 2631-1, 1997):

$$\text{Motion sickness dose value} = a_{rms}t^{1/2}$$

where $a_{rms}$ is the r.m.s. value of the frequency-weighted acceleration (m s$^{-2}$ r.m.s.) and $t$ is the exposure period (seconds). The percentage of unadapted adults who are expected to vomit is given by 1/3 MSDV. (These relationships have been derived from exposures in which up to 70% of persons vomited during exposures lasting between 20 min and 6 h).

The motion sickness dose value has been used for the prediction of sickness on various marine craft (ships, hovercraft, and hydrofoil) in which vertical oscillation has been shown to be a prime cause of sickness (Lawther and Griffin, 1988b). Vertical oscillation is not the principal cause of sickness in many road vehicles (Turner and Griffin, 1999; Griffin and Newman, 2004) and some other environments: the above expression should not be assumed to be applicable to the prediction of sickness in all environments.

## 5 HAND-TRANSMITTED VIBRATION

Prolonged and regular exposure of the fingers or the hands to vibration or repeated shock can give rise to

**Table 7  Type of Motion Cue Mismatch Produced by Various Provocative Stimuli**

| | Category of Motion Cue Mismatch | |
|---|---|---|
| | Visual (A) / Vestibular (B) | Canal (A) / Otolith (B) |
| TYPE 1<br>  A and B simultaneously give contradictory or uncorrelated information | Watching waves from a ship | Making head movements while rotating (Coriolis or cross-coupled stimulation) |
| | Use of binoculars in a moving vehicle | Making head movements in an abnormal environment which may be constant (e.g., hyper- or hypogravity) or fluctuating (e.g., linear oscillation) |
| | Making head movements when vision is distorted by an optical device | Space sickness |
| | ''Pseudo-Coriolis'' stimulation | Vestibular disorders (e.g., Ménière's disease, acute labyrinthitis, trauma labyrinthectomy) |
| TYPE IIa<br>  A signals in absence of expected B signal | Cinerama sickness | Positional alcohol nystagmus |
| | Simulator sickness | Caloric stimulation of semicircular canals |
| | ''Haunted swing'' | Vestibular disorders (e.g., pressure vertigo, cupulolithiasis) |
| | Circular vection | |
| TYPE IIb<br>  B signals in absence of expected A signals | Looking inside a moving vehicle without external visual reference (e.g., below deck in a boat) | Low-frequency ($<0.5$ Hz) translational oscillation |
| | Reading in a moving vehicle | Rotating linear acceleration vector (e.g., ''barbecue spit'' rotation, rotation about an off-vertical axis) |

*Source*: Adapted from Benson (1984).

various signs and symptoms of disorder. The precise extent and interrelation between the signs and symptoms are not fully understood, but five types of disorder may be identified (see Table 8).

The various disorders may be interconnected: more than one disorder can affect a person at the same time and it is possible that the presence of one disorder facilitates the appearance of another. The onset of each disorder is dependent on several variables, such as the vibration characteristics, the dynamic response of the fingers or hand, individual susceptibility to damage, and other aspects of the environment. The terms *vibration*

**Table 8  Five Types of Disorder Associated with Hand-Transmitted Vibration Exposures**

| Type | Disorder |
|---|---|
| Type A | Circulatory disorders |
| Type B | Bone and joint disorders |
| Type C | Neurological disorders |
| Type D | Muscle disorders |
| Type E | Other general disorders (e.g., central nervous system) |

*Source*: From Griffin (1990).
*Note*: Some combination of these disorders is sometimes referred to as the hand–arm vibration syndrome (HAVS).

*syndrome* and *hand–arm vibration syndrome* (HAVS) are sometimes used to refer to one or more of the effects listed in Table 8.

## 5.1  Sources of Hand-Transmitted Vibration

The vibration on tools varies greatly depending on tool design and method of use, so it is not possible to categorize individual tool types as *safe* or *dangerous*. However, Table 9 lists examples of tools and processes that are sometimes a cause for concern.

## 5.2  Effects of Hand-Transmitted Vibration

### 5.2.1  Vascular Disorders

The first published cases of the condition now most commonly known as *vibration-induced white finger* (VWF) are acknowledged to be those reported in Italy by Loriga in 1911. A few years later, cases were documented at limestone quarries in Indiana. Vibration-induced white finger has subsequently been reported to occur in many other widely varied occupations in which there are exposures of the fingers to vibration (see Taylor and Pelmear, 1975; Wasserman et al., 1982; Griffin, 1990).

***Signs and Symptoms***  Vibration-induced white finger (VWF), is characterized by intermittent whitening (i.e., blanching) of the fingers (Griffin and Bovenzi,

**Table 9  Examples of Tools and Processes Potentially Associated with Vibration Injuries**

| Type of Tool | Examples of Tool Type |
|---|---|
| Percussive metal-working tools | Riveting tools |
| | Caulking tools |
| | Chipping tools |
| | Chipping hammers |
| | Fettling tools |
| | Hammer drills |
| | Clinching and flanging tools |
| | Impact wrenches |
| | Swaging |
| | Needle guns |
| Grinders and other rotary tools | Pedestal grinders |
| | Hand-held grinders |
| | Hand-held sanders |
| | Hand-held polishers |
| | Flex-driven grinders/polishers |
| | Rotary burring tools |
| Percussive hammers and drills used in mining, demolition and road construction | Hammers |
| | Rock drills |
| | Road drills, etc. |
| Forest and garden machinery | Chain saws |
| | Antivibration chain saws |
| | Brush saws |
| | Mowers and shears |
| | Barking machines |
| Other processes and tools | Nut runners |
| | Shoe-pounding-up machines |
| | Concrete vibro-thickeners |
| | Concrete leveling vibrotables |
| | Motorcycle handle bars |

2002). The finger tips are usually the first to blanch, but the affected area may extend to all of one or more fingers with continued vibration exposure. Attacks of blanching are precipitated by cold and therefore usually occur in cold conditions or when handling cold objects. The blanching lasts until the fingers are rewarmed and vasodilation allows the return of the blood circulation.

Many years of vibration exposure often occur before the first attack of blanching is noticed. Affected persons often have other signs and symptoms, such as numbness and tingling. Cyanosis and, rarely, gangrene have also been reported. It is not yet clear to what extent these other signs and symptoms are causes of, caused by, or unrelated to attacks of white finger.

***Diagnosis***  There are other conditions that can cause similar signs and symptoms to those associated with VWF. Vibration-induced white finger cannot be assumed to be present merely because there are attacks of blanching. It will be necessary to exclude other known causes of similar symptoms (by medical examination) and also necessary to exclude so-called primary Raynaud's disease (also called *constitutional white finger*). This exclusion cannot yet be achieved with complete confidence, but if there is no family history of the symptoms, if the symptoms did not occur before the first significant exposure to vibration, and if the symptoms and signs are confined to areas in contact with the vibration (e.g., the fingers, not the ears), they will often be assumed to indicate vibration-induced white finger.

Diagnostic tests for vibration-induced white finger can be useful, but at present they are not infallible indicators of the disease. The measurement of finger systolic blood pressure following finger cooling and the measurement of finger rewarming times following cooling can be useful, but many others tests are in use (see Griffin and Bovenzi, 2002).

The severity of the effects of vibration is sometimes recorded by reference to the *stage* of the disorder. The staging of vibration-induced white finger is often based on verbal statements made by the affected person recalling an attack of finger blanching, but it may be influenced by evidence in photographs taken during an attack. In the Stockholm Workshop staging system, the staging is influenced by both the frequency of attacks of blanching and the areas of the digits affected by blanching (see Table 10).

A scoring system is used to record the areas of the digits affected by blanching (see Figure 10). The scores correspond to areas of blanching on the digits commencing with the thumb. On the fingers a score of 1 is given for blanching on the distal phalanx, a score of 2 for blanching on the middle phalanx, and a score of 3 for blanching on the proximal phalanx. On the thumbs the scores are 4 for the distal phalanx and 5 for the proximal phalanx. The blanching score may be based on statements from the affected person or on the visual observations of a designated observer (e.g., a nurse).

**Table 10  Stockholm Workshop Scale for Classification of Vibration-Induced White Finger**

| Stage | Grade | Description |
|---|---|---|
| 0 | — | No attacks |
| 1 | Mild | Occasional attacks affecting only the tips of one or more fingers |
| 2 | Moderate | Occasional attacks affecting distal and middle (rarely also proximal) phalanges of one or more fingers |
| 3 | Severe | Frequent attacks affecting all phalanges of most fingers |
| 4 | Very severe | As in stage 3, with trophic skin changes in the finger tips |

*Source*: From Gemne et al., (1987).
*Note*: If a person has stage 2 in two fingers of the left hand and stage 1 in a finger on the right hand, the condition may be reported as 2L(2)/1R(1). There is no defined means of reporting the condition of digits when this varies between digits on the same hand. The scoring system is more helpful when the extent of blanching is to be recorded.
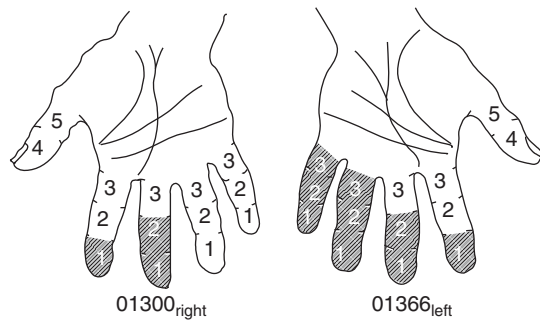
**Figure 10** Method of scoring areas of digits affected by blanching (from Griffin, 1990). The blanching scores for the hands shown are $01300_{right}$, $01366_{left}$.

### 5.2.2 Neurological Disorders

Neurological effects of hand-transmitted vibration (e.g., numbness, tingling, elevated sensory thresholds for touch, vibration, temperature and pain, and reduced nerve conduction velocity) are considered to be separate effects of vibration and not merely symptoms of vibration-induced white finger (Griffin and Bovenzi, 2002). A method of reporting the extent of vibration-induced neurological effects of vibration has been proposed (see Table 11). This staging is not currently related to the results of any specific objective test: the *sensorineural stage* is a subjective impression of a physician based on the statements of the affected person or the results of any available clinical or scientific testing. Neurological disorders are sometimes identified by screening tests using measures of sensory function, such as the thresholds for feeling vibration, heat, or coldness on the fingers.

### 5.2.3 Muscular Effects

The research literature includes reports of muscle atrophy among users of vibrating tools. Workers exposed to hand-transmitted vibration sometimes report difficulty with their grip, including reduced dexterity, reduced grip strength, and locked grip. Many of the reports are derived from symptoms reported by exposed persons rather than signs detected by physicians and could

**Table 11 Proposed Sensorineural Stages of Effects of Hand-Transmitted Vibration**

| Stage | Symptoms |
|---|---|
| $0_{SN}$ | Exposed to vibration but no symptoms |
| $1_{SN}$ | Intermittent numbness with or without tingling |
| $2_{SN}$ | Intermittent or persistent numbness, reduced sensory perception |
| $3_{SN}$ | Intermittent or persistent numbness, reduced tactile discrimination, and/or manipulative dexterity |

Source: From Brammer et al., (1987).

be a reflection of neurological problems (Griffin and Bovenzi, 2002).

Muscle activity may be of great importance to tool users since a secure grip can be essential to the performance of the job and the safe control of the tool. The presence of vibration on a handle may encourage the adoption of a tighter grip than would otherwise occur and a tight grip may increase the transmission of vibration to the hand. If the chronic effects of vibration result in reduced grip, this may sometimes help to protect operators from further effects of vibration, but interfere with both work and leisure activities.

### 5.2.4 Articular Disorders

Many surveys of the users of hand-held tools have found evidence of bone and joint problems, most often among men operating percussive tools such as those used in metal-working jobs and mining and quarrying. It is speculated that some characteristic of such tools, possibly the low-frequency shocks, is responsible. Some of the reported injuries relate to specific bones and suggest the existence of cysts, vacuoles, decalcification, or other osteolysis and degeneration or deformity of the carpal, metacarpal, or phalangeal bones. Osteoarthrosis and olecranon spurs at the elbow and other problems at the wrist and shoulder are also documented.

Notwithstanding the evidence of many research publications, there is not universal acceptance that vibration is a common cause of articular problems and there is currently no dose–effect relation that predicts their occurrence. In the absence of specific information, it seems that adherence to current guidance for the prevention of vibration-induced white finger may provide reasonable protection.

### 5.2.5 Other Effects

Effects of hand-transmitted vibration may not be confined to the fingers, hands, and arms: many studies have found a high incidence of problems such as headaches and sleeplessness among tool users and have concluded that these symptoms are caused by hand-transmitted vibration. Although these are real problems to those affected, they are subjective effects that are not accepted as real by all researchers. Some current research is seeking a physiological basis for such symptoms. It would appear that caution is appropriate, but it is reasonable to assume that the adoption of the modern guidance to prevent vibration-induced white finger will also provide some protection from any other effects of hand-transmitted vibration within, or distant from, the hand.

### 5.3 Preventative Measures

Protection from the effects of hand-transmitted vibration requires actions from management, tool manufacturers, technicians, and physicians at the workplace and from tool users. Table 12 summarizes some of the actions that may be appropriate.

When there is reason to suspect that hand-transmitted vibration may cause injury, the vibration at tool–hand interfaces should be determined (by measurement or

**Table 12 Some Preventative Measures to Consider When Persons are Exposed to Hand-Transmitted Vibration**

| Group | Action |
| --- | --- |
| Management | Seek technical advice |
| | Seek medical advice |
| | Warn exposed persons |
| | Train exposed persons |
| | Review exposure times |
| | Policy on removal from work |
| Tool manufacturers | Measure tool vibration |
| | Design tools to minimize vibration |
| | Ergonomic design to reduce grip force, etc. |
| | Design to keep hands warm |
| | Provide guidance on tool maintenance |
| | Provide warning of dangerous vibration |
| Technical at workplace | Measure vibration exposure |
| | Provide appropriate tools |
| | Maintain tools |
| | Inform management |
| Medical | Pre-employment screening |
| | Routine medical checks |
| | Record all signs and reported symptoms |
| | Warn workers with predisposition |
| | Advise on consequences of exposure |
| | Inform management |
| Tool user | Use tool properly |
| | Avoid unnecessary vibration exposure |
| | Minimize grip and push forces |
| | Check condition of tool |
| | Inform supervisor of tool problems |
| | Keep warm |
| | Wear gloves when safe to do so |
| | Minimize smoking |
| | Seek medical advice if symptoms appear |
| | Inform employer of relevant disorders |

Source: Adapted from Chapter 19 of *The Handbook of Human Vibration* (Griffin, 1990).

by seeking information from other sources, e.g., the tool manufacturer). It will then be possible to predict whether the tool or process is likely to cause injury and whether any other tool or process could give a lower vibration severity. The duration of exposure to vibration should also be quantified. Reduction of exposure time may include the provision of exposure breaks during the day and, if possible, prolonged periods away from vibration exposure. For any tool or process having a vibration magnitude sufficient to cause injury, there should be a system to quantify and control the maximum daily duration of exposure to vibration in any individual.

Gloves are sometimes recommended as a means of reducing the adverse effects of vibration on the hands. International Standard 10819 (ISO, 1996) defines the requirements of *anti-vibration gloves*, but the standard has limitations and cannot be considered a reliable indication of whether a glove is beneficial (Griffin, 1998a). When using the frequency weightings in current standards, most commonly available gloves do *not* normally provide effective attenuation of the vibration on most tools. Gloves and cushioned handles may reduce the transmission of high frequencies of vibration, but current standards imply that these frequencies are not usually the primary cause of disorders. Gloves may protect the hand from other forms of mechanical injury (e.g., cuts and scratches) and protect the fingers from temperature extremes. Warm hands are less likely to suffer an attack of finger blanching and some consider that maintaining warm hands while exposed to vibration may also lessen the damage caused by the vibration.

Workers who are exposed to vibration magnitudes sufficient to cause injury should be warned of the possibility of vibration injuries and educated on the ways of reducing the severity of their vibration exposures. They should be advised of the symptoms to look out for and told to seek medical attention if the symptoms appear. There should be pre-employment medical screening wherever a subsequent exposure to hand-transmitted vibration may reasonably be expected to cause vibration injury. Medical supervision of each exposed person should continue throughout employment at suitable intervals, possibly annually.

### 5.4 Standards for Evaluation of Hand-Transmitted Vibration

There are standards method for measuring, evaluating, and assessing hand-transmitted vibration.

#### 5.4.1 Vibration Measurement

International standards 5349-1 (ISO, 2001) and 5349-2 (ISO, 2002) give general methods of measuring hand-transmitted vibration on tools and processes. Care is required to obtain representative measurements of tool vibration with appropriate operating conditions. There can be difficulties in obtaining valid measurements using some commercial instrumentation (especially when there are high shock levels). It is wise to determine acceleration spectra and inspect the acceleration time-histories before accepting the validity of any measurements.

#### 5.4.2 Vibration Evaluation

All current national and international standards use the same frequency weighting (called $W_h$) to evaluate hand-transmitted vibration over the approximate frequency range 8 to 1000 Hz (Figure 11; Griffin, 1997). This weighting is applied to measurements of vibration acceleration in each of the three axes of vibration at the point of entry of vibration to the hand. More recent standards suggest the overall severity of hand-transmitted vibration should be calculated from root-sums-of-squares of the frequency-weighted acceleration in the three axes.
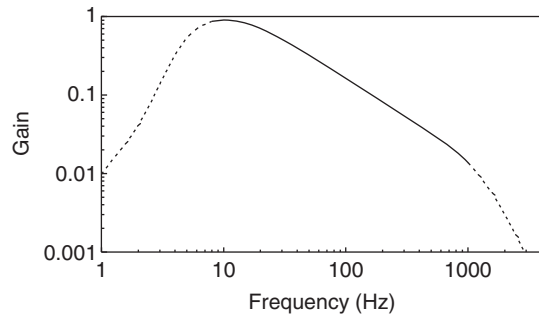
**Figure 11** Frequency weighting $W_h$ for evaluation of hand-transmitted vibration.

The standards imply that if two tools expose the hand to vibration for the same period of time, the tool having the lowest frequency-weighted acceleration will be least likely to cause injury or disease.

Occupational exposures to hand-transmitted vibration can have widely varying daily exposure durations—from a few seconds to many hours. Often, exposures are intermittent. To enable a daily exposure to be reported simply, the standards refer to an equivalent 8-h exposure:

$$a_{\mathrm{hw(eq,8h)}} = A(8) = a_{\mathrm{hw}} \left[ \frac{t}{T_{(8)}} \right]^{1/2}$$

where $t$ is the exposure duration to an r.m.s. frequency-weighted acceleration, $a_{\mathrm{hw}}$, and $T(8)$ is 8 h (in the same units as $t$).

### 5.4.3 Vibration Assessment According to International Standard 5349 (ISO, 2001)

In an informative annex of ISO 5349-1 (ISO, 2001) there is a suggested relation between the lifetime exposure to hand-transmitted vibration, $D_y$ (in years), and the 8-h energy-equivalent daily exposure $A(8)$ for the conditions expected to cause 10% prevalence of finger blanching (Figure 12):

$$D_y = [A(8)]^{-1.06}$$

The percentage of affected persons in any group of exposed persons will not always correspond to the values shown in Figure 12: the frequency weighting, the time-dependency and the dose-effect information are based on less than complete information and they have been simplified for practical convenience. Additionally, the number of persons affected by vibration will depend on the rate at which persons enter and leave the exposed group. The complexity of the above equation implies far greater precision than is possible: a more convenient estimate of the years of exposure (in the range 1 to 25 years) required for 10% incidence of finger blanching is:

$$D_y = \frac{30.0}{A(8)}$$
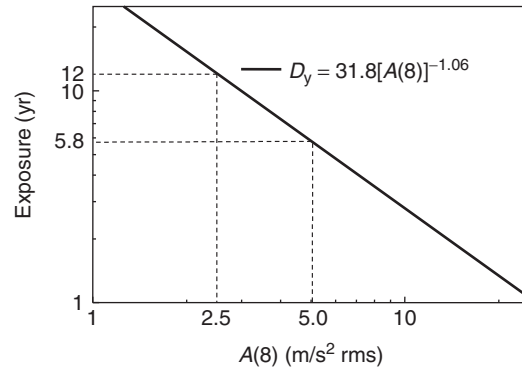


**Figure 12** Relation between daily $A(8)$ and years of exposure expected to result in 10% incidence of finger blanching according to ISO 5349 (ISO, 2001). A 10% probability of finger blanching is predicted after 12 years at the EU exposure action value and after 5.8 years at the EU exposure limit value.

This equation gives the same result as the equation in the standard (to within 14%) and there is no information suggesting it is less accurate.

The informative annex to International Standard 5349 (ISO, 2001, p. 15) states: "Studies suggest that symptoms of the hand-arm vibration syndrome are rare in persons exposed with an 8-h energy-equivalent vibration total value, $A(8)$, at a surface in contact with the hand, of less than 2 m/s$^2$ and unreported for $A(8)$ values less than 1 m/s$^2$." However, this sentence should be interpreted with caution in view of the very considerable doubts over the frequency weighting and time-dependence in the standard (Griffin et al., 2003).

### 5.4.4 EU Machinery Safety Directive

The Machinery Safety Directive of the European Community (2006/42/EC) requires that instruction handbooks for hand-held and hand-guided machinery specify the equivalent acceleration to which the hands or arms are subjected where this exceeds a stated value (currently a frequency-weighted acceleration of 2.5 m s$^{-2}$ r.m.s.) (European Parliament and the Council of the European Union, 2006). Very many hand-held vibrating tools exceed this value. Standards defining test conditions for the measurement of vibration on many tools (e.g., chipping and riveting hammers, rotary hammers and rock drills, grinding machines, pavement breakers, chain saws) have been defined, e.g., ISO 8662 (1988) and ISO 28927 (2009).

### 5.4.5 EU Physical Agents Directive (2002)

For hand-transmitted vibration, the EU Physical Agents Directive defines an 8-h equivalent exposure action value of 2.5 m s$^{-2}$ r.m.s. and an 8-h equivalent exposure limit value of 5.0 m s$^{-2}$ r.m.s. (Figure 13) (European Parliament and the Council of the European Union, 2002). The Directive says workers shall not be exposed above the exposure limit value. If the
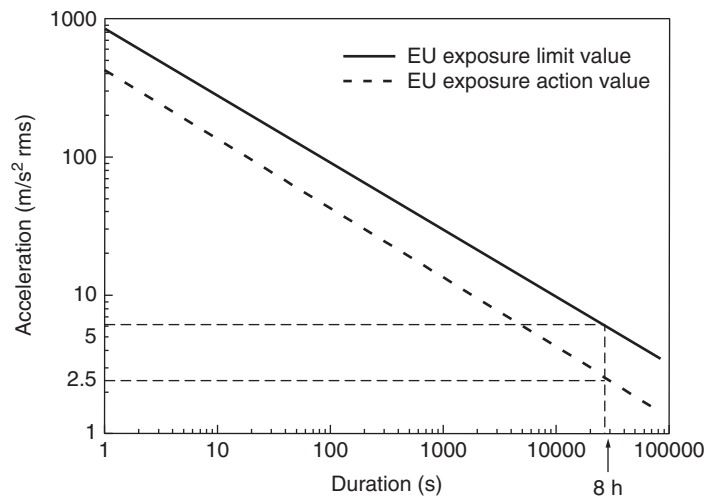
**Figure 13** Hand-transmitted vibration exposure limit value [$A(8) = 5.0$ m s$^{-2}$ r.m.s.] and 'exposure action value' [$A(8) = 2.5$ m s$^{-2}$ r.m.s.] in the EU Physical Agents (vibration) Directive.

exposure action values are exceeded, the employer shall establish and implement a program of technical and/or organizational measures intended to reduce to a minimum exposure to mechanical vibration and the attendant risks. The Directive requires that workers exposed to mechanical vibration in excess of the exposure action values shall be entitled to appropriate health surveillance. However, health surveillance is not restricted to situations where the exposure action value is exceeded: health surveillance is required if there is any reason to suspect that workers may be injured by the vibration, even if the action value is not exceeded.

According to ISO 5349-1 (ISO, 2001), the onset of finger blanching would be expected in 10% of persons after 12 years at the EU exposure action value and after 5.8 years at the exposure limit value (see Figure 12). The exposure action value and the exposure limit value in the EU Directive do not define safe exposures to hand-transmitted vibration (Griffin, 2004).

## REFERENCES

Allen, R. W., Jex, H. R., and Magdaleno, R. E. (1973), "Manual Control Performance and Dynamic Response during Sinusoidal Vibration," AMRL-TR-73-78, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, OH.

Benson, A. J. (1984), "Motion Sickness," in *Vertigo*, M. R. Dix and J. S. Hood, Eds., Wiley, New York.

Benson, A. J., Barnes, G. R. (1978), "Vision during Angular Oscillation: The Dynamic Interaction of Visual and Vestibular Mechanisms," *Aviation, Space And Environmental Medicine*, Vol. 49, No. 1, Section II, pp. 340–345.

Bovenzi, M. (2009), "Metrics of Whole-Body Vibration and Exposure-Response Relationships for Low Back Pain in Professional Drivers: A Prospective Cohort Study," *International Archives of Occupational and Environmental Health*, Vol. 82, pp. 893–917.

Bovenzi, M., and Hulshof, C. T. J. (1998), "An Updated Review of Epidemiologic Studies on the Relationship between Exposure to Whole-Body Vibration and Low Back Pain," *Journal of Sound and Vibration*, Vol. 215, No. 4, pp. 595–611.

Brammer, A. J., Taylor, W., and Lundborg, G. (1987), "Sensorineural Stages of the Hand-Arm Vibration Syndrome," *Scandinavian Journal of Work, Environment and Health*, Vol. 13, No. 4, pp. 279–283.

British Standards Institution (BSI) (1987), "Measurement and Evaluation of Human Exposure to Whole-Body Mechanical Vibration and Repeated Shock," BS 6841, BSI, London.

British Standards Institution (BSI) (2008), "Guide to the Evaluation of Human Exposure to Vibration in Buildings. Part 1: Vibration Sources Other Than Blasting," BS 6472-1, BSI, London.

Corbridge, C., and Griffin, M. J. (1991), "Effects of Vertical Vibration on Passenger Activities: Writing and Drinking," *Ergonomics*, Vol. 34, No. 10, pp. 1313–1332.

Corbridge, C., Griffin, M. J., and Harborough, P. (1989), "Seat Dynamics and Passenger Comfort," *Proceedings of the Institution of Mechanical Engineers*, Vol. 203, pp. 57–64.

Dupuis, H., and Zerlett, G. (1986), *The Effects of Whole-Body Vibration*, Springer-Verlag, Berlin.

European Parliament and the Council of the European Union (2002), "On the Minimum Health and Safety Requirements Regarding the Exposure of Workers to the Risks Arising from Physical Agents (Vibration)," Directive 2002/44/EC, *Official Journal of the European Communities*, L177/13-19, July 6, 2002.

European Parliament and the Council of the European Union (2006), "Directive of the European Parliament and of the Council on Machinery, and Amending Directive 95/16/EC," *Official Journal of the European Communities*, L157, pp. 24–86.

Fairley, T. E., and Griffin, M. J. (1989), "The Apparent Mass of the Seated Human Body: Vertical Vibration," *Journal of Biomechanics*, Vol. 22, No. 2, pp. 81–94.

Gemne, G., Pyykko, I., Taylor, W., and Pelmear, P. (1987), "The Stockholm Workshop Scale for the Classification of Cold-Induced Raynaud's Phenomenon in the Hand-Arm Vibration Syndrome (Revision of the Taylor-Pelmear Scale)," *Scandinavian Journal of Work*, *Environment and Health*, Vol. 13, No. 4, pp. 275–278.

Griffin, M. J. (1976), "Eye Motion during Whole-Body Vertical Vibration," *Human Factors*, Vol. 18, No. 6, pp. 601–606.

Griffin, M. J. (1990), *Handbook of Human Vibration*, Academic, London.

Griffin, M. J. (1991), "Physical Characteristics of Stimuli Provoking Motion Sickness," Paper 3, in *Motion Sickness: Significance In Aerospace Operations And Prophylaxis*, AGARD Lecture Series LS-175, NATO Research and Technology Organisation, ISBN 92-835-0634-0.

Griffin, M. J. (1997), "Measurement, Evaluation, and Assessment of Occupational Exposures to Hand-Transmitted Vibration," *Occupational and Environmental Medicine*, Vol. 54, No. 2, pp. 73–89.

Griffin, M. J. (1998a), "Evaluating the Effectiveness of Gloves in Reducing the Hazards of Hand-Transmitted Vibration," *Occupational and Environmental Medicine*, Vol. 55, No. 5, pp. 340–348.

Griffin, M. J. (1998b), "A Comparison of Standardized Methods for Predicting the Hazards of Whole-Body Vibration and Repeated Shocks," *Journal of Sound and Vibration*, Vol. 215, No. 4, pp. 883–914.

Griffin, M. J. (2004), "Minimum Health and Safety Requirements for Workers Exposed to Hand-Transmitted Vibration and Whole-Body Vibration in the European Union; a Review," *Occupational and Environmental Medicine*, Vol. 61, pp. 387–397.

Griffin, M. J., and Bovenzi, M. (2002), "The Diagnosis of Disorders Caused by Hand-Transmitted Vibration: Southampton Workshop 2000," *International Archives of Occupational and Environmental Health*, Vol. 75, (1–2), pp. 1–5.

Griffin, M. J., and Hayward, R. A. (1994), "Effects of Horizontal Whole-Body Vibration on Reading," *Applied Ergonomics*, Vol. 25, No. 3, pp. 165–169.

Griffin, M. J., and Newman, M. M. (2004), "Effects of the Visual Field on Motion Sickness in Cars," *Aviation, Space and Environmental Medicine*, Vol. 75, pp. 739–748.

Griffin, M. J., Bovenzi, M., and Nelson, C. M. (2003), "Dose Response Patterns for Vibration-Induced White Finger," *Occupational and Environmental Medicine*, Vol. 60, pp. 16–26.

International Organization for Standardization (ISO) (1974), Guide For The Evaluation Of Human Exposure To Whole-Body Vibration, *International Standard*, ISO 2631 (E).

International Organization for Standardization (ISO) (1985), "Evaluation of Human Exposure to Whole-Body Vibration_Part 1: General Requirements," International standard 2631/1, ISO, Geneva.

International Organization for Standardization (ISO) (1988), "Hand-Held Portable Tools_Measurement of Vibration at the Handle_Part 1: General," ISO 8662-1, ISO, Geneva.

International Organization for Standardization (ISO) (1996), "Mechanical Vibration and Shock_Hand-Arm Vibration_Method for the Measurement and Evaluation of the Vibration Transmissibility of Gloves at the Palm of the Hand," ISO 10819, ISO, Geneva.

International Organization for Standardization (ISO) (1997), "Mechanical Vibration and Shock_Evaluation of Human Exposure to Whole-Body Vibration. Part 1: General Requirements," ISO 2631-1, ISO, Geneva.

International Organization for Standardization (ISO) (2001), "Mechanical Vibration_Measurement and Evaluation of Human Exposure to Hand-Transmitted Vibration_Part 1: General Requirements," ISO 5349-1:2001(E), ISO, Geneva.

International Organization for Standardization (ISO) (2002), "Mechanical Vibration_Measurement and Evaluation of Human Exposure to Hand-Transmitted Vibration_Part 2: Practical Guidance for Measurement at the Workplace," ISO 5349-2:2001 E, ISO, Geneva.

International Organization for Standardization (ISO) (2003), "Mechanical Vibration and Shock_Evaluation of Human Exposure to Whole-Body Vibration_Part 2: Vibration in Buildings (1 Hz to 80 Hz)," ISO 2631-2, ISO, Geneva.

International Organization for Standardization (ISO) (2008), "Acoustics_Preferred Reference Quantities for Acoustic Levels," ISO 1683, ISO, Geneva.

International Organization for Standardization (ISO) (2009), "Hand-Held Portable Power Tools_Test Methods for Evaluation of Vibration of Vibration Emission _ Part 1: Angle and Vertical Grinders" ISO 28927-1, ISO, Geneva.

Lawther, A., and Griffin, M. J. (1987), "Prediction of the Incidence of Motion Sickness from the Magnitude, Frequency, and Duration of Vertical Oscillation," *Journal of the Acoustical Society of America*, Vol. 82, No. 3, pp. 957–966.

Lawther, A., and Griffin, M. J. (1988a), "A Survey of the Occurrence of Motion Sickness amongst Passengers at Sea," *Aviation, Space and Environmental Medicine*, Vol. 59, No. 5, pp. 399–406.

Lawther, A., and Griffin, M. J. (1988b), "Motion Sickness and Motion Characteristics of Vessels at Sea," *Ergonomics*, Vol. 31, No. 10, 1373–1394.

Levison, W. H., and Harrah, C. B. (1977), "Biomechanical and Performance Response of Man in Six Different Directional Axis Vibration Environments," AMRL-TR-77-71, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, OH.

Lewis, C. H., and Griffin, M. J. (1977), "The Interaction of Control Gain and Vibration with Continuous Manual Control Performance," *Journal of Sound and Vibration*, Vol. 55, No. 4, pp. 553–562.

Lewis, C. H., and Griffin, M. J. (1979), "The Effect of Character Size on the Legibility of Numeric Displays during Vertical Whole-Body Vibration," *Journal of Sound and Vibration*, Vol. 67, No. 4, pp. 562–565.

Lewis, C. H., and Griffin, M. J. (2002), "Evaluating Vibration Isolation of Soft Seat Cushions Using an Active Anthropodynamic Dummy," *Journal of Sound and Vibration*, Vol. 253, No. 1, pp. 295–311.

Loriga, G. (1911), "Il Lavoro con i Martelli Pneumatici [The Use of Pneumatic Hammers]", *Bollettino del Ispettorato del Lavoro*, Vol. 2, pp. 35–60.

Matsumoto, Y., and Griffin, M. J. (2000), "Comparison of Biodynamic Responses in Standing and Seated Human Bodies," *Journal of Sound of Vibration*, Vol. 238, No. 4, pp. 691–704.

Matsumoto, Y., and Griffin, M. J. (2001), "Modelling the Dynamic Mechanisms Associated with the Principal Resonance of the Seated Human Body," *Clinical Biomechanics*, Vol. 16, Supplement No. 1, pp. S31–S44.

McLeod, R. W., and Griffin, M. J. (1989), "A Review of the Effects of Translational Whole-Body Vibration on Continuous Manual Control Performance," *Journal of Sound and Vibration*, Vol. 133, No. 1, pp. 55–115.

McLeod, R. W., and Griffin, M. J. (1990), "Effects of Whole-Body Vibration Waveform and Display Collimation on the Performance of a Complex Manual Control Task," *Aviation, Space and Environmental Medicine*, Vol. 61, No. 3, pp. 211–219.

Meddick, R. D. L., and Griffin, M. J. (1976), "The Effect of Two-Axis Vibration on the Legibility of Reading Material," *Ergonomics*, Vol. 19, No. 1, 21–33.

Morioka, M., and Griffin, M. J. (2006a), "Magnitude-Dependence of Equivalent Comfort Contours for Fore-and-Aft, Lateral and Vertical Whole-Body Vibration," *Journal of Sound and Vibration*, Vol. 298, No. 3, pp. 755–772.

Morioka, M., and Griffin, M. J. (2006b), "Magnitude-Dependence of Equivalent Comfort Contours for Fore-and-Aft, Lateral and Vertical Hand-Transmitted Vibration," *Journal of Sound and Vibration*, Vol. 295, pp. 633–648.

Morioka, M., and Griffin, M. J. (2008), "Absolute Thresholds for the Perception of Fore-and-Aft, lateral, and Vertical Vibration at the Hand, the Seat, and the Foot," *Journal of Sound and Vibration*, Vol. 314, pp. 357–370.

Moseley, M. J., and Griffin, M. J. (1986a), "A Design Guide for Visual Displays and Manual Tasks in Vibration Environments. Part I: Visual Displays," Technical Report No. 133, Institute of Sound and Vibration Research, University of Southampton.

Moseley, M. J., and Griffin, M. J. (1986b), "Effects of Display Vibration and Whole-Body Vibration on Visual Performance," *Ergonomics*, Vol. 29, No. 8, pp. 977–983.

Moseley, M. J., Lewis, C. H., and Griffin, M. J. (1982), "Sinusoidal and Random Whole-Body Vibration: Comparative Effects on Visual Performance," *Aviation, Space and Environmental* Medicine, Vol. 53, No. 10, pp. 1000–1005.

National Institute of Occupational Safety and Health, (NIOSH), Bernard, B. P. (Ed.) (1997), "Musculoskeletal Disorders and Workplace Factors: A Critical Review of Epidemiologic Evidence for Work-Related Disorders of the Neck, Upper Extremities, and Low Back." DHHS (NIOSH) Publication No. 97-141, U.S. Department of Health and Human Services, NIOSH, Washington, DC.

Nawayseh, N., and Griffin, M. J. (2003), "Non-Linear Dual-Axis Biodynamic Response to Vertical Whole-Body Vibration," *Journal of Sound and Vibration*, Vol. 268, pp. 503–523.

Nixon, C. W., and Sommer, H. C. (1963), "Influence of Selected Vibrations upon Speech (Range Of 2 Cps-20 Cps and Random)," AMRL-TDR-63-49, Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, OH.

O'Hanlon, J. G., and Griffin, M. J. (1971), "Some Effects of the Vibration of Reading Material upon Visual Performance," Technical Report No. 49, Institute of Sound and Vibration Research, University of Southampton, Southampton, UK.

Paddan, G. S., and Griffin, M. J. (1988a), "The Transmission of Translational Seat Vibration to the Head_1. Vertical Seat Vibration," *Journal of Biomechanics*, Vol. 21, No. 3, pp. 191–197.

Paddan, G. S., and Griffin, M. J. (1988b), "The Transmission of Translational Seat Vibration to the Head_II. Horizontal Seat Vibration," *Journal of Biomechanics*, Vol. 21, No. 3, 199–206.

Reason, J. T. (1970, June), "Motion Sickness: A Special Case of Sensory Rearrangement," *Advancement of Science*, Vol. 26, pp. 386–393.

Reason, J. T. (1978), "Motion Sickness Adaptation: A Neural Mismatch Model," *Journal of the Royal Society of Medicine*, Vol. 71, pp. 819–829.

Reason, J. T., and Brand, J. J. (1975), *Motion Sickness*. Academic, London.

Sherwood, N., and Griffin, M. J. (1990), "Effects of Whole-Body Vibration on Short-Term Memory," *Aviation, Space and Environmental Medicine*, Vol. 61, No. 12, pp. 1092–1097.

Sherwood, N., and Griffin, M. J. (1992), "Evidence of Impaired Learning during Whole-Body Vibration," *Journal of Sound and Vibration*, Vol. 152, No. 2, pp. 219–225.

Shoenberger, R. W. (1974), "An Investigation of Human Information Processing during Whole-Body Vibration," *Aerospace Medicine*, Vol. 45, No. 2, pp. 143–153.

Shoenberger, R. W., and Wilburn, D. L. (1973), "Tracking Performance during Whole-Body Vibration with Side-Mounted and Centre-Mounted Control Sticks," AMRL-TR-72-120, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, OH.

Taylor, W., and Pelmear, P. L. (Editors) (1975), *Vibration White Finger in Industry*. Academic Press, London.

Torle, G. (1965), "Tracking Performance under Random Acceleration: Effects of Control Dynamics," *Ergonomics*, Vol. 8, No. 4, pp. 481–486.

Turner, M., and Griffin, M. J. (1999), "Motion Sickness in Public Road Transport: The Relative Importance of Motion, Vision and Individual Differences," *British Journal of Psychology*, Vol. 90, pp. 519–530.

Wasserman, D., Taylor, W., Behrens, V., Samueloff, S., and Reynolds, D. (1982), "Vibration White Finger Disease in U.S. Workers Using Pneumatic Chipping and Grinding Handtools. I: Epidemiology," Technical Report, DHSS (NIOSH) Publication No. 82-118, U.S. Department of Health and Human Services, National Institute for Occupational Safety and Health, Washington, DC.

Wei, L., and Griffin, M. J. (1998a), "Mathematical Models for the Apparent Mass of the Seated Human Body Exposed to Vertical Vibration," *Journal of Sound and Vibration*, Vol. 212, No. 5, pp. 855–874.

Wei, L., and Griffin, M. J. (1998b), "The Prediction of Seat Transmissibility from Measures of Seat Impedance," *Journal of Sound and Vibration*, Vol. 214, No. 1, pp. 121–137.

Wells, M. J., and Griffin, M. J. (1984), "Benefits of Helmet-Mounted Display Image Stabilisation under Whole-Body Vibration," *Aviation, Space and Environmental Medicine*, Vol. 55, No. 1, pp. 13–18.

Wilson, R. V. (1974), "Display Collimation under Whole-Body Vibration," *Human Factors*, Vol. 16, No. 2, pp. 186–195.

# CHAPTER 23

# SOUND AND NOISE: MEASUREMENT AND DESIGN GUIDANCE*

**John G. Casali**
**Virginia Polytechnic Institute and State University**
**Blacksburg, Virginia**

## 1 INTRODUCTION

Sound along with its subset, noise, which is often defined as unwanted sound, is a phenomenon that confronts human factors professionals in many settings and applications. A few examples are (1) an auditory warning signal, for which the proper sound parameters must be selected for maximizing detection, identification, and localization; (2) a situation wherein the speech communication that is critical between operators is compromised in its intelligibility by environmental noise, and therefore redesign of the communications system and/or acoustic environment is needed; (3) a residential community is intruded upon by the noise from vehicular traffic or a nearby industrial plant, causing annoyance and sleep arousal and necessitating abatement; (4) an in-vehicle auditory display that warns of dangerous conditions must convey urgency and localization cues; (5) a worker is exposed to hazardous noise on the job, and to prevent hearing loss, an appropriate hearing protection device (HPD) must be selected; and (6) a soldier's ears must be protected from exposure to gunfire with an HPD, but at the same time, he or she must be able to detect enemy threat-related sounds. To deal effectively with examples of these types, the human factors engineer must understand the basics of sound, instrumentation, and techniques for its measurement and quantification, analyses of acoustic measurements for ascertaining the audibility of signals and speech as well as the risks to hearing, and countermeasures to combat the deleterious effects of noise. In this chapter these and related matters are addressed from a human factors engineering perspective while several important noise-related standards and regulations are also covered.

---

* Portions of this chapter are based in part on Casali and Gerges (2006) and Robinson and Casali (2003).

At the outset it should be noted that the science of acoustics and the study of sound and noise within it is very broad and comprises a vast body of research and standards literature. Thus, as the subject of a single chapter, this topic cannot be covered in great depth herein. It is therefore an intent of this chapter to introduce several major topics concerning sound/noise, particularly as it impacts humans, and to point the reader to other publications for detail on specific topics. As for the area of sound/noise as a whole, three excellent, broad coverage texts are Kryter (1994), Crocker (1998), and Berger et al. (2003).

## 2 SOUND AND NOISE

Most aspects of acoustics rely on accurate quantification and evaluation of the sound itself; therefore, a basic understanding of sound parameters and sound measurement is needed before delving into application-oriented issues.

### 2.1 Fundamental Parameters

*Sound* is a disturbance in a medium (in industry, home, or recreational settings, most commonly air or a conductive structure such as a floor or wall) that has mass and elasticity. For example, an exhaust fan on the roof of an industrial plant has blades that rotate in the air, creating noise which may propagate into the surrounding community. Because the blades are coupled to the air medium, they produce pressure waves that consist of alternating compressions (above ambient air pressure) and rarefactions (below ambient pressure) of air molecules, the *frequency* ( $f$ ) of which is the number of above/below ambient pressure cycles per second, or *hertz* ( $Hz$ ). The reciprocal of frequency, $1/f$, is the *period* of the waveform. The waveform propagates outward from the fan as long as it continues to rotate, and the disturbance in air pressure that occurs in relation to ambient air pressure is heard as sound, in this case "fan roar." The linear distance traversed by the sound wave in one complete cycle of vibration is the *wavelength*:

$$\lambda = c/f \qquad (1)$$

Wavelength ( $\lambda$ in meters or feet) depends on the sound frequency ( $f$ in hertz) and velocity ( $c$ in meters per second or feet per second; in air at 68°F and pressure of 1 atmosphere (atm), 344 m/s or 1127 ft/s) in the medium. The speed of sound is influenced by the temperature of the medium, and in air it increases about 1.1 ft/s for each increase of 1°F.

*Vibrations* are oscillations in solid media and are often associated with the production of sound waves. *Noise* can be loosely defined as a subset of sound; that is, noise is sound that is undesirable or offensive in some aspect. However, the distinction is largely situation- and listener-specific, as perhaps best stated in the old adage "one person's music is another's noise."

Unlike some common ergonomics-related stressors such as repetitive motions or awkward lifting maneuvers, noise is a physical stimulus that is readily measurable and quantifiable using transducers (microphones)

and instrumentation [sound level meters (SLMs) and their variants] that are commonly available. Aural exposure to noise and the damage potential therefrom are functions of the *total energy* transmitted to the ear. In other words, the energy is equivalent to the product of the noise intensity and duration of the exposure. Several metrics that relate to the energy of the noise exposure have been developed, most with an eye toward accurately reflecting the exposures that occur in industrial or community settings. These metrics are covered in Section 3.2, but first, the most basic unit of measurement must be understood, the *decibel*.

### 2.2 Physical Quantification: Sound Levels and the Decibel Scale

The unit of *decibel*, one-tenth of a *bel*, is the most common metric applied to the quantification of noise amplitude. The decibel (dB) is a measure of *level*, defined as the logarithm of the ratio of a quantity to a reference quantity of the same type. In acoustics, it is applied to sound level, of which there are three types.

*Sound power level*, the most basic quantity, is typically expressed in decibels and is defined as

$$\text{Sound power level (dB)} = 10 \log_{10} \text{Pw}_1/\text{Pw}_r \qquad (2)$$

where $\text{Pw}_1$ is the acoustic power of the sound in watts or other power unit and $\text{Pw}_r$ is the acoustic power of a reference sound in watts, usually taken to be the acoustic power at hearing threshold for a young, healthy ear at the frequency of maximum sensitivity, the quantity $10^{-12}$ W.

*Sound intensity level*, following from power level, is typically expressed in decibels and is defined as

$$\text{Sound intensity level (dB)} = 10 \log_{10} I_1/I_r \qquad (3)$$

where $I_1$ is the acoustic intensity of the sound in watts per square meter or other intensity unit and $I_r$ is the acoustic intensity of a reference sound in watts per square meter, usually taken to be the acoustic intensity at hearing threshold, or the quantity $10^{-12}$ W/m². 

Within the last decade, sound measurement instruments to measure sound *intensity* level have become commonplace, albeit expensive and relatively complex. Sound *power* level, by contrast, is not directly measurable but can be computed from empirical measures of sound intensity level or sound pressure level. On the other hand, sound *pressure* level is directly measurable by using relatively straightforward instruments and is by far the most common metric used in practice.

*Sound pressure level* (SPL), abbreviated in formulas as $L_P$, is also typically expressed in decibels. Since power is directly proportional to the square of the pressure, SPL is defined as

$$\text{Sound pressure level (SPL or } L_P; \text{ dB)}$$
$$= 10 \log_{10} P_1^2/P_r^2 = 20 \log_{10} P_1/P_r \qquad (4)$$

where $P_1$ is the pressure level of the sound in micro-pascals (μPa) or other pressure unit and $P_r$ is the

pressure level of a reference sound in micropascals, usually taken to be the pressure at hearing threshold, or the quantity 20 μPa, or 0.00002 Pa. Other equivalent reference quantities are 0.0002 dyn/cm$^2$ and 20 μbars.

The application of the decibel scale to acoustic measurements yields a convenient means of collapsing the vast range of sound pressures which would be required to accommodate sounds that can be encountered into a more manageable, compact range. As shown in Figure 1, using the logarithmic compression produced by the decibel scale, the range of typical sounds from human hearing threshold to the threshold of tactile "feeling" is 120 dB, while the linear pressure scale applied to the same range of sounds produces a range of 1,000,000 Pa. Of course, sounds *do* occur that are higher than 120 dB (e.g., artillery fire) or lower than 0 dB (below normal threshold on an audiometer). A comparison of decibel values of example sounds to their pressure values (in pascals) is also depicted in Figure 1.

In considering changes in sound level measured in decibels, a few numerical relationships emanating from the decibel formulas above are often helpful in practice. An increase (decrease) in SPL by 6 dB is equivalent to a doubling (halving) of the sound pressure. Similarly, on the power or intensity scales, an increase (decrease) of 3 dB is equivalent to a doubling (halving) of the sound power or intensity. The latter relationship gives

rise to what is known as the *equal-energy rule* or *trading relationship*. Because sound represents energy which is itself a product of intensity and duration, an original sound that increases (decreases) by 3 dB is equivalent in total energy to the same original sound that does not change in decibel value but decreases (increases) in its duration by half (twice).

## 2.3 Basic Computations with Decibels

There are many practical instances in which it is helpful to predict the combined result of several individual sound sources that have been measured separately in decibels. This can be performed for random, uncorrelated sound sources using the equation

$$L_{P\text{combo}}(\text{dB})$$
$$= 10 \, \log_{10} \left( 10^{L_{P1}/10} + 10^{L_{P2}/10} + \cdots + 10^{L_{Pn}/10} \right) \tag{5}$$

and it applies for any decibel weighting (dBA, dBC, etc., as explained later) or for any bandwidth (such as one-third octave, full octave, etc.). For example, suppose that an industrial plant currently exposes workers in a work area to a time-weighted average (TWA) of 83.0 dBA, which is below the Occupational Safety and Health Administration (OSHA, 1983) action level (85.0 dBA)



**Sound Pressure Level (dB) re 20 μPa**
(*logarithmic scale*)
Range: Factor of 120 dB

| | dB | Pa | |
|---|---|---|---|
| Jackhammer, operator's position | 120 | 20 | |
| Chainsaw, cutting, operator's position | 110 | 10 | |
| Drag race car, unmuffled, 100 ft | 100 | 5 / 2 | DANGER |
| Power mower, muffled, operator position | 90 | 1 / 0.5 | |
| Electric razor, at ear | 80 | 0.2 | RISK |
| Vacuum cleaner, operator's position | 70 | 0.1 / 0.05 | |
| Conversation, 3 ft apart | 60 | 0.02 | ANNOYANCE |
| Computer fan, operator's position | 50 | 0.01 / 0.005 | |
| Quiet bedroom | 40 | 0.002 | |
| Recording studio | 30 | 0.001 / 0.0005 | |
| Very soft whisper, at ear | 20 | 0.0002 | |
| Wristwatch, ticking, at arm's length | 10 | 0.0001 / 0.00005 | |
| Threshold of hearing for healthy ear | 0 | 0.00002 | |

**Sound Pressure (Pa)**
(*linear scale*)
Range: Factor of 1,000,000 Pa

**Figure 1**   Sound pressure level in decibels and sound pressure in pascals for typical sounds.

at which a hearing conservation program would be required by law. Two new pieces of equipment are proposed for purchase and installation in this area: a new single-speed conveyor that has a constant noise output of 78.0 dBA and a new compressor that has a constant output of 82.5 dBA. The combined sound level will be approximately

$$
\begin{aligned}
L_{P\,\text{combo}} & (\text{dB}) \\
& = 10\,\log_{10}\left(10^{83.0/10} + 10^{78.0/10} + 10^{82.5/10}\right) \\
& = 86.4\,\text{dBA}
\end{aligned}
$$

Thus, by purchasing this conveyor, the plant would move from a noise exposure level (83.0 dBA) that is in compliance with OSHA to one that is not (86.4 dBA). This is one illustration why industries should adhere to a "buy quiet" policy, so that noise exposure problems are not created unknowingly by equipment purchases.

Subtraction of decibels works in the same manner as addition:

$$
L_{P\,\text{difference}}(\text{dB}) = 10\,\log_{10}\left(10^{L_{P1}/10} - 10^{L_{P2}/10}\right) \quad (6)
$$

Using the example above, if the compressor were eliminated from the situation, the overall combined noise level would be the combination of the three sources as computed to be 86.4 dBA, reduced by the absence of the compressor at 82.5 dBA:

$$
\begin{aligned}
L_{P\,\text{difference}}\,(\text{dB}) & = 10\,\log_{10}\left(10^{86.4/10} - 10^{82.5/10}\right) \\
& = 84.1\,\text{dBA}
\end{aligned}
$$

With this result, the plant area noise level moves back into OSHA compliance under the action level of 85.0 dBA, but just by about 1.0 dBA. To err on the safe side, especially to accommodate the potential of any upward fluctuations in noise level, this plant's management should still look to reduce the noise further or install a hearing conservation program.

There are a few rules of thumb that arise from the computations shown above. One is that when two sound sources are approximately equivalent in SPL, the combination of the two will be about 3 dB larger than the decibel level of the *higher* source. Another is that as the difference between two sounds exceeds about 13 dB, the contribution of the lower level sound to the combined sound level is negligible (i.e., about 0.2 dB). In relation to this, when it is desirable to measure a sound of interest in isolation but it cannot be physically separated from a background noise, the question becomes: To what extent is the background noise influencing the accuracy of the measurement? In many cases, such as in some manufacturing plants, the background noise cannot be turned off but the sound of interest can. If this is the case, then the sound of interest is measured in the background noise, and then the background noise is measured alone. If the background noise measurement differs from the combined measurement by more than 13 dB, then it has not influenced the measurement of the sound of interest in a significant manner. If the difference is

smaller, then equation (6) can be applied to correct the measurement, effectively by removing the background noise's contribution. Some standards use a difference of 10 dB as a guideline for when to apply the background noise correction.

Finally, it is important to recognize that due to the limits in precision and reliability of decibel measurements, for the applications discussed in this chapter (and most others in acoustics as well), it is unnecessary to record decibel calculations that result from the formulas herein to greater than one decimal point, and it is usually sufficient to round final results to the nearest 0.5 dB or even to integer values. However, to avoid interim rounding error, it is important to carry the significant figures through each step of the formulas until the end result is obtained (Ostergaard, 2003).

## 3 MEASUREMENT AND QUANTIFICATION OF SOUND AND NOISE EXPOSURES

### 3.1 Basic Instrumentation

Measurement and quantification of sound levels and noise exposure levels provide the fundamental data for assessing hearing exposure risk, speech and signal-masking effects, hearing conservation program needs, and engineering noise control strategies. A vast array of instrumentation is available; however, for most of the aforementioned applications, a basic understanding of three primary instruments (SLMs, dosimeters, and real-time spectrum analyzers) and their data output will suffice. In instances where noise is highly impulsive in nature, such as gunfire, and/or development of situation-specific engineering noise control solutions is anticipated, more specialized instruments may be necessary.

Because sound is propagated as pressure waves that vary over space and in time, a complete acoustic record of a noise exposure or a sound event that has a prolonged duration requires simultaneous measurements at all points of interest in the sound field. This measurement should occur over a representative, continuous time period to document the noise level exhaustively in the space. Obviously, this is typically cost- and time-prohibitive, so one must resort to sampling strategies for establishing the observation points and intervals. The analyst must also decide whether detailed, discrete-time histories with averaging over time and space are needed (such as with a noise-logging *dosimeter*), if discrete samples taken with a short-duration moving time average (with a basic *sound level meter*) will suffice, or if frequency-band-specific SPLs are needed for selecting noise abatement materials (with a *spectrum analyzer*). A discussion of these three primary types of sound measurement instruments and the noise descriptors that can be obtained therefrom follows.

### 3.1.1 Sound Level Meter

Most sound measurement instruments derive from the basic SLM, a device for which there are four grades and associated performance tolerances that become more stringent as the grade number *decreases*, described by

American National Standards Institute (ANSI) S1.4-1983(R2006) (ANSI, 2006). Type 0 instruments have the most stringent tolerances and are for laboratory use only. Other grades include type 1, intended for precision measurement in the field or laboratory; type 2, intended for general field use, especially where frequencies above 10,000 Hz are not prevalent; and type S, a special-purpose meter that may perform at grade 1, 2, or 3 but may not include all of the operational functions of the particular grade. A grade of type 2 or better is needed for measuring occupational exposures and community noise and to obtain data for most court proceedings. For SLMs which provide time-averaged or integrated SPLs or, optionally, sound exposure levels for OSHA or other noise-monitoring requirements, ANSI S1.43-1997(R2007) should be consulted (ANSI, 2007a). This standard specifies SLM characteristics that are essential to the accurate measurement of steady, intermittent, fluctuating, and impulsive sounds, particularly when the measurement obtained is over a time interval as opposed to instantaneously.

A block diagram of the functional components of a generic SLM appears in Figure 2. At the top, a microphone/preamplifier senses the pressure changes caused by an airborne sound wave and converts the pressure signal into a voltage signal. Because the pressure fluctuations of a sound wave are small in magnitude, the corresponding voltage signal must be preamplified and then input to an amplifier, which boosts the signal before it is processed further. The passband, the range of frequencies that are passed through and processed, of a

high-quality SLM contains frequencies from about 10 to 20,000 Hz, but depending on the frequency weighting used, not all frequencies are treated in the same way. A selectable frequency-weighting network, or filter, is then applied to the signal. These networks most commonly include the A-, B-, and C-weighting functions shown in Figure 3b. For OSHA noise-monitoring measurements and for many community noise applications, the A scale, which deemphasizes the low frequencies and to a smaller extent the high frequencies, is used. In addition to the common A scale (which approximates the 40-phon level of hearing) and C scale (100-phon level), other selections may be available. If no weighting function is selected on the meter, the notation dBZ or dB(linear) is used, and all frequencies are processed without weighting factors. The actual weighting functions for the three suffix notations A, B, and C are superimposed on the phon contours of Figure 3a and are also depicted in Figure 3b as actual frequency-weighting functions.

Next (not shown), the signal is squared to reflect the fact that SPL in decibels is a function of the square of the sound pressure. The signal is then applied to an exponential averaging network, which defines the meter's dynamic response characteristics. In effect, this response creates a moving-window, short-time average display of the sound waveform. The two most common settings are FAST, which has a time constant of 0.125 s, and SLOW, which has a time constant of 1.0 s. These time constants were established decades ago to give analog needle indicators a rather sluggish response (particularly on the SLOW setting) so that they could be read by the human eye even when highly fluctuating sound pressures were measured. Under the FAST or SLOW dynamics, the meter indicator rises exponentially toward the decibel value of an applied constant SPL. For OSHA measurements, the SLOW setting is used, and this setting is also best when the average value (as it is changing over time) is desired. The FAST setting is more appropriate when the variability or range of fluctuations of a time-varying sound is desired. On certain SLMs, a third time constant, IMPULSE, may also be included for measurement of sounds that have sharp transient characteristics over time and are generally less than 1 s in duration, exemplified by gunshots or impact machinery such as drop forges. The IMPULSE setting has an exponential rise time constant of 35 ms and a decay time of 1.5 s. It is useful to afford the observer the time to view the maximum value of a burst of sound before it decays and is more commonly applied in community and business machine noise measurements than in industrial settings.

Because sound often consists of symmetrical pressure fluctuations above and below ambient air pressure for which the arithmetic average is zero, a *root mean square* (rms) averaging procedure is applied when FAST, SLOW, or IMPULSE measurements are taken, and the result is displayed in decibels. In effect, each pressure (or converted voltage) value is squared, the arithmetic mean of all squared values is then obtained, and finally the square root of the mean is computed to provide the rms value.
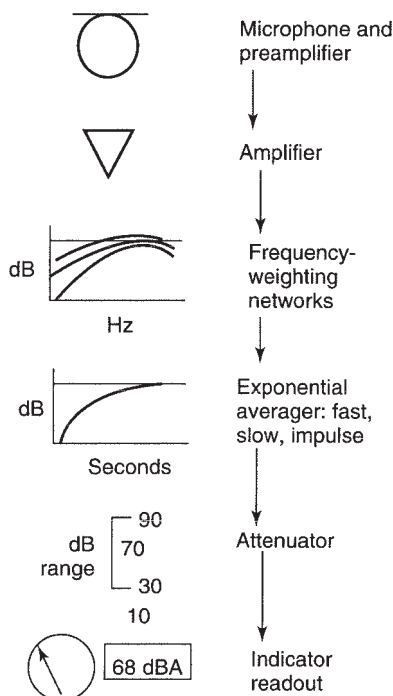


**Figure 2**  Functional components of a sound level meter.
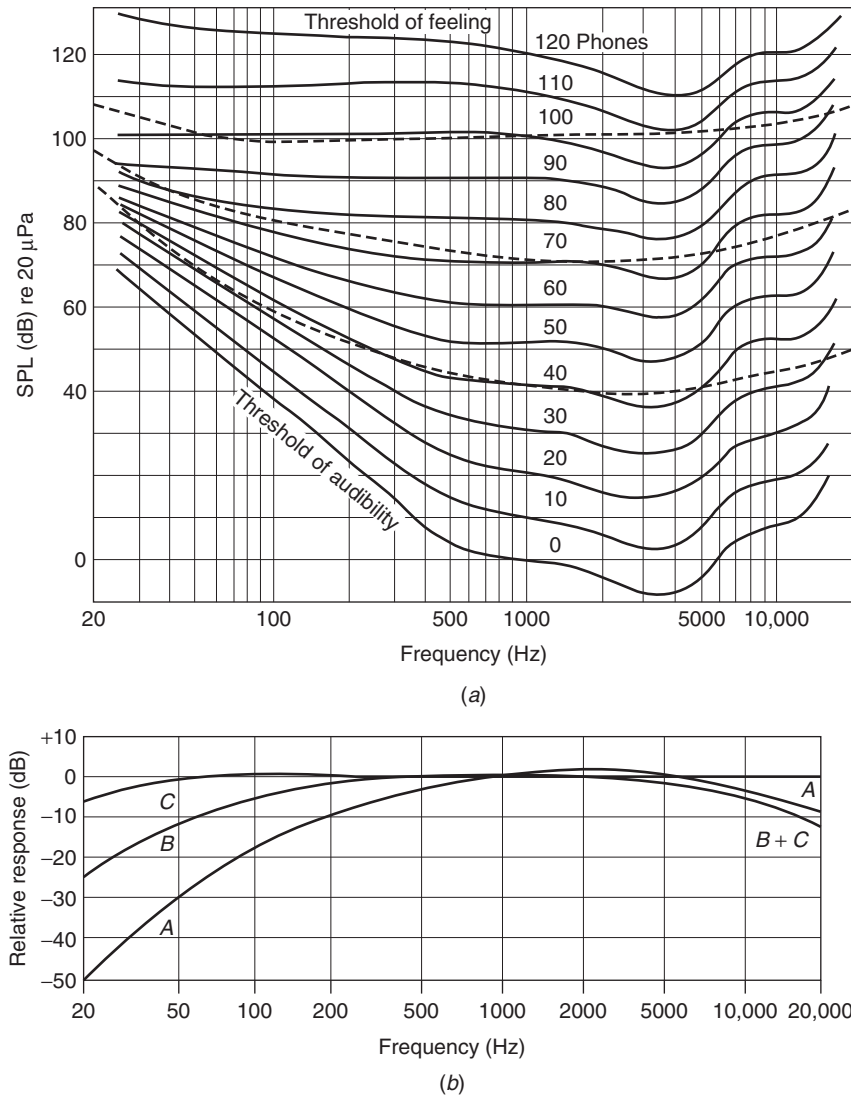
**Figure 3** (*a*) Equal-loudness contours based on the psychophysical phon scale, with sound level meter frequency-weighting curves superimposed; (*b*) decibels vs. frequency values of A, B, and C sound level meter weighting curves. (Adapted with permission from Earshen, 1986.)

Some SLMs include an unweighted PEAK setting that does not utilize the rms computation but instead provides an indication of the actual peak SPL reached during a pressure impulse. This measurement mode is necessary for certain applications: for instance, to determine if the OSHA limit of 140 dB for impulsive exposure is exceeded. A type 1 or 2 meter must be capable of measuring a 50-μs pulse. It is important to note that the aforementioned rms-based IMPULSE dynamics setting is unsuitable for measurement of PEAK SPLs.

With regard to the final component of a SLM shown in Figure 2, the indicator display or readout, much debate has existed over whether an analog (needle pointer or bar "thermometer-type" linear display) or

digital (numeric) display is best. Ergonomics research indicates that, although the digital readout affords higher precision of information to be presented in a smaller space, its disadvantage is that the least significant digit becomes impossible to read when the sound level is fluctuating rapidly. Also, it is more difficult with a digital readout for the observer to capture the maximum and minimum values of a sound, as is often desirable using the FAST or IMPULSE response. On the other hand, if very precise measurements down to a fraction of a decibel are needed, the digital indicator is preferable as long as the meter incorporates an appropriate time integrating/averaging feature or "hold" setting so that the data values can be captured. Because of the advantages and

disadvantages of each type of display, some contemporary SLMs include both analog and digital readouts.

***Microphone Considerations***   Most SLMs have interchangeable microphones that offer varying frequency response, sensitivity, and directivity characteristics (Peterson, 1979). The *response* of the microphone is the ratio of electrical output (in volts) to the sound pressure at the diaphragm of the microphone. Sound pressure is commonly expressed in pascals for free-field conditions (where there are no sound reflections resulting in reverberation), and the free-field voltage response of the microphone is given as millivolts per pascal. When specifications for *sensitivity* or *output level* are given, the response is usually based on a pure-tone sound wave input. Typically, the output level is provided in decibels re 1 V at the microphone electrical terminals, and the reference sensitivity is 1 V/Pa.

Most microphones that are intended for general sound measurements are essentially *omnidirectional* (i.e., nondirectional) in their response for frequencies below about 1000 Hz. The 360° response pattern of a microphone is called its *polar response*, and the pattern is generally symmetrical about the axis perpendicular to the diaphragm. Some microphones are designed to be highly directional, of which one example is the cardiode design, which has a heart-shaped polar response wherein the maximum sensitivity is for sounds whose direction of travel causes them to enter the microphone at 0° (or the *perpendicular incidence response*), and minimum sensitivity is for sounds entering at 180° behind the microphone. The response at 90°, where sound waves travel and enter parallel to the diaphragm, is known as the *grazing incidence response*. Another response pattern, the *random-incidence response*, represents the mean response of the microphone for sound waves that strike the diaphragm from all angles with equal probability. This response characteristic is the most versatile, and thus it is the response pattern used most often in the United States. Frequency responses for various microphone incidence patterns are depicted in Figure 4.

Because most U.S. SLM microphones are omnidirectional and utilize the random-incidence response, it is best for an observer to point the microphone at the primary noise source and hold it at an angle of incidence from the source at approximately 70°. This will produce a measurement most closely corresponding to the random-incidence response. On the other hand, free-field microphones have their flattest response at normal incidence (0°), while pressure microphones have their flattest response at grazing incidence (90°), and both should be pointed accordingly with respect to the noise source. Care must be taken to avoid shielding the microphone with the body or other structures. The response of microphones can also vary with temperature, atmospheric pressure, and humidity, with temperature usually being the most critical factor. Most microphone manufacturers supply correction factors for variations in decibel readout due to temperature effects. Atmospheric effects are generally significant only when measurements are made in aircraft or at very high altitudes, and humidity has a negligible effect except at very high levels. In any case, microphones must not be exposed
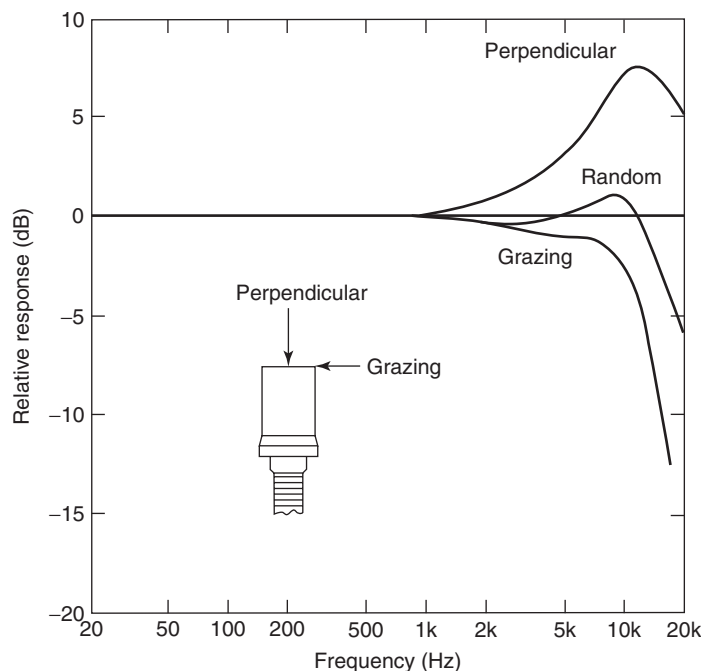


**Figure 4**   Frequency response of a hypothetical microphone for three angles of incidence. (Adapted with permission from Peterson, 1979.)

to moisture or large magnetic fields, such as those produced by transformers. When used in windy conditions, a foam windscreen should be placed over the microphone. This will reduce the contaminating effects of wind noise while influencing the frequency response of the microphone only slightly at high frequencies. The windscreen offers the additional benefit of protection of the microphone from damage due to being struck and/or from airborne foreign matter.

**Sound Level Meter Applications** It is important to note that the basic SLM is intended to measure sound levels at a given moment in time, although certain specialized devices can perform integration or averaging of sound levels over an extended period of time. When the nonintegrating/nonaveraging SLM is used for long-term noise measurements, such as over a workday, it is necessary to sample and make multiple manual data entries on a record to characterize the exposure. Being difficult, both in terms of reading the meter and recording sound level data, this technique is usually best limited to area measurements and is not applied to an individual's exposure sampling. Furthermore, the sampling process becomes more difficult as the fluctuations in a noise become more rapid and/or random in nature. SLMs are useful for determining the levels of human speech in both rms or peak values, calibration of laboratory experiments, calibration of audiometers (with special attachments), and community noise event-related measurements.

### 3.1.2 Dosimeter

The *audio-dosimeter* is a portable battery-powered device that is derived directly from a SLM but also features the ability to obtain special measures of noise exposure (discussed later) that relate to regulatory compliance and hearing hazard risk. Some versions are weather resistant and can be used outdoors to log a record of noise in a community setting, including both event-related, short-term measures and long-term averages and other statistical data.

Dosimeters for industrial use are very compact and are generally worn on the belt or in the pocket of an employee, with the microphone generally clipped to the lapel or shoulder of a shirt or blouse. The intent is to obtain a noise exposure history over the course of a full or partial work shift and to obtain, at a minimum, a readout of the TWA exposure and noise dose for the period measured. Depending on the features, the dosimeter may provide a running histogram of noise levels on a short-time-interval (such as 1-min) basis, compute statistical distributions of the noise exposures for the period, flag and record exposures that exceed OSHA maxima of 115 dBA continuous or 140 dB PEAK, and compute average metrics using 3 dB, 5 dB, or even other time-versus-level exchange rates. The dosimeter eliminates the need for the observer to set up a discrete sampling scheme, follow a noise-exposed worker, or monitor continuously an instrument that is staged outdoors, all of which are necessary with a conventional SLM. Dosimeters are special versions of integrating/averaging SLMs, which are governed by ANSI S1.43-1997(R2007), as referenced in ANSI (2007a) herein.

### 3.1.3 Spectrum Analyzer

A *spectrum analyzer* is an advanced SLM which incorporates selective frequency-filtering capabilities to provide an analysis of the noise level as a function of frequency. In other words, the noise is broken down into its frequency components and a distribution of the noise energy in all measured frequency bands is available. Bands are delineated by upper and lower edge or cutoff frequencies and a center frequency. Different widths and types of filters are available, with the most common width being the *octave filter*, wherein the center frequencies of the filters are related by multiples of 2 (i.e., 31.5, 63, 125, 250,..., 4000, 8000, and 16,000 Hz), with the most common type being the center-frequency proportional, wherein the width of the filter depends on the center frequency (as in an octave filter set, in which the passband width equals the center frequency divided by $2^{1/2}$). The *octave band*, commonly called the 1/1-octave filter, has a center frequency (CF) that is equal to the geometric mean of the upper ($f_u$) and lower ($f_l$) cutoff frequencies. The formulas to compute the center frequency for the octave filter, as well as the band-edge frequencies, are

$$\text{CF} = (f_u f_l)^{1/2}$$
$$\text{Upper cutoff, } f_u = \text{CF} \cdot 2^{1/2} \qquad (7)$$
$$\text{Lower cutoff, } f_l = \text{CF}/2^{1/2}$$

More precise spectral resolution can be obtained with other center-frequency proportional filter sets with narrower bandwidths, the most common being the 1/3 octave, and with constant-percentage bandwidth filter sets, such as 1% or 2% filters. Note that in both types the filter bandwidth increases as the center frequency increases. Still other analyzers have constant-bandwidth filters, such as 20-Hz-wide bandwidths which are of constant width regardless of center frequency. Whereas in the past most spectrum analyzer filters have been analog devices with "skirts" or overshoots extending slightly beyond the cutoff frequencies, digital computer-based analyzers are now very common. These "computational" filters use fast Fourier transform (FFT) or other algorithms to compute sound level in a prespecified band of fixed resolution. FFT devices can be used to obtain very high resolutions of noise spectral characteristics using bandwidths as low as 1 Hz. However, in most measurement applications, a 1/1- or 1/3-octave analyzer will suffice unless the noise has considerable power in near-tonal components that must be isolated. One caution is in order: If a noise fluctuates in time and/or frequency, an integrating/averaging analyzer should be used to achieve good accuracy of measurements. It is important that the averaging period be long in comparison to the variability of the noise being sampled.

*Real-time analyzers* incorporate parallel banks of filters (not FFT driven) that can process all frequency bands simultaneously, and the signal output may be controlled by a SLOW, FAST, or other time constant setting, or it may be integrated or averaged over a fixed

time period to provide $L_{\text{OSHA}}$, $L_{\text{eq}}$, or other average-type measurements to be discussed later.

***Spectrum Analyzer Applications***   While occupational noise is monitored with a dosimeter or SLM for the purpose of OSHA noise exposure compliance (using A-weighted broadband measurement) or the assessment of hearing protection adequacy (using C-weighted broadband measurement), both of these applications can also be addressed (in some cases more accurately) with the use of spectral measurements of the noise level. For instance, the OSHA Occupational Noise Exposure Standard (OSHA, 1983) allows the use of octave-band measurements reduced to broadband dBA values to determine if noise exposures exceed dBA limits defined in Table G-9 of the standard. Furthermore, Appendix B of the standard concerns hearing protector adequacy and allows the use of an octave-band method for determining, on a spectral rather than a broadband basis, whether a hearing protector is adequate for a particular noise spectrum. It is also noteworthy that spectral analysis can help the hearing conservationist discriminate noises as to their hazard potential even though they may have similar A-weighted SPLs. This is illustrated in Figure 5, where both noises would be considered to be of equal hazard by the OSHA-required dBA measurements (since they both are 90 dBA), but the 1/3-octave analysis demonstrates that the lowermost noise is more hazardous, as evidenced by the heavy concentration of energy in the midrange and high frequencies.

One of the most important applications of the spectrum analyzer is to obtain data that will provide the basis for engineering noise control solutions. For instance, to select an absorption material for lining interior surfaces of a workplace, the spectral content of the noise must be known so that the appropriate density, porosity, and thickness of material may be selected.

Spectrum analyzers are also necessary for performing the frequency-specific measurements needed to predict either signal audibility or speech intelligibility in noisy situations, according to the techniques discussed in Section 7. Furthermore, they can be applied for calibration of signals for laboratory experiments and audiometers, for determining the frequency response and other quality-related metrics for systems designed for music and speech rendition, and for determining certain acoustic parameters of indoor spaces, such as reverberation time.

***dBC–dBA***   Lacking a spectrum analyzer, one can obtain a very rough indication of the dominant spectral content of a noise by using a SLM and taking measurements in both dBA and dBC for the same noise. If the dBC–dBA value is large, that is, about 5 dB or more, then it can be concluded that the noise has considerable low-frequency content. If, on the other hand the dBC–dBA value is negative, the noise clearly has strong midrange components, since the A-weighting curve exhibits slight amplification in the range 2000–4000 Hz. Such rules of thumb rely on the differences in the C- and A- weighting curves shown in Figure 3*b*. However, they should not be relied upon in lieu of a spectrum analysis if the noise is believed to have high-frequency or narrowband components that need noise control attention.

### 3.1.4   Acoustic Calibrators

Each of the instruments described above contains a microphone that transduces the changes in pressure and inputs this signal into the electronics. Although modern sound measurement equipment is generally stable and reliable, calibration is necessary to match the microphone to the instrument so that the accuracy of the measurement is assured. Because of its susceptibility to varying environmental conditions and damage due
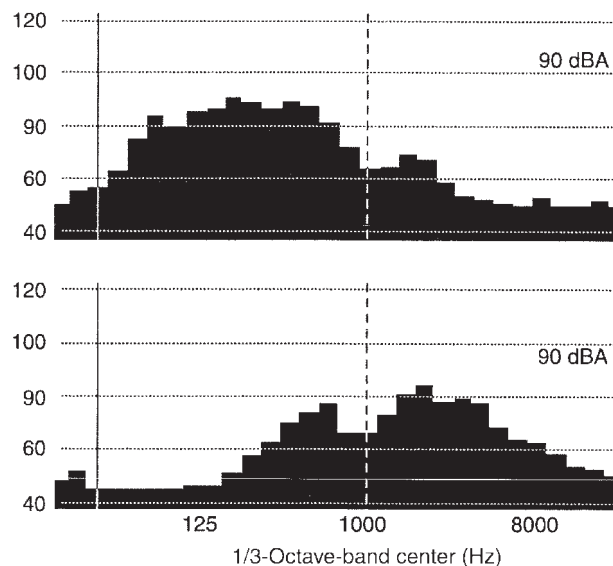


**Figure 5**   Spectral differences for two different noises that have the same dBA value.

to rough handling, moisture, and magnetic fields, the microphone is generally the weakest link in the measurement equipment chain. Therefore, an acoustic calibrator should be applied before and after each measurement with a SLM. The pretest calibration ensures that the instrument is indicating the correct SPL for a standard reference calibrator output at a specified SPL and frequency (e.g., 94 dB at 1000 Hz). The posttest calibration is done to determine if the instrumentation, including the microphone, has drifted during the measurement and, if so, if the drift is large enough to invalidate the data obtained. Calibrators may be electronic transducer-type devices with loudspeaker outputs from an internal oscillator, or "pistonphones," which use a reciprocating piston in a closed cavity to produce sinusoidal pressure variations as the cylinder volume changes. Both types include adapters that allow the device to be mated to microphones of different diameters. Calibrators should be sent to the manufacturer at least annually for bench calibration and certification.

There are many other issues that bear on the proper application of sound level measurement equipment, such as microphone selection and placement, averaging time and sampling schemes, and statistical data reduction techniques, all of which are beyond the scope of this chapter. For further coverage of these topics, the reader is referred to the acoustics texts of Harris (1991) and Berger et al. (2003).

## 3.2 Sound and Noise Metrics

### 3.2.1 Exchange or Trading Rates

Because both sound amplitude and sound duration determine the energy of an *exposure*, average-type measures are based on simple algorithms or *exchange rates*, which trade amplitude for time and vice versa. For example, most noise regulations, OSHA (1983) or otherwise, stipulate that a worker's exposure may not exceed a maximum daily accumulation of noise energy. In other words, in OSHA terms the product of duration and intensity must remain under the regulatory cap or *permissible exposure limit* (PEL) of 90 dBA TWA for an 8-h work period, which is equivalent to a 100% noise dose. Much debate has occurred over the past several decades about which exchange rate is most appropriate for prediction of hearing damage risk, and most countries currently use either a 3- or 5-dB relationship. The OSHA exchange rate is 5 dB, which means that an increase (decrease) in decibel level by 5 dB is equivalent (in exposure) to a doubling (halving) of time. For instance, using the OSHA PEL of 90 dBA for 8 h, if a noise is at 95 dBA, the allowable exposure per workday is half of 8 h, or 4 h. If a noise is at 85 dBA, the allowable exposure time is twice 8 h, or 16 h. These allowable reference exposure durations (*T* values) are provided in Table A-1 of the OSHA (1983) regulation or they may be computed using the formula for *T,* which appears below as equation (14). The 5-dB exchange rate is predicated in part on the theory that intermittent noise is less damaging than continuous noise because some recovery from temporary hearing loss occurs during quiet periods. Arguments against it include the fact that an

exchange of 5 dB for a factor of 2 in time duration has no real physical basis in terms of energy equivalence. Furthermore, there is some evidence that the quiet periods of intermittent noise exposures are insufficient in length to allow for recovery to occur. The 5-dB exchange rate is used for all measures associated with OSHA regulations, including the most general average measure of $L_{OSHA}$, the TWA referenced to an 8-h duration, and noise dose in percent.

Most European countries use a 3-dB exchange rate, also known as the aforementioned *equal-energy rule*. In this instance, a doubling (halving) of sound intensity, which corresponds to a 3-dB increase (decrease), equates (in energy) to a doubling (halving) of exposure duration. The equal energy concept stems from the fact that if sound intensity is doubled or halved, the equivalent sound intensity level change is 3 dB. An exposure to 90 dBA for 8 h using a 3-dB exchange rate is equivalent to a 120-dBA exposure of only 0.48 min. Because each increase in decibels by 10 corresponds to a 10-fold increase in intensity, the 30-dB increase from 90 to 120 dBA represents a 1000-fold ($10^3$) increase in sound intensity, from 0.001 to 1 W/m$^2$. The 90-dBA exposure period is 8 h, or 480 min, and this must be reduced by the same factor as the SPL increase, so $480/1000 = 0.48$ min, or 29 s. The 3-dB exchange rate is used for all measures associated with the equivalent continuous sound level, $L_{eq}$.

### 3.2.2 Average and Integrated SPLs

As discussed earlier, conventional SLMs provide "momentary" decibel measurements that are based on very short moving-window exponential averages using FAST, SLOW, or IMPULSE time constants. However, since the majority of noises fluctuate over time, one of several types of average measurements, discussed below, is usually most appropriate as a descriptor of the central tendency of the noise. Averages may be obtained in one of two ways: (1) by observing and recording conventional SLM readouts using a short-time-interval sampling scheme and then manually computing the average value from the discrete values or (2) by using a SLM or dosimeter which automatically calculates a running-average value using microprocessor circuitry which provides either a true continuous integration of the area under the sound pressure curve or which obtains discrete samples of the sound at a very fast rate and computes the average, per ANSI (2007a) S1.43-1997(R2007). Generally, average measures obtained by method 2 yield more representative values because they are based on continuous or near-continuous sampling of the waveform, which the human observer cannot perform well even with continuous vigilance.

The average metrics discussed below are generally considered as the most useful for evaluating noise hazards in industry, annoyance potential in the community, and other sounds in the laboratory or in the field which fluctuate over time. In most cases for industrial hearing conservation as well as community noise annoyance purposes, the metrics utilize the A-weighting scale. For precise spectral measurements with no frequency weighting, the decibel unweighted (linear) scale may be

applied in the measurements. The equations are all in a form where the data values are considered to be discrete sound levels. Thus, they can be applied to data from conventional SLMs or dosimeters. For continuous sound levels (or when the equations are used to describe true integrating meter functioning), the $\sum$ sign in the equations would be replaced by the integral sign $\int_0^T$ and the $t_i$ replaced by $dt$. Variables used in the equations are as follows:

$L_i$ = decibel level in measurement interval $i$

$N$ = number of intervals

$T$ = total measurement time period

$t_i$ = length of measurement interval $i$

$Q$ = exchange rate (dB)

$q = Q/\log_{10}(2)$ $\begin{cases} \text{for 3-dB exchange, } q = 10.0 \\ \text{for 4-dB exchange, } q = 13.29 \\ \text{for 5-dB exchange, } q = 16.61 \end{cases}$

The general form equation for *average SPL*, or $L_{\text{average}}$, $L_{\text{av}}$, is

$$L_{\text{av}}(Q) = q\,\log_{10}\left[\frac{1}{T}\sum_{i=1}^{N}\left(10^{Li/q}\,t_i\right)\right] \qquad (8)$$

The *equivalent continuous sound level*, $L_{\text{eq}}$, equals the continuous sound level which when integrated or averaged over a specific time would result in the same energy as a variable sound level over the same time period. The equation for $L_{\text{eq}}$, which uses a 3-dB exchange rate, is

$$L_{\text{eq}} = L_{\text{av}}(3) = 10\,\log_{10}\left[\frac{1}{T}\sum_{i=1}^{N}\left(10^{Li/10}t_i\right)\right] \qquad (9)$$

In applying the $L_{\text{eq}}$, the individual $L_i$ values are usually in dBA. Equation (9) may also be used to compute the overall equivalent continuous sound level (for a single site or worker) from individual $L_{\text{eq}}$ values that are obtained over contiguous time intervals by substituting the $L_{\text{eq}}$ values in the $L_i$ variable. The $L_{\text{eq}}$ values are often expressed with the time period over which the average is obtained; for instance, $L_{\text{eq}}(24)$ is an equivalent continuous level measured over a 24-h period. Another average measure that is derived from $L_{\text{eq}}$ and often used for community noise quantification is $L_{\text{dn}}$, which is simply a 24-h $L_{\text{eq}}$ measurement with a 10-dB penalty added to all nighttime noise levels from 10 P.M. to 7 A.M. The rationale for the penalty is that humans are more disturbed by noise, especially due to sleep arousal, during nighttime periods.

The equation for the *OSHA average noise level*, $L_{\text{OSHA}}$, which uses a 5-dB exchange rate, is

$$L_{\text{OSHA}}(5) = 16.61\,\log_{10}\left[\frac{1}{T}\sum_{i=1}^{N}\left(10^{L_{iA}/16.61}t_i\right)\right] \qquad (10)$$

where $L_{iA}$ is in dBA, slow response.

OSHA's TWA is a special case of $L_{\text{OSHA}}$ which requires that the total time period always be 8 h, that time is expressed in hours, and that sound levels below 80 dBA, termed the *threshold level*, are not included in the measurement:

$$\text{TWA} = 16.61\,\log_{10}\left[\frac{1}{8}\sum_{i=1}^{N}\left(10^{L_{iA}/16.61}t_i\right)\right] \qquad (11)$$

where $L_{iA}$ is in dBA, slow response, and $T$ is always 8 h. Only $L_{iA} \geq 80$ dBA is included.

OSHA's *noise dose* is a percentage representation of the noise exposure, where 100% is the maximum allowable dose, corresponding to a 90-dBA TWA referenced to 8 h. Dose utilizes a *criterion sound level*, which is presently 90 dBA, and a *criterion exposure period*, which is presently 8 h. A noise dose of 50% corresponds to a TWA of 85 dBA, and this is known as the OSHA *action level*. Calculation of dose, $D$, is as follows:

$$D = \frac{100}{T_c}\sum_{i=1}^{N}\left(10^{(L_{iA}-L_c)/q}t_i\right) \qquad (12)$$

where $L_{iA}$ is in dBA, slow response, $L_c$ is the criterion sound level, and $T_c$ is the criterion exposure duration. Only $L_{iA} \geq 80$ dBA is included.

Noise dose $D$ can also be expressed as follows for a constant sound level over the workday:

$$D = 100\left(\frac{C_1}{T_1} + \frac{C_2}{T_2} + \cdots + \frac{C_n}{T_n}\right) \qquad (13)$$

where $C_i$ is the total time (h) of actual exposure at $L_i$, $T_i$ is total time (h) of reference-allowable exposure at $L_i$, from Table G-16a of OSHA, (1983), and $C_i/T_i$ represents a partial dose at sound level $i$.

The *reference allowable* exposure $T$ for a given sound level can also, in lieu of consulting Table G-16a in OSHA (1983), be computed as

$$T = 8/2^{(L-90)/5} \qquad (14)$$

where $L$ is the measured dBA level.

Two other useful equations to compute dose $D$ from TWA and vice versa are

$$D = 100 \times 10^{(\text{TWA}-90)/16.61} \qquad (15)$$

$$\text{TWA} = [16.61\,\log_{10}(D/100)] + 90 \qquad (16)$$

where $D$ is the dose in percent. TWA can also be found for each value of dose $D$ in Table A-1 of OSHA (1983).

A final measure that is particularly useful for quantifying the exposure due to single or multiple occurrences of an acoustic event (such as a complete operating cycle of a machine, a vehicle drive-by, or an aircraft flyover) is the *sound exposure level* (SEL). The SEL represents a sound 1 s in length that imparts the

same acoustic energy as a varying or constant sound that is integrated over a specified time interval $t_i$ in seconds. Over $t_i$, an $L_{eq}$ is obtained which indicates that SEL is used only with a 3-dB exchange rate. A reference duration of 1 s is applied for $t_0$ in the following equation for SEL:

$$\text{SEL} = L_{eq} + 10 \log_{10}(t_i/t_0) \qquad (17)$$

where $L_{eq}$ is the equivalent SPL measured over time period $t_i$.

Detailed example problems and solutions using the formulas above may be found in Casali and Robinson (1999).

## 4 INDUSTRIAL NOISE REGULATION AND ABATEMENT

### 4.1 Need for Attention to Noise

In this section the discussion will concentrate on the management of noise in industry, because that is the major source of noise exposure for most people, and as such, it constitutes a very common threat toward noise-induced hearing loss (NIHL). Many of the techniques for measurement, engineering control, and hearing protection also apply to other exposures, such as those encountered in recreational or military settings. The need for attention to industrial noise is indicated when (1) noise creates sufficient intrusion and operator distraction such that job performance and even job satisfaction are compromised; (2) noise creates interference with important communications and signals, such as interoperator communications, machine- or process-related aural cues, alerting/emergency signals, or military tactics and missions; and/or (3) noise exposures constitute a hazard for NIHL in workers.

### 4.2 OSHA (and Other) Noise Exposure Regulated Limits

In U.S. workplaces, while a few industrial hearing conservation programs were voluntarily implemented by a few industries in the 1940s and 1950s (Berger, 2003a), legal limits on general industrial noise exposure and application of hearing protection were not promulgated into law until May of 1971, and this occurred with the Occupational Noise Exposure Standard of the Occupational Safety and Health Act. The OSHA noise standard was the first requirement, based on exposure levels, for noise abatement and hearing protection devices (HPDs) in the general industry (OSHA, 1971a), and a similar law was promulgated for construction (OSHA, 1971b), these settings being where the great majority of U.S. citizens were and continue to be at risk for hearing loss due to noise exposure.

In 1983, some 12 years after the original OSHA legislation on noise in industry, the legal advent of the OSHA Hearing Conservation Amendment (OSHA, 1983) for General Industry immediately caused the proliferation of HPDs in U.S. industrial workplaces because this amendment required a choice of HPDs to be *supplied* to any worker exposed to above an 85 dBA TWA,

or 50% noise dose, for an 8-h workday, with the measurement taken on the "slow" scale and using a 5-dB exchange rate between exposure dBA level and time of exposure. Other industries, including airline, truck and bus carriers, railroads, and oil and gas well drilling, developed separate, and generally less comprehensive, noise and hearing conservation regulations than those of OSHA (1971a, 1983) for general industry, and unfortunately, to date, there has never been an analogue to the OSHA Hearing Conservation Amendment of 1983 adopted into law for the construction industry, where noise levels are often high and auditory warning signals (e.g., vehicle backup alarms) are prevalent (Casali and Alali, 2009). Finally, in the mining industry, noise exposure limits and hearing protection were addressed in that industry's regulation, first under the Federal Coal Mine Health and Safety Act of 1969 and later under the Federal Mine Safety and Health Amendments Act of 1977. In 1999, the Mine Safety and Health Administration (MSHA) issued a more comprehensive noise regulation that governed all forms of mining (MSHA, 1999).

In regard to combating the hearing loss problem in OSHA terms, if the noise dose (in general industry) exceeds the OSHA action level of 50%, which corresponds to an 85-dBA TWA, the employer must institute a *hearing conservation program* (HCP) which consists of several facets (OSHA, 1983). If the criterion level of 100% dose is exceeded (which corresponds to the permissible exposure level (PEL) of 90-dBA TWA for an 8-h day), the regulations specifically state that steps must be taken to reduce the employee's exposure to the PEL or below via administrative work scheduling and/or the use of engineering controls. It is stated specifically that HPDs must be provided if administrative and/or engineering controls fail to reduce the noise to the PEL. Therefore, in applying the letter of the law, HPDs are only intended to be relied on when administrative or engineering controls are infeasible or ineffective. The final OSHA noise-level requirement pertains to impulsive or impact noise, which is not to exceed a PEAK SPL limit of 140 dB.

### 4.3 Hearing Conservation Programs

#### 4.3.1 Shared Responsibility: Management, Workers, and Government

A successful HCP depends on the shared commitment of management and labor as well as the quality of services and products provided by external noise control consultants, audiology or medical personnel who conduct the hearing measurement program, and vendors (e.g., hearing protection suppliers). Involvement and interaction of corporate positions such as the plant safety engineer, ergonomist, occupational nurse, noise control engineer, purchasing director, and manufacturing supervisor are important. Furthermore, government regulatory agencies, such as OSHA and MSHA, have a responsibility to maintain and disseminate up-to-date noise exposure regulations and HCP guidance, to conduct regular in-plant compliance checks of noise exposure and quality of HCPs, and to provide enforcement where noise control and/or hearing protection is inadequate. Finally, the

"end user" of the HCP, that is, the worker, must be an informed and motivated participant. For instance, if a fundamental component of the HCP is the personal use of HPDs, the effectiveness of the program in preventing NIHL will depend most heavily on the worker's commitment to wear the HPD properly and consistently. Failure by any of these groups to carry out their responsibilities can result in HCP failure and worker hearing loss. Side benefits of a successful HCP may include a marked reduction in noise-induced distractions and interference on the job and an improvement in worker comfort and morale.

### 4.3.2 Hearing Conservation Program Components

Hearing conservation in industry should be thought of as a strategic, programmatic effort that is initiated, organized, implemented, and maintained by the employer, with cooperation from other parties as indicated above. A well-accepted approach is to address the noise exposure problem from a *systems* perspective, wherein empirical noise measurements provide data input which drives the implementation of countermeasures against the noise (including engineering controls, administrative strategies, and personal hearing protection). Subsequently, noise and audiometric data, which reflect the effectiveness of those countermeasures, serve as feedback for program adjustments and improvements. A brief discussion of the major elements of a HCP, as dictated by OSHA (1983), follows.

*Monitoring*   Noise exposure monitoring is intended to identify employees for inclusion in the HCP and to provide data for the selection of HPDs. The data are also useful for identifying areas where engineering noise control solutions and/or administrative work scheduling may be necessary. All OSHA-related measurements, with the exception of the PEAK SPL limit, are to be made using a SLM or dosimeter (of at least ANSI type 2) set on the dBA scale, SLOW response, using a 5-dB exchange rate, and incorporating all sounds whose levels are from 80 to 130 dBA. It is unspecified, but it must be assumed that sounds above 130 dBA should also be monitored. (Of course, such noise levels represent OSHA noncompliance since the maximum allowable continuous sound level is 115 dBA.) Appendix G of the OSHA regulation suggests that monitoring be conducted at least once every one or two years. Related to the noise-monitoring requirement is that of notification. Employees must be given the opportunity to observe the noise-monitoring process, and they must be notified when their exposures exceed the 50% dose (85-dBA TWA) level.

*Audiometric Testing Program*   All employees whose noise exposures are at the 50% dose level or above must be included in a pure-tone audiometric testing program wherein a baseline audiogram is completed within six months of the first exposure, and subsequent tests are done on an annual basis. Annual audiograms are compared against the baseline to determine if the worker has experienced a *standard threshold shift* (STS), which

is defined by OSHA (1983). The annual audiogram may be adjusted for age-induced hearing loss (presbycusis) using the gender-specific correction data in Appendix F of the regulation. All OSHA-related audiograms must include 500, 1000, 2000, 3000, 4000, and 6000 Hz, in comparison to most clinical audiograms, which typically extend from 125 to 8000 Hz. If an STS is revealed, a licensed physician or audiologist must review the audiogram and determine the need for further audiological or otological evaluation, the employee must be notified of the STS, and the selection and proper use of HPDs must be revisited.

*Training Program and Record Keeping*   An essential component of an HCP is a training program for all noise-exposed workers. Training elements to be covered include the effects of noise on hearing; purpose, selection, and use of HPDs; and purpose and procedures of audiometric testing. Also, accurate records must be kept of all noise exposure measurements, at least from the last two years, as well as audiometric test results for the duration of the worker's employment. It is important, but not required by OSHA, that noise and audiometric data be used as feedback for improving the program. For example, noise exposure records may be used to identify machines that need maintenance attention, to assist in the relocation of noisy equipment during plant layout efforts, to provide information for future equipment procurement decisions, and to target plant areas that are in need of noise control intervention. Some employers plot noise levels on a "contour map," delineating floor areas by their decibel levels. When monitoring indicates that the noise level in a particular contour has changed, it is taken as a sign that the machinery and/or work process has changed in the area and that further evaluation may be needed.

*Hearing Protection Devices*   The OSHA Hearing Conservation Amendment (OSHA, 1983) requires that a selection of HPDs that are suitable for the noise and work situation must be made available to all employees whose TWA exposures meet or exceed 85 dBA. Such HPDs are also useful outside the workplace, for the protection of hearing against noises produced by power tools, lawn care equipment, recreational vehicles, target shooting and hunting, spectator events, ordnance and various military weapons, as well as many other exposures. Complete overviews of conventional HPDs which provide noise attenuation via passive (nonelectronic) means may be found in Berger (2003b) and Gerges and Casali (2007). Following is a brief overview of the basic types of devices, primarily adopted from the book chapter by Gerges and Casali (2007).

   *Earplugs* consist of vinyl, silicone, spun fiberglass, cotton/wax combinations, and open-cell or closed-cell foam products that are inserted into the ear canal to form a noise-blocking seal. Proper fit to the user's ears and training in insertion procedures are critical to the success of earplugs. A related device is the *semi-insert* or *ear canal cap*, which consists of earplug-like pods that are positioned at the rim of the ear canal and held in place by a lightweight headband. The headband is

useful for storing the device around the neck when the user moves out of the noise. *Earmuffs* consist of earcups, usually of a rigid plastic material with an absorptive liner, that enclose the outer ear completely and seal around it with foam- or fluid-filled cushions. A headband connects the earcups, and on some models this band is adjustable so that it can be worn over the head, behind the neck, or under the chin, depending on the presence of other headgear, such as a welder's mask. In general terms, as a group, earplugs provide better attenuation than earmuffs below about 500 Hz and equivalent or greater protection above 2000 Hz (Gerges and Casali, 2007). At intermediate frequencies, earmuffs sometimes have the advantage in attenuation. Earmuffs are generally more easily fit by the user than either earplugs or canal caps, and depending on the temperature and humidity of the environment, the earmuff can be uncomfortable (in hot or high-humidity environments) or a welcome ear insulator (in a cold environment). Semi-inserts (canal caps) generally offer less attenuation and comfort than earplugs or earmuffs, but because they are readily storable around the neck, they are convenient for those workers who frequently move in and out of noise.

Conventional styles of passive earplugs and earmuffs generally exhibit a spectral profile of attenuation that is nonlinear with respect to sound frequency; that is, the attenuation generally increases with increased frequency. At any given frequency, the most attenuation that any HPD can provide, also termed its "theoretical limit," is the bone conduction threshold of the wearer. This is because at sound levels above the bone conduction threshold the HPD is "flanked" by the bone conduction pathway to the sensory portion of the ear and sound enters the ear through structural-borne vibrations of the skull (Berger, 2003b).

Beginning in the mid-1980s, conventional passive HPDs were augmented with new features, such as attenuation spectra which are uniform or "flat" as a function of frequency, attenuation capabilities that increase as a function of increases in incident sound level (termed "amplitude-sensitive" or "level-dependent" devices), adjustable attenuation as achieved with adjustable continuous valves or discrete dampers inside a vent running through the HPD, and passive noise attenuators achieved using one-end-closed tube structures that provide quarter-wave resonance to cancel offending noise in a narrow frequency band. All of these passive HPD augmentations, along with research results on their performance, are covered in detail in Casali (2010a). Furthermore, battery-powered or "active" electronic HPD features began to appear in the 1980s, including active noise cancellation that incorporates "anti-noise" of inverted phase relationship with the noise to be cancelled (primarily effective below about 1000 Hz), electronically-modulated sound transmission devices which provide microphone pickup of sounds external to the HPD and output-limited, amplified pass-through of signals and speech within a certain passband through the HPD, and military/law enforcement-oriented tactical communications and protection systems (TCAPS) which provide covert two-way communications, signal pass-through capabilities, and gunfire-responsive protection. All versions of these active electronic HPDs are reviewed in Casali (2010b), along with research data on their performance in various applications.

Regardless of their type, HPD effectiveness depends heavily on the proper fitting and use of the devices (Park and Casali, 1991). Therefore, the employer is required to provide training in the fitting, care, and use of HPDs to all employees affected (OSHA, 1983). Hearing protector use becomes mandatory when the worker has not undergone the baseline audiogram, has experienced an STS, or has a TWA exposure that meets or exceeds 90 dBA. In the case of the worker with an STS, the HPD must attenuate the noise to 85 dBA TWA or below. Otherwise, the HPD must reduce the noise to at least 90 dBA TWA.

The protective effectiveness or adequacy of an HPD for a given noise exposure must be determined by applying the attenuation data as currently required by the U.S. Environmental Protection Agency (EPA, 1979) to be included on protector packaging. These data are obtained from psychophysical threshold tests at nine 1/3-octave bands with centers from 125 to 8000 Hz that are performed on human subjects, and the difference between the thresholds with the HPD on and without it constitutes the attenuation at a given frequency. Spectral attenuation statistics (means and standard deviations) and the single-number noise reduction rating (NRR) which is computed therefrom are provided. The ratings are the primary means by which end users compare different HPDs on a common basis and make determinations of whether adequate protection and OSHA compliance will be attained for a given noise environment.

The most accurate method of determining HPD adequacy is to use octave-band measurements of the noise and the spectral mean and standard deviation attenuation data to determine the protected exposure level under the HPD. This is called the *National Institute for Occupational Safety and Health* (*NIOSH*) *long method* or *octave-band method*. Computational procedures appear in NIOSH (1975). Because this method requires octave-band measurements of the noise, preferably with each noise band's data in TWA form, the data requirements are large and the method is not widely applied in industry. However, because the noise spectrum is compared against the attenuation spectrum of the HPD, a "matching" of exposure to protector can be obtained; therefore, the method is considered to be the most accurate available.

The NRR represents a means of collapsing the spectral attenuation data into one broadband attenuation estimate that can easily be applied against broadband dBC or dBA TWA noise exposure measurements. In calculation of the NRR, the mean attenuation is reduced by two standard deviations; this translates into an estimate of protection theoretically achievable by 98% of the population (EPA, 1979). The NRR is intended primarily to be subtracted from the dBC exposure TWA to estimate the protected exposure level in dBA:

$$\text{Workplace TWA (dBC)} - \text{NRR}$$
$$= \text{protected TWA (dBA)} \qquad (18)$$

Unfortunately, because OSHA regulations require that noise exposure monitoring be performed in dBA, the dBC values may not be readily available to the hearing conservationist. In the case where the TWA values are in dBA, the NRR can still be applied, albeit with some loss of accuracy. With dBA data, a 7-dB "safety" correction is applied to the NRR to account for the largest typical differences between C- and A-weighted measurements of industrial noise, and the equation is

$$\text{Workplace TWA (dBA)} - (\text{NRR} - 7)$$
$$= \text{protected TWA (dBA)} \quad (19)$$

Although the methods above are promulgated by OSHA (1983) for determining HPD adequacy for a given noise situation, a word of caution is needed. The data appearing on HPD packaging are obtained under optimal laboratory conditions with properly fitted protectors and trained human subjects. In no way does the "experimenter-fit" protocol and other aspects of the currently required (by the EPA) test procedure, ANSI S3.19-1974 (ANSI, 1974), represent the conditions under which HPDs are selected, fit, and used in the workplace (Park and Casali, 1991). *Therefore, the attenuation data used in the octave-band or NRR formulas shown above are, in general, inflated and cannot be assumed as representative of the protection that will be achieved in the field*. The results of a review of research studies in which manufacturers' on-package NRRs were compared against NRRs computed from actual subjects taken with their HPDs from field settings are shown in Figure 6. Clearly, the differences between laboratory and field estimates of HPD attenuation are large and the hearing conservationist must take this into account

when selecting protectors. Efforts by ANSI Working Group S12/WG11 has focused on the development of an improved testing standard, ANSI S12.6-1997(R2008) (ANSI, 2008), which has an important human factors provision in its "method B" for subject (not experimenter) fitting of the HPD and relatively naïve (not trained) subjects, as is the provision in ANSI S3.19-1974 (ANSI, 1974) and "method A" of the newer standard, ANSI S12.6-1997(R2008) (ANSI, 2008).

The method B testing protocol of ANSI S12.6-1997(R2008) has been demonstrated to yield attenuation data that are more representative of those achievable under workplace conditions wherein a high-quality HCP is operated. However, at press time for this chapter, the EPA was in the process of attempting to promulgate a comprehensive new federal law to govern the testing and labeling of all hearing protectors of various types. In the EPA's proposed rule (EPA, 2009), the basic test for obtaining the passive spectral attenuation of an HPD which is proposed for use in developing the HPD's noise reduction label relies on the use of the method A fitting technique of ANSI S12.6-1997(R2008), which has been demonstrated in the same or similar forms to produce attenuation values that are substantially *higher* than those achieved by actual users in field use (e.g., Berger, 2003b; Park and Casali, 1991). It remains to be seen, however, whether method A or B will be adopted in the EPA rule or whether any new rule will even be adopted into law. It is also important to note that the EPA's proposed rule (EPA, 2009) does have, for the first time in proposed U.S. law, provisions for much more complete testing and labeling of augmented hearing protectors that offer special capabilities, as compared to that which is available in the current EPA rule (EPA, 1979), which relies exclusively on the ANSI S3.19-1974 standard that
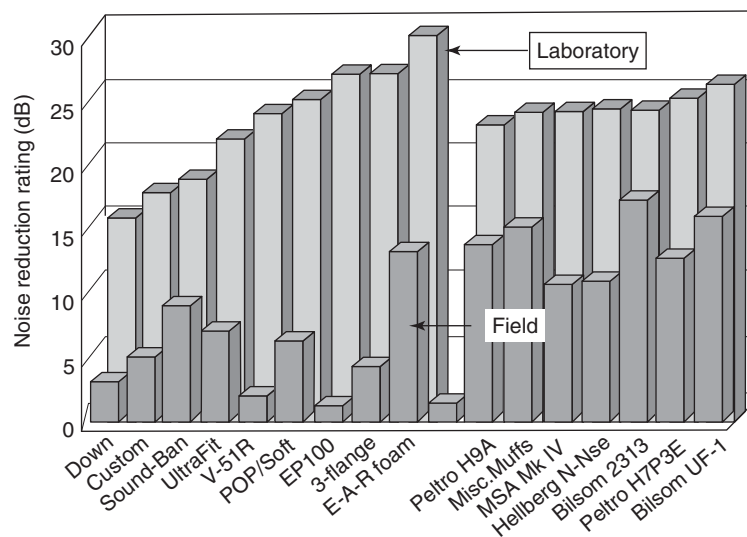


**Figure 6** Comparison of hearing protection device NRRs by device type: manufacturers' laboratory data vs. real-world "field" data. (Adapted with permission from Berger, 2003b.)

is limited to testing of conventional passive attenuation. In other words, assuming that the EPA's proposed rule is promulgated into law, the capabilities of active noise cancellation, active sound transmission, impulsive (e.g., gunfire) protection circuits, passive amplitude-sensitive devices, and other special augmentations will be tested and labeled as to their performance.

### 4.4 Engineering Noise Control

As discussed above, hearing protection and/or administrative controls are not a panacea for combating the risks posed by noise. They should not supplant noise control engineering; in fact, the best solution, in part because it does not rely on employee behavior, is to reduce the noise itself, preferably at the emission source. The physical reduction of the noise energy, either at its source, in its path, or at the worker, should be a major focus of noise management programs. However, in many cases where noise control is ineffective, infeasible (as on an airport taxi area), or prohibitively expensive, HPDs become the primary countermeasure.

There are many techniques used in noise control, and the specific approach must be tailored to the noise problem at hand. Spectrum analyzer measurements are typically used by noise control engineers in the selection of control strategies. Example noise control strategies are (1) *isolation of the source* via relocation, enclosure, or vibration damping using metal or air springs (below about 30 Hz) or elastomer (above 30 Hz) supports; (2) *reduction at the source or in the path* using mufflers or silencers on exhausts, reducing cutting, fan, or impact speeds, dynamically balancing rotating components, reducing fluid flow speeds and turbulence, absorptive foam or fiberglass on reflective surfaces to reduce reverberation, shields to reflect and redirect noise (especially high frequencies), and lining or wrapping of pipes and ducts; (3) *replacement or alteration of machinery*, including belt drives as opposed to noisier gears, electrical rather than pneumatic tools, and shifting frequency outputs such as by using centrifugal fans (low frequencies) rather than propeller or axial fans (high frequencies), keeping in mind that low frequencies propagate further than high frequencies, but high frequencies are more hazardous to hearing; and (4) *application of quieter materials*, such as rubber liners in parts bins, conveyors, and vibrators, resilient hammer faces, bumpers on material handling equipment, nylon slides or rubber tires rather than metal rollers, and fiber rather than metal gears. Further discussion of these and other techniques may be found in Driscoll and Royster (2003) and in Harris (1991), and an illustration of implementation possibilities in an industrial plant appears in Figure 7.

A final approach that has recently become available to industry is *active noise reduction* (ANR), in which an electronic system is used to transduce an offensive noise in a sound field and then process and feed back the noise into the same sound field such that it is exactly $180°$ out of phase with, but of equal amplitude to, the original noise. The superposition of the out-of-phase *anti-noise* with the original noise causes physical cancellation of the noise in a target zone of the workplace. For highly repetitive, predictable noises, synthesis of the anti-noise, as opposed to transduction and reintroduction, may also be used in a feedforward fashion. At frequencies below about 1000 Hz, the ANR technique is most effective, which is fortuitous since the passive noise control materials to combat low-frequency noise, such as absorptive liners and barriers, are typically heavy, bulky, and expensive. At higher frequencies and their corresponding shorter wavelengths, the processing and phase relationships become more complicated and cancellation is less successful, although the technology is improving rapidly and the bandwidth of effective ANR cancellation is increasing (Casali et al., 2004; Casali, 2010b).

In designing and implementing noise control hardware, it is important that ergonomics be taken into account. For instance, in a sound-treated booth to house an operator, the ventilation system, lighting, visibility outward to the surrounding work area, and other considerations relating to operator comfort and performance must be considered. With regard to noise-isolating machine enclosures, access provisions should be designed so as not to compromise the operator–machine interface. In this regard, it is important that both operation and maintenance needs be met. If noise control hardware creates difficulties for the operators in carrying out their jobs, they may tend to modify or remove it, rendering it ineffective.

## 5 AUDITORY EFFECTS OF NOISE

### 5.1 Hearing Loss in the United States

Noise-induced hearing loss (NIHL) is one of the most widespread occupational maladies in the United States, if not the world. In the early 1980s, it was estimated that over 9 million workers were exposed to noise levels averaging over 85 dBA for an 8-h workday (EPA, 1981). Today, this number is likely to be higher because the control of noise sources, in both type and number, has not kept pace with the proliferation of industrial and service sector development. Due in part to the fact that before the first OSHA noise exposure regulation of 1971 there were no U.S. federal regulations governing noise exposure in general industry, many workers over 50 years of age now exhibit hearing loss that results from the effects of occupational noise.

Of course, the total noise exposure from both occupational and nonoccupational sources determines the NIHL that a victim experiences. Of the estimated 28 million Americans who exhibit significant hearing loss due to a variety of etiologies, such as pathology of the ear and hereditary tendencies, over 10 million have losses that are directly attributable to noise exposure [National Institutes of Health (NIH), 1990]. Therefore, the noise-related losses are *preventable* in nearly all cases. The majority of losses are due to on-the-job exposures, but leisure noise sources do contribute a significant amount of energy to the total noise exposure of some people. Although the effects of noise exposure are serious and must be reckoned with by the safety professional, one
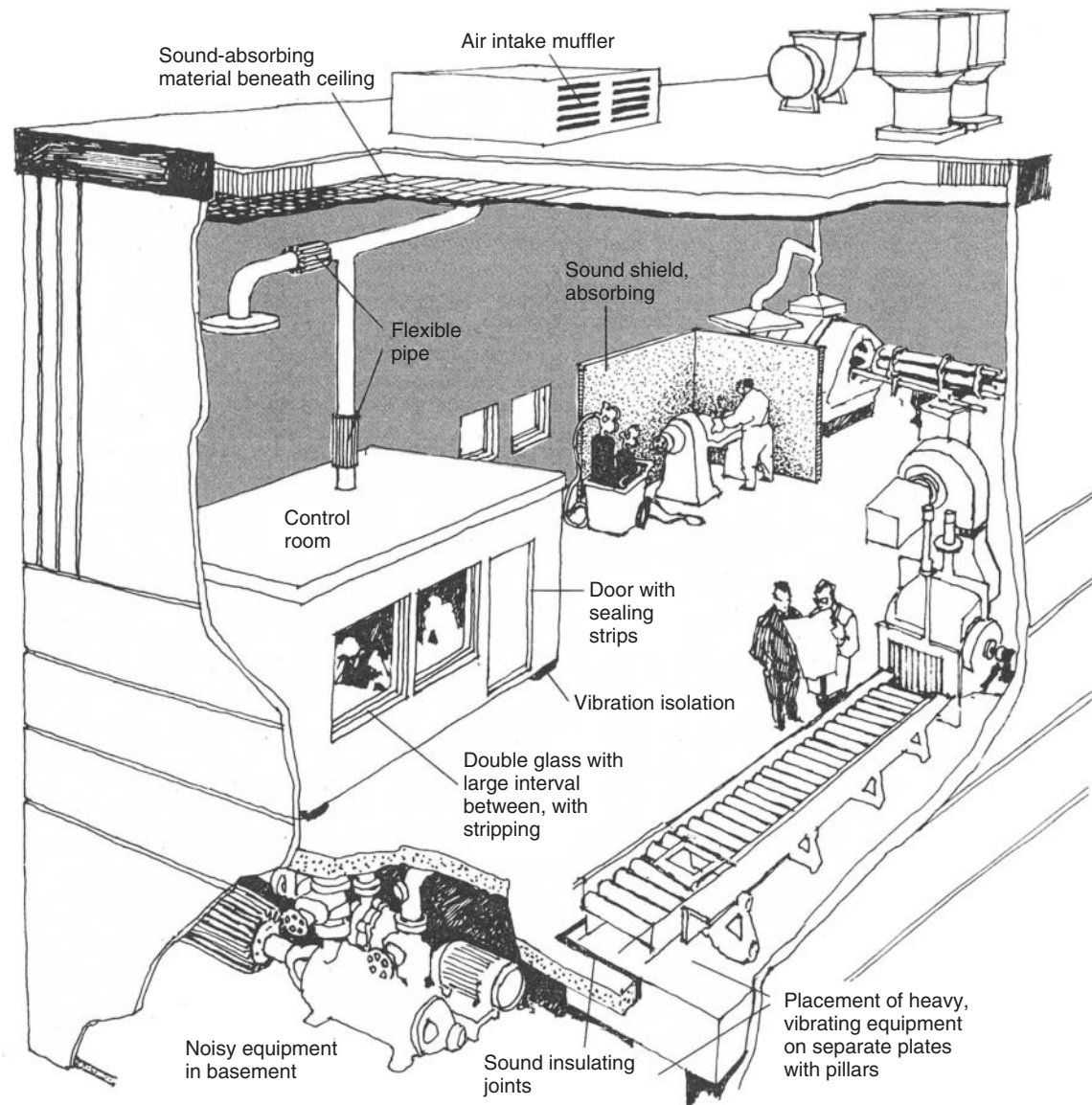
**Figure 7**    Noise control implementation in an industrial plant. (Adapted with permission from OSHA, 1980.)

fact is encouraging: Process/machine-produced noise, as well as most sources of leisure noise, are physical stimuli that can be avoided, reduced, or eliminated; therefore, NIHL is preventable with effective abatement and protection strategies. Total elimination of NIHL should thus be the only acceptable goal.

Noise-induced hearing loss is also a staggering problem in the military, especially during periods of war. Recent estimates from 2007 showed that since the Afghanistan war began in 2001 and the Iraq war in 2003, approximately 52% of combat soldiers

experienced moderately severe hearing loss or worse, primarily attributable to combat-related exposures [Defense Occupational and Environmental Health Readiness Data Repository (DOEHRS-DR), 2007]. Recent Army reports portray an even bleaker picture, with over *one-third* of U.S. soldiers who return from service in these two wars exhibiting permanent noise-induced hearing loss that is believed to be associated with military operations (Ahroon, 2007). Furthermore, the problem of noise-induced hearing loss is the *most common* military disability, as evidenced by over $1.2

billion spent on personnel hearing-related injuries in fiscal year 2006 alone (Saunders and Griest, 2009). This staggering cost included $786+ million for medical and hearing-assistive expenses associated with the noise-induced hearing impairment and $418+ million for the debilitating, life-pervasive tinnitus malady (persistent ringing or whistling in the ears). In fiscal year 2007, the U.S. Veterans Administration dispensed 348,920 hearing aids to veterans at a cost of approximately $141 million (Saunders and Griest, 2009). By comparison to these staggering annual expenditures relating to noise-induced hearing loss in the military, in industry the annual cost of disability payments for hearing-related injuries in about 30 million workers was about $242.4 million in 2001 (NIOSH, 2001). Part of the problem in military-related hearing loss is that warfighters may be inhibited from using hearing protection devices for fear that they may compromise their ability to maintain stealth, operate tactically, and hear threats. Therefore, improvements in hearing protection designs, particularly those that maintain or perhaps enhance the warfighter's situation awareness while simultaneously providing protection against gunfire and other noises, are much needed. More information on this subject may be found in the reviews of Casali (2010a,b).

## 5.2 Types and Etiologies of Noise-Induced Hearing Loss

Although the major concern of the industrial hearing conservationist is to prevent employee hearing loss that stems from occupational noise exposure, it is important to recognize that hearing loss may also emanate from a number of sources other than noise, including infections and diseases specific to the ear, most frequently originating in the middle or conductive portion; other bodily diseases, such as multiple sclerosis, which injures the neural part of the ear; ototoxic drugs, of which the mycin family is a prominent member; exposure to certain chemicals and industrial solvents; hereditary factors; head trauma; sudden hyperbaric- or altitude-induced pressure changes; and aging of the ear (presbycusis). Furthermore, not all noise exposure occurs on the job. Many workers are exposed to hazardous levels during leisure activities, from such sources as automobile/motorcycle racing, personal stereo headsets and car stereos, firearms, and power tools. The effects of noise on hearing are generally subdivided into acoustic trauma and temporary or permanent threshold shifts (Melnick, 1991).

### 5.2.1 Acoustic Trauma

Immediate organic damage to the ear from an extremely intense acoustic event such as an explosion is known as *acoustic trauma*. The victim will notice the loss immediately, and it often constitutes a permanent injury. The damage may be to the conductive chain of the ear, including rupture of the tympanum (eardrum) or dislodging of the ossicular chain (small bones and muscles) of the middle ear. Conductive losses can, in many cases, be compensated for with a hearing aid and/or surgically corrected. Neural damage may also occur, involving a dislodging of the hair cells and/or breakdown of the neural organ (Organ of Corti) itself. Unfortunately, neural loss is irrecoverable and not typically compensable with a hearing aid. Acoustic trauma represents a severe injury, but fortunately, its occurrence is relatively uncommon, even in industrial settings. However, it can occur due to sudden explosive-induced trauma in the military setting.

### 5.2.2 Noise-Induced Threshold Shift

A *threshold shift* is defined as an elevation of hearing level from a person's baseline hearing level and it constitutes a loss of hearing sensitivity. *Noise-induced temporary threshold shift* (NITTS), sometimes referred to as "auditory fatigue," is by definition recoverable with time away from the noise. Thus, elevation of threshold is temporary and usually can be traced to an overstimulation of the neural hair cells (actually, the stereocilia) in the Organ of Corti. Although the person may not notice the temporary loss of sensitivity, NITTS is a cardinal sign of overexposure to noise. It may occur over the course of a full workday in noise or even after a few minutes of exposure to very intense noise. Although the relationships are somewhat complex and individual differences are rather large, NITTS does depend on the level, duration, and spectrum of the noise as well as on the audiometric test frequency in question (Melnick, 1991).

With *noise-induced permanent threshold shift* (NIPTS), there is no possibility of recovery. NIPTS can manifest suddenly as a result of acoustic trauma; however, noises that cause NIPTS most typically constitute exposures that are repeated over a long period of time and have a cumulative effect on hearing sensitivity. In fact, the losses are often quite insidious in that they occur in small steps over a number of years of overexposure and the person may not be aware until it is too late. This type of exposure produces permanent neural damage, and although there are some individual differences as to magnitude of loss and audiometric frequencies affected, the typical pattern for NIPTS is a prominent elevation of threshold at the 4000-Hz audiometric frequency (sometimes called the 4-kHz notch), followed by a spreading of loss to adjacent frequencies of 3000 and 6000 Hz. From a classic study on workers in the jute weaver industry, Figure 8 depicts the temporal profile of NIPTS as the family of audiometric threshold shift curves, with each curve representing a different number of years of exposure. As noise exposure continues over time, the hearing loss spreads over a wider frequency bandwidth inclusive of midrange and high frequencies and encompassing the range of most auditory warning signals. In some cases, the hearing loss renders it unsafe or unproductive for the victim to work in certain occupational settings where the hearing of certain signals are requisite to the job. Unfortunately, the power of the consonants of speech sounds, which heavily influence the intelligibility of human speech, also lie in the frequency range that is typically affected by NIPTS, compromising the victim's ability to understand speech. This is the tragedy of NIPTS in that the worker's ability to communicate is hampered, often
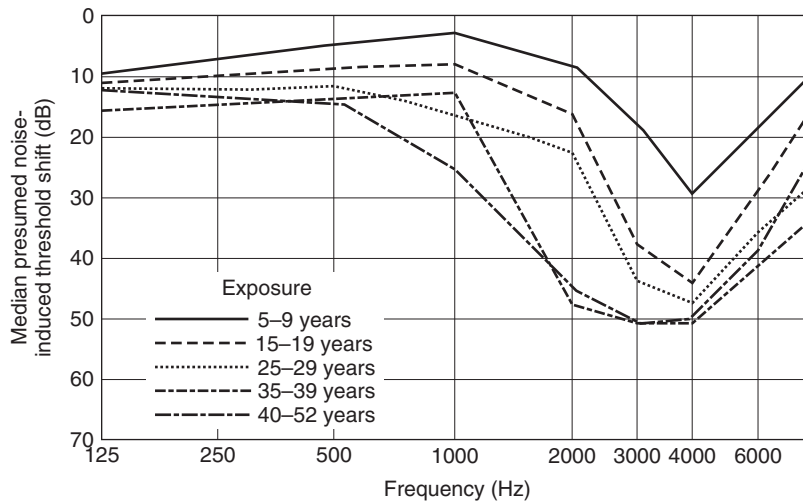
**Figure 8** Cumulative auditory effects of years of noise exposure in a jute-weaving industry. (Adapted with permission from Taylor et al., 1964.)

severely and always irrecoverably. Hearing loss is a particularly troubling disability because its presence is not overt; therefore, the victim is often unintentionally excluded from conversations and may miss important auditory signals because others either are unaware of the loss or simply forget about the need to compensate for it.

### 5.3  Concomitant Auditory Injuries

Following exposure to high-intensity noise, some people will notice that ordinary sounds are perceived as "muffled," and in some cases, they may experience a ringing or whistling sound in the ears, known as *tinnitus*. These manifestations should be taken as serious indications that overexposure has occurred and that protective action should be taken if similar exposures are encountered in the future. Tinnitus may also occur by itself or in conjunction with NIPTS. Some people report that tinnitus is always present, pervading their lives. It thus has the potential to be quite disruptive and in severe cases debilitating.

More rare than tinnitus, but typically quite debilitating, is the malady known as *hyperacusis*, which refers to hearing that is extremely sensitive to sound. Hyperacusis can manifest in many ways, but a number of victims report that their hearing became painfully sensitive to sounds of even normal levels after exposure to a particular noise event. Therefore, at least for some, hyperacusis can be traced directly to noise exposure. Sufferers often must use HPDs when performing normal activities, such as walking on city streets, visiting movie theaters, or washing dishes in a sink, because such activities produce sounds that are painfully loud to them. It should be noted that hyperacusis sufferers often exhibit normal audiograms, even though their reaction to sound is one of hypersensitivity.

## 6  PERFORMANCE, NONAUDITORY, AND PERCEPTUAL EFFECTS OF NOISE

### 6.1  Performance and Nonauditory Health Effects

#### 6.1.1  Task Performance Effects

It is important to recognize that, among other deleterious effects, noise can degrade operator task performance. Research studies concerning the effects of noise on performance are primarily laboratory based and task/noise specific; therefore, extrapolation of the results to actual industrial settings is somewhat risky (Sanders and McCormick, 1993). Nonetheless, on the negative side, noise is known to mask task-related acoustic cues as well as to cause distraction and disruption of "inner speech"; on the positive side, noise may at least initially heighten operator arousal and thereby improve performance on tasks that do not require substantial cognitive processing (Poulton, 1978). To obtain reliable effects of noise on performance, except on tasks that rely heavily on short-term memory, the level of noise must be fairly high, usually 95 dBA or greater. Tasks that are simple and repetitive often show no deleterious performance effects (and sometimes improvements) in the presence of noise, whereas difficult tasks that rely on perception and information processing on the part of the operator will often exhibit performance degradation (Sanders and McCormick, 1993). It is generally accepted that unexpected or aperiodic noise causes greater degradation than predictable, periodic, or continuous noise, and the startle response created by sudden noise can be disruptive.

#### 6.1.2  Nonauditory Health Effects

Noise has been linked to physiological problems other than those of the hearing sense, including hypertension, heart irregularities, extreme fatigue, and digestive

disorders. Most physiological responses of this nature are symptomatic of stress-related disorders. Because the presence of high noise levels often induces other stressful feelings (such as sleep disturbance and interference with conversing in the home and fear of missing oncoming vehicles or warning signals on the job), there are second-order effects of noise on physiological functioning that are difficult to predict. The reader is referred to Kryter (1994) for a detailed discussion of nonauditory health effects of noise.

## 6.2 Annoyance Effects of Noise

Noise has frequently given rise to vigorous complaints in many settings, ranging from office environments to aircraft cabins to homes. Such complaints are manifestations of what is known as noise-induced *annoyance*, which has given rise to a host of products, such as white/pink noise generators for masking undesirable noise sources, noise-canceling headsets, and noise barriers for reducing sound propagation over distances and through walls. In the populated community, noise is a common source of disturbance, and for this reason many communities, both urban and rural, have noise ordinances and/or zoning restrictions which regulate the maximum noise levels that can result from certain sources and/or in certain land areas. In communities that have no such regulations, residents who are disturbed by noise sources such as industrial plants or spectator events often have no other recourse than to bring civil lawsuits for remedy (Casali, 1999). The principal rationale for limiting noise in communities is to reduce sleep and speech interference and to avoid annoyance (Driscoll et al., 2003). Some of the measurement units and instrumentation discussed in this chapter are useful for community and other noise annoyance applications, while more detailed information on the subject may be found in Fidell and Pearsons (1997), Casali (1999), and Driscoll et al. (2003).

## 6.3 Loudness and Related Scales of Measurement

One of the most readily identified aspects of a sound or noise and one that relates to a majority of complaints, be it a theater actor's voice which is too quiet or a background noise which is too intense, is that of *loudness*. As discussed above, the decibel is useful for quantifying the amplitude of a sound on a physical scale; however, it does not yield an absolute or relative basis for quantifying the human *perception* of sound amplitude, commonly called *loudness*. However, there are several psychophysical scales that are useful for measuring loudness, the two most prominent being *phons* and *sones*.

### 6.3.1 Phons

The decibel level of a 1000-Hz tone, which is judged by human listeners to be equally loud to a sound in question, is the phon level of the sound. The phon levels of sounds of different intensities are shown in Figure 3*a*; this family of curves is referred to as the *equal-loudness contours*. On any given curve, the combinations of

sound level and frequency along the curve produce sound experiences of equal loudness to the normal-hearing listener. Note that at 1000 Hz on each curve the phon level is equal to the decibel level. The threshold of hearing for a young, healthy ear is represented by the 0-phon-level curve. The young, healthy ear is sensitive to sounds between about 20 and 20,000 Hz, although, as shown by the curve, it is not equally sensitive to all frequencies. At low- and midlevel sound intensities, low-frequency and to a lesser extent high-frequency sounds are perceived as less intense than sounds in the range 1000–4000 Hz, where the undamaged ear is most sensitive. But as phon levels move to higher values, the ear becomes more linear in its loudness perception for sounds of different frequencies. It is because the ear exhibits this nonlinear behavior that the frequency-weighting responses for dBA, dBC, and so on, were developed, as discussed in Section 3.1.1.

### 6.3.2 Sones

Although the phon scale provides the ability to equate the loudness of sounds of various frequencies, it does not afford an ability to describe how much louder one sound is than another. For this, the *sone* scale is needed (Stevens, 1936). One sone is defined as the loudness of a 1000-Hz tone of 40-dB SPL. In relation to 1 sone, 2 sones are twice as loud, 3 sones are three times as loud, $\frac{1}{2}$ sone is half as loud, and so on. Phon level ($L_P$) and sones are related by the following formula for sounds at or above a 40-phon level:

$$\text{Loudness (sones)} = 2^{(L_P - 40)/10} \qquad (20)$$

According to equation (20), 1 sone equals 40 phons and the number of sones doubles with each 10-phon increase above 40; therefore, it is straightforward to conduct a comparative estimate of loudness levels of sounds with different decibel levels. The rule of thumb is that each 10-dB increase in a sound (i.e., one that is above 40 dB to begin with) will result in a doubling of its loudness. For instance, a home theater room that is currently at 50 dBA may be comfortable for listening to movies and classical music. However, if a new air-conditioning system increases the noise level in the room by 10 dBA, the occupants will experience a perceptual doubling of loudness and will probably complain about the interference with speech and music in the room. Once again, the compression effect of the decibel scale yields a measure that does not reflect the much larger influence that an increase in sound level will have on the human perception of loudness.

***Precise Calculation of Sone Levels by the Stevens Method*** It should be evident that sone levels can be calculated directly from psychological measurements in phons [per equation (20)] but not from physical measurements of SPL in decibels without special conversions. This is because the phon-based loudness and SPL relationship changes as a function of the sound frequency, and the magnitude of this change depends on the intensity of the sound. The *Stevens*

*method*, also known as the *ISO spectral method*, is fully described in Rossing (1990). Briefly, this method requires measurement of the dB (linear) level in 10 standard octave or 1/3-octave bands, with centers at 31, 63, 125, 250, 500, 1000, 2000, 4000, 8000, and 16000 Hz. Then, for each band measurement, the loudness index, $S_i$, is computed from Figure 9 as follows:

$$\text{Loudness level (sones)} = S_{max} + 0.3 \sum S_i \quad (21)$$

where $S_i$ is the loudness index from Figure 9, $S_{max}$ is the largest of the loudness indices, and $\sum S_i$ is the sum of the loudness indices for all bands except $S_{max}$. Using this "precise" method, the effect is to include the loudest band of noise at 100%, while the totality of the other bands is included at 30%. Obviously, because the noise

must be measured in octave or 1/3-octave bands, the method is measurement intensive and requires special instrumentation (i.e., a real-time spectrum analyzer).

***Approximation of Sone Levels from dBA***    In contrast to the Stevens method, the loudness of a sound in sones can be computed from dBA values, albeit with substantially less spectral precision. In this method, only a SLM (as compared to a spectrum analyzer) is needed, and measurements are captured in dBA. Then 1.5 sones is equated to 30 dBA, and the number of sones is doubled for each 10-dBA increase over 30 dBA. For example, 40 dBA = 3 sones, 50 dBA = 6 sones, 55 dBA = 8 sones, 60 dBA = 12 sones, 65 dBA = 16 sones, 70 dBA = 24 sones, 75 dBA = 32 sones, 80 dBA = 48 sones, 85 dBA = 64 sones, and 90 dBA = 96 sones (Rossing, 1990). This method is particularly accurate
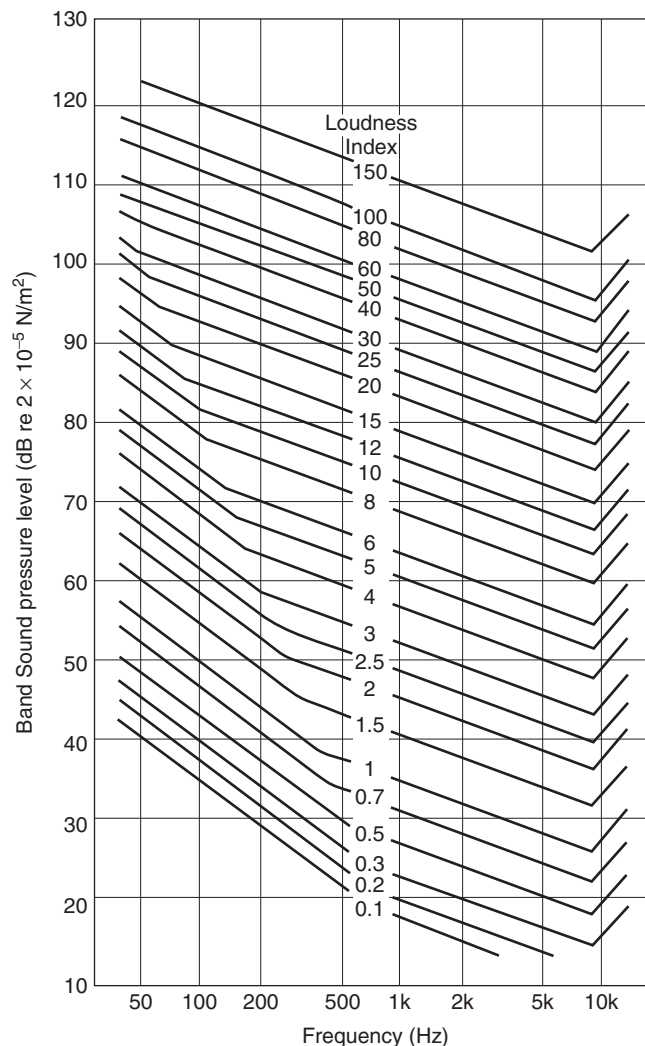


**Figure 9**    Chart for calculating loudness indices from decibel levels in various frequency bands for use in computing sone levels. (Adapted with permission from Rossing, 1990.)

at low to moderate sound levels since the ear responds in similar sensitivity to the A-weighting curve at these levels.

***Practical Applications of the Sone*** Despite its practicality, the sone scale is not widely used (an exception is that household ventilation fans typically have voluntary sone ratings). However, it is the most useful scale for comparing different sounds as to their loudnesses as perceived by humans. Given its interval qualities, the sone is more useful than decibel measurements when attempting to compare the loudness of different products' emissions; for example, a vacuum cleaner that emits 60 sones is twice as loud as one of 30 sones. The sone also is useful in conveying sound loudness experiences to lay groups. An example of such use for illustrating the perceptual impacts of a community noise disturbance (automobile racetrack) to a civil court jury may be found in Casali (1999).

### 6.3.3 Modifications of the Sone

A modification of the sone scale (Mark VI and subsequently Mark VII sones) was proposed by Stevens (1972) to account for the fact that most real sounds are more complex than pure tones. Utilizing the general form equation (22) below, this method incorporates octave-band, 1/2-octave band, or 1/3-octave band noise measurements and adds to the sone value of the most intense frequency band a fractional portion of the sum of the sone values of the other bands ($\sum S$):

$$\text{Loudness (sones)} = S_m + k\left(\sum S - S_m\right) \qquad (22)$$

where $S_m$ is the maximum sone value in any band, $k$ is a fractional multiplier that varies with bandwidth (octave, $k = 0.3$; 1/2-octave, $k = 0.2$; 1/3-octave, $k = 0.15$), and $\sum S$ is the sum of the sone values of the other bands.

### 6.3.4 Zwicker's Method of Loudness

The concept of the critical band for loudness formed the basis for Zwicker's (1960) method of loudness quantification. The critical band is the frequency band within which the loudness of a band of continuously distributed sound of equal SPL is independent of the width of the band. The critical bands widen as frequency increases. A graphical method is used for computing the loudness of a complex sound based on critical band results obtained and graphed by Zwicker. The noise spectrum is plotted and lines are drawn to depict the spread of a masking effect. The result is a bounded area on the graph which is proportional to total loudness. The method is relatively complex, and Zwicker (1960) should be consulted for computational detail.

### 6.3.5 Noisiness Units

As descriptive terms, *noisiness* and *loudness* are related but not synonymous. Noisiness can be defined as the "subjective unwantedness" of a sound. Perceived noisiness may be influenced by a sound's loudness, tonality,

duration, impulsiveness, and variability (Kryter, 1994). Whereas a low level of loudness might be perceived as enjoyable or pleasing, a low level of unwantedness (i.e., noisiness) is by definition undesirable. Equal-noisiness contours, analogous to equal-loudness contours, have been developed based on a unit (analogous to the phon) called the *perceived noise level* ($\text{PN}_{\text{dB}}$), which is the SPL in decibels of a 1/3-octave band of random noise centered at 1000 Hz, which sounds equally noisy to the sound in question. Also, an N (later D) SLM weighting curve was developed for measuring the perceived noise level of a sound. A subjective noisiness unit analogous to the sone, the *noy*, is used for comparing sounds as to their relative noisiness. One noy is equal to 40 $\text{PN}_{\text{dB}}$, and 2 noys are twice as noisy as 1 noy, 5 noys are five times as noisy, and so on. Similar to the behavior of sones as discussed above for loudness, an increase of about 10 $\text{PN}_{\text{dB}}$ is equivalent to a doubling of the perceived noisiness of a sound.

## 7 SIGNAL DETECTION AND SPEECH COMMUNICATIONS IN NOISE

### 7.1 General Concepts in Signal and Speech Audibility

#### 7.1.1 Signal-to-Noise Ratio Influence

One of the most noticeable effects of noise is its interference with speech communications and the hearing of nonverbal signals. Operators often complain that they must shout to be heard and that they cannot hear others trying to communicate with them. Similarly, noise interferes with the detection of signals such as alarms for general area evacuation and warnings in buildings, annunciators, on-equipment alarms, and machine-related sounds which are relied upon for feedback to industrial workers. In a car or truck, the hearing of external signals, such as emergency vehicle sirens or train horns or in-vehicle warning alarms or messages, may be compromised by the ambient noise levels. The ratio (actually the signed algebraic difference) of the speech or signal level to the noise level, termed the *signal (or speech)-to-noise ratio* (S/N), is a critical parameter in determining whether speech or signals will be heard in noise. A S/N value of +5 dB means that the signal is 5 dB greater than the noise; a S/N value of −5 dB means that the signal is 5 dB lower than the noise.

#### 7.1.2 Masking and Masked Threshold

Technically, *masking* is defined as the increase (in decibels) of the threshold of a desired signal or speech (the *masked sound*) to be raised in the presence of an interfering sound (the *masking sound* or *masker*). For example, in the presence of noisy traffic alongside a busy street, an auditory pedestrian crossing signal's volume must be sufficiently higher than the traffic noise level to enable a pedestrian to hear it, whereas a lower volume will be audible (and possibly more comfortable) when no traffic is present. It is also possible for one signal to mask another signal if both are active at the same time. The *masked threshold* is often defined

in psychophysical terms as the SPL required for 75% correct detection of a signal when that signal is presented in a two-interval task wherein, on a random basis, one of the two intervals of each task trial contains the signal and the noise and the other contains only noise. In a controlled laboratory test scenario, a signal that is about 6 dB above the masked threshold will result in nearly perfect detection performance (Sorkin, 1987). In the remainder of this chapter, various aspects of the masking phenomenon are discussed and methods for calculating a masked signal threshold or, in the case of speech, estimates of intelligibility are presented. Throughout, it is important to remember that the masked threshold is, in fact, a *threshold*; it is *not* the level at which the signal is *clearly* audible. For the ensuing discussion, a functional definition of an auditory threshold is the SPL at which the stimulus is *just audible* to a person listening intently for it in the specified conditions. If the threshold is determined in "silence," as is the case during an audiometric examination, it is referred to as an *absolute threshold*. If, on the other hand, the threshold is determined in the presence of noise, it is referred to as a *masked threshold*.

### 7.2 Analysis of Signal Detectability in Noise

Fundamentally, *detection* of an auditory signal is prerequisite to any other function performed on or about that signal, such as *discrimination* of it from other signals, *identification* of its source, *recognition* of its intended meaning or urgency, *localization* of its placement in azimuth, elevation, and distance, and/or *judgment* of its speed or approach or retreat. Although the S/N ratio is one of the most critical parameters that determine a signal's detectability in a noise, there are many other factors as well. These include the spectral content of the signal and noise (especially in relation to the *critical bandwidth*), temporal characteristics of signal and noise (especially in relation to the contrast between them), duration of the signal's presentation, listener's hearing ability, demands on the listener's attention, criticality of the situation at hand, and the attenuation of hearing protectors, if used. These factors are discussed in detail in Robinson and Casali (2003) and Casali and Gerges (2006). The ensuing discussion concentrates on the most important issue of spectral content of the signal and noise and how that content impacts masking effects on audibility.

#### 7.2.1 Spectral Considerations and Masking

Generally speaking, the greater the decibel level of the background noise relative to the signal (inclusive of speech), the more difficult it will be to hear the signal. Conversely, if the level of the background noise is reduced and/or the level of the signal is increased, the masked signal will be more readily audible. In some cases, ambient noise can be reduced through engineering controls, and in the same or other cases it may be possible to increase the intensity of the signals. Although most off-the-shelf auditory warning devices have a preset output level, it is possible to increase the effective level of the devices by distributing multiple alarms or warning devices throughout a coverage area instead of relying on one centrally located device. This approach can also be used for variable-output systems such as public address loudspeakers since simply increasing the output of such systems often results in distortion of the amplified speech signal, thereby reducing intelligibility. Simply increasing the signal level without adding more sound sources can have the undesirable side effect of increasing the noise exposures of people in the area of the signal if the signal is sounded too often. If the signal levels are extremely high (e.g., over 105 dB), exposed persons could experience temporary threshold shifts or tinnitus if they are in the vicinity of the device when it is sounding. As to a working decibel range for auditory warning signals, a recommendation from the International Organization for Standardization (ISO, 2003) standard "Danger Signals for Public and Work Areas—Auditory Danger Signals" is that the signal shall not be less than 65 dBA or more than 118 dBA in the signal reception area.

One problem directly related to the level of the background noise is distortion within the inner ear. At very high noise levels, the cochlea becomes overloaded and cannot accurately transduce/discriminate different forms of acoustic energy (e.g., signal and noise) reaching it, resulting in the phenomenon known as *cochlear distortion*. In order for a signal, including speech, to be audible at very high noise levels, it must be presented at a higher level, relative to the background noise, than would be necessary at lower noise levels. This is one reason why it is best to make reduction of the background noise a high priority in occupational or other environments.

In addition to manipulating the levels of the auditory displays, alarms, warnings, and background noise, it is also possible to increase the likelihood of detection of an auditory display or alarm by manipulating its spectrum so that it contrasts with the background noise and other common workplace sounds. In a series of experiments, Wilkins and Martin (1982, 1985) found that the contrast of a signal with both the background noise and irrelevant signals was an important parameter in determining the detectability of a signal. For example, in an environment characterized by high-frequency noise such as sawing and/or planing operations in a wood mill, it might be best to select a warning device with strong low-frequency components, perhaps in the range of 500–800 Hz. On the other hand, for low-frequency noise such as might be encountered in the vicinity of large-capacity, slow-rotation ventilation fans, an alarm with strong midfrequency components in the range of 1000–1500 Hz might be a better choice.

***Upward Spread of Masking*** When considering masking of a tonal signal by a tonal noise or a narrow band of noise, masking is greatest in the immediate vicinity of the masking tone or, in the case of a band-limited noise, the center frequency of the band. (This is one reason why increasing the contrast in frequency between the signal and noise can increase the audibility of a signal.) However, the masking effect does spread out above and below this frequency, being greater at the frequencies above the frequency of the masking noise

than at frequencies below the frequency of the masking noise (Wegel and Lane, 1924; Egan and Hake, 1950). This phenomenon, referred to as the *upward spread of masking*, becomes more pronounced as the level of the masking noise increases, probably due to cochlear distortion. In practical situations, masking by pure tones would seldom be a problem, except in instances where the noise contains strong tonal components or if two warnings with similar frequencies were activated simultaneously. Although less pronounced, upward spread of masking does occur when band-limited noises are used as maskers. This phenomenon is illustrated in Figure 10.

***Masking with Broadband Noise*** A very common form of masking characteristic of typical industrial workplaces or building spaces such as conference rooms or auditoria occurs when a signal or speech is masked by a broadband noise. Experiments on broadband noise masking commonly employ white or pink noise. White noise sounds very much like static on a radio tuned to a frequency that is between broadcast channels, and it consists of equal energy by hertz, while pink noise sounds like the roar of a waterfall, consisting of a 3-dB-per-octave decrease in energy as frequency increases in hertz. In examining the masking of pure-tone stimuli by white noise, Hawkins and Stevens (1950) found that masking was directly proportional to the level of the masking noise, irrespective of the frequency of the

masked tone. In other words, if a given background white noise level increased the threshold of a 2500-Hz tone by 35 dB, the threshold of a 1000-Hz tone would also be increased by 35 dB. Furthermore, they found that for the noise levels investigated masking increased linearly with the level of the white noise, meaning that if the level of the masking noise were increased 10 dB, the masked thresholds of the tones also increased by 10 dB. The bottom line is that broadband noise such as white or pink noise, due to its inclusion of all frequencies, serves as a very effective masker of tonal signals and speech. Thus, its abatement often needs to be of high priority. On the other hand, white or pink noises may be useful as in intentional maskers to mask the distractions created by conversations and phone ringers among open-plan offices, although it is debatable whether one noise should be added to combat another noise in this sense.

### 7.2.2 Signal Audibility Analysis Method Based on Critical Band Masking

Fletcher (1940) developed what would become *critical band theory*, which has formed the fundamental basis for explaining how signals are masked by narrowband noise. According to this theory, the ear behaves as if it contains a series of overlapping auditory filters, with the bandwidth of each filter being proportional to its center frequency. When masking of pure tones by broadband noise is considered, only a narrow "critical band" of the noise centered at the frequency of the tone is effective as
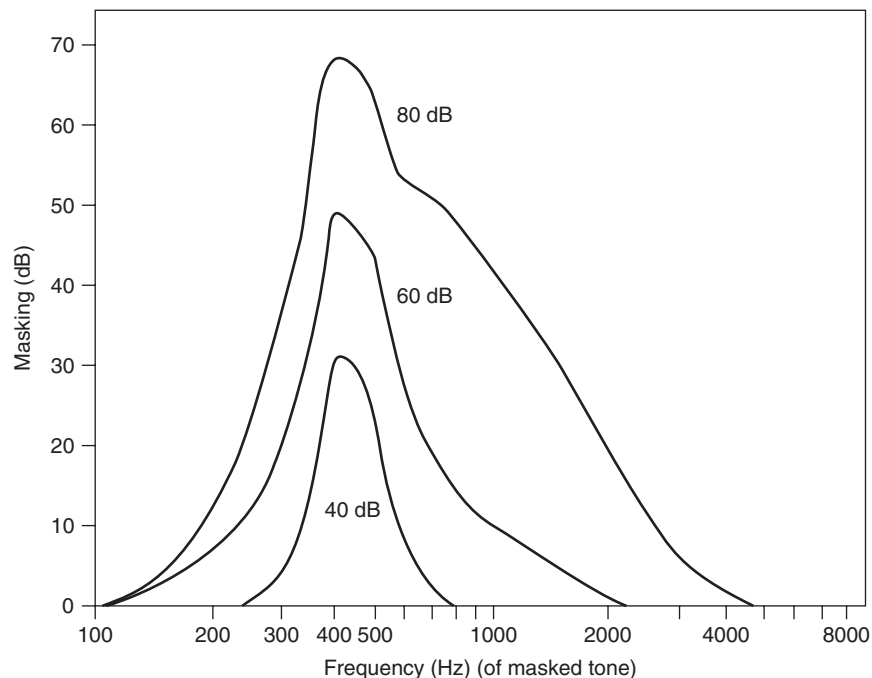


**Figure 10** Upward spread of masking of a pure tone by three levels (40, 60, and 80 dB) of a 90-Hz-wide band of noise centered at 410 Hz. The ordinate (*y* axis) is the amount (in decibels) by which the absolute threshold of the masked tone is raised by the masking noise, and the abscissa is the frequency of the masked tone. (Adapted with permission from Egan and Hake, 1950.)

a masker and the width of the band is dependent only on the frequency of the tone being masked. In other words, the masked threshold of a pure tone could be predicted simply by knowing the frequency of the tone and the spectrum level (decibels per hertz) of the masking noise, assuming that the noise spectrum is reasonably flat in the region around the tone. Thus, the masked threshold of a tone in white noise would simply be

$$L_{mt} = L_{ps} + 10 \log (BW) \qquad (23)$$

where $L_{mt}$ is the masked threshold, $L_{ps}$ is the spectrum level of the masking noise, and BW is the width of the auditory filter centered around the tone. Strictly speaking, this relationship applies only when the masking noise is flat (equal energy by hertz) and when the masked signal has a duration greater than 0.1 s. However, an acceptable approximation may be obtained for other noise conditions as long as the spectrum level in the critical band does not vary by more than 6 dB (Sorkin, 1987). In many environments, the background noise is likely to be sufficiently constant and can often be presumed to be flat in the critical band for a given signal. The exception to this assumption is a situation where the noise has prominent tonal components and/or fluctuates a great deal.

The spectrum level of the noise in each of the 1/3-octave bands containing the signal components *is not* the same as the band level measured using an octave-band or 1/3-octave-band analyzer. Spectrum level refers to the level per hertz, or the level that would be measured if the noise were measured using a filter that is 1 Hz wide. If it is assumed that the noise is flat within the bandwidth of the 1/3-octave-band filter, the spectrum level can be estimated using the equation

$$L_{ps} = 10 \log (10^{L_{pb}/10}/BW_{1/3}) \qquad (24)$$

where $L_{ps}$ is the spectrum level of the noise within the 1/3-octave band, $L_{pb}$ is the SPL measured in the 1/3-octave band in question, and $BW_{1/3}$ is the bandwidth of the 1/3-octave band, calculated by multiplying the center frequency ($f_c$) of the band by 0.232.

Finally, the bandwidth of the auditory filter can be approximated by multiplying the frequency of the masked signal/tone by 0.15 (Patterson, 1982; Sorkin, 1987). If the signal levels measured in one or more of the 1/3-octave bands considered exceed these masked threshold levels, the signal should be audible. A computational example using the critical band method appears in Robinson and Casali (2003).

### 7.2.3 Signal Audibility Analysis Method Based on ISO Standard 7731-2003(E)

The Department of Defense, National Fire Protection Association, Society of Automotive Engineers, Underwriters' Laboratories, ANSI, and ISO are examples of organizations that have promulgated standards to guide the design of auditory warning signals for specific applications, such as on-vehicle warnings, sirens,

on-firefighter alarms, evacuation alarms, and fire alarms. However, for performing an *analysis* of most any acoustic alarm as to its predicted audibility in a specific noise, perhaps the most comprehensive standard is ISO 7731-2003(E), "Danger Signals for Public and Work Areas—Auditory Danger Signals" (ISO, 2003). (This standard provides guidelines for calculation of the masked threshold of audibility but also specifies the spectral content and minimum signal-to-noise ratios (S/N) of the signals and requires special considerations for people suffering from hearing loss or those wearing HPDs.)

Application of ISO 7731 (2003) is best illustrated by an example. A warning signal that is quite common is a standard backup alarm typically found on commercial trucks and construction/industrial equipment. It has strong tonal components in the range 1000–2000 Hz and significant harmonic components at higher frequencies (Casali and Alali, 2009). The alarm has a 1-s period and a 50% duty cycle (i.e., it is "on" for 50% of its period). The levels in all other 1/3-octave bands are sufficiently below those in the bands mentioned as to be inconsequential. The levels needed for audibility of this signal will be determined for application in a hypothetical masking noise spectrum represented by its 1/3-octave and octave band levels, shown in columns 2 and 4, respectively, in Table 1.

1.  Starting at the lowest octave-band or 1/3-octave-band level available, the masked threshold ($L_{mt1}$) for a signal in that band is

$$L_{mt1} = L_{pb1} \qquad (25)$$

where $L_{pb1}$ is the SPL measured in the octave band or 1/3-octave band in question.

2.  For each successive octave-band or 1/3-octave-band filter $n$, the masked threshold ($L_{mtn}$) is the noise level in that band or the masked threshold in the preceding band less a constant, whichever is *greater*:

$$L_{mtn} = \max(L_{pbn}; L_{mtn-1} - C) \qquad (26)$$

where $C$ equals 7.5 dB for octave-band data or 2.5 dB for 1/3-octave-band data.

For an auditory signal to be *"clearly audible,"* ISO 7731 requires that at least one of the following be met: (1) the dBA level of the signal must exceed the dBA level of the ambient noise by more than 15 dB, (2) the signal level must exceed the masked threshold by at least 10 dB in at least one octave band, or (3) the signal level must exceed the masked threshold by at least 13 dB in at least one 1/3-octave band. Furthermore, the spectral content of the signal must include frequency components in the range of 500–2500 Hz, and it is recommended that there be two dominant components in the subset range of 500–1500 Hz. Furthermore, to accommodate persons with hearing loss or using hearing protection, "sufficient" signal energy below 1500 Hz is recommended.

**Table 1  Masked Threshold Calculations according to ISO 7731-2003(E) for 1/3-Octave-Band and Octave-Band Methods**

| (1) Center Frequency (Hz)[a] | (2) $\frac{1}{3}$-Octave-Band Level (dB) | (3) Masked Threshold (dB)[b] | (4) Octave-Band Level (dB) | (5) Masked Threshold (dB)[b] |
|---|---|---|---|---|
| 25 | 52.0 | 52.0 | | |
| **31.5** | 50.7 | 50.7 | 54.7 | 54.7 |
| 40 | 42.9 | 48.2 | | |
| 50 | 56.4 | 56.4 | | |
| **63** | 86.8 | 86.8 | 88.5 | 88.5 |
| 80 | 83.7 | 84.3 | | |
| 100 | 79.7 | 81.8 | | |
| **125** | 83.7 | 83.7 | 87.1 | 87.1 |
| 160 | 82.8 | 82.8 | | |
| 200 | 76.5 | 80.3 | | |
| **250** | 81.4 | 81.4 | 85.1 | 85.1 |
| 315 | 81.6 | 81.6 | | |
| 400 | 76.3 | 79.1 | | |
| **500** | 77.3 | 77.3 | 80.7 | 80.7 |
| 630 | 73.1 | 74.8 | | |
| 800 | 74.4 | 74.4 | | |
| **1,000** | 79.6 | **79.6** | 81.5 | **81.5** |
| 1,250 | 73.4 | **77.1** | | |
| 1,600 | 82.6 | 82.6 | | |
| **2,000** | 80.1 | **80.1** | 87.9 | **87.9** |
| 2,500 | 85.3 | **85.3** | | |
| 3,150 | 83.7 | 83.7 | | |
| **4,000** | 85.7 | 85.7 | 90.9 | 90.9 |
| 5,000 | 88.0 | 88.0 | | |
| 6,300 | 74.2 | 85.5 | | |
| **8,000** | 77.3 | 83.0 | 79.1 | 83.4 |
| 10,000 | 58.7 | 80.5 | | |
| 12,500 | 67.4 | 78.0 | | |
| **16,000** | 48.7 | 75.5 | 67.6 | 75.9 |
| 20,000 | 53.3 | 73.0 | | |

*Source*: ISO (2003).
[a]Frequencies in boldface type are octave-band center frequencies.
[b]Thresholds in boldface type are the masked thresholds for the signal components of the backup alarm described in the text.

While the aforementioned broadband dBA measurement is sufficient per ISO 7731, the 1/3-octave band or full-octave band procedures which are computed by equations (25) and (26) and exemplified by the data in Table 1 are preferred, due to their higher spectral precision. These procedures (unlike the aforementioned critical band procedure) presume that the auditory filter

width is equal to the 1/3-octave band or to the octave band and also takes upward spread of masking into account by comparing the level in the band in question to the level in the preceding band. For example, for the 1250-Hz row of column 3, it can be seen that the masked threshold of the previous 1/3-octave band (1000 Hz) determines, via equation (26), the masked threshold (77.1 dB) of the 1250-Hz band due to upward masking effects. The masked thresholds for each 1/3-octave band and octave band of noise for the example are shown in columns 3 and 5, respectively, in Table 1. For the purposes of the example signal (a backup alarm), only the thresholds for the 1/3-octave bands centered at 1000, 1250, 2000, and 2500 Hz and the threshold for the octave bands centered at 1000 and 2000 Hz are relevant, because these are the signal's dominant component bands, and they overlap the standard's spectral requirements noted above. (But if the signal had possessed significant energy below 1000 Hz, then the 1/3-octave bands centered at 500, 630, and 800 Hz would require attention, as would the octave band centered at 500 Hz.)

The conclusion is that if the signal levels measured in one or more of these bands exceed the calculated masked threshold levels (as indicated by boldface type), then the backup alarm is predicted to be *barely audible.* More importantly, to next determine the necessary sound level output of the alarm to render it "clearly audible" per ISO 7731, to simplify we will assume that the backup alarm's dominant frequency bands (1000, 1250, 2000, and 2500 Hz) themselves cannot change but their decibel output can be raised. Thus, based on the 1/3-octave analysis, in order for the alarm to be reliably audible, *the signal level would have to be at least the following in at least one of these four 1/3-octave bands*: centered at 1000 Hz: $79.6 + 13 = 92.6$ dB; at 1250 Hz: $77.1 + 13 = 90.1$ dB; at 2000 Hz: $80.1 + 13 = 93.1$ dB; at 2500 Hz: $85.3 + 13 = 98.3$ dB. Or, based on the octave analysis, in order for the alarm to be reliably audible, *the signal level would have to be at least the following in at least one of these two octave bands*: centered at 1000 Hz: $81.5 + 10 = 91.5$ dB; at 2000 Hz: $87.9 + 10 = 97.9$ dB. Of course, these results are based on ISO 7731's criteria for *clear audibility* and are well above the levels required for threshold audibility.

The ISO 7731 (2003) standard provides a procedure which may be used to calculate masked thresholds with and without HPDs. Calculating a protected masked threshold for a particular signal requires (1) subtracting the attenuation of the HPD from the noise spectrum to obtain the noise spectrum effective when the HPD is worn; (2) calculation of a masked threshold for each signal component using the procedures outlined in the preceding discussion, which results in the signal component levels that would be *just audible* to the listener when the HPD is worn; and (3) adding the attenuation of the HPD to the signal component thresholds to provide an estimate of the environmental (exterior to the HPD) signal component levels that would be required to produce the under-HPD threshold levels calculated in step 2. Although not difficult, this procedure does

require a reasonably reliable estimate of the actual attenuation provided by the HPD. The manufacturer's data supplied with the HPD are unsuitable for this purpose because they overestimate the real-world performance of the HPD, as explained in Section 4.2.3. Furthermore, if a 1/3-octave band masking computation is desired, the manufacturer's attenuation data, which are available for only nine selected 1/3-octave bands, are insufficient for the computation. Finally, the standard does not take the listener's hearing level into account. It is simply assumed that if the calculated masked thresholds are above the listeners' absolute thresholds, the signals should be audible, and if hearing impairment is at issue, signals should include sufficient energy below 1500 Hz.

As alluded to previously, use of the ISO 7731 standard for prediction of masked threshold for auditory signals is not limited to the octave and 1/3-octave calculations discussed herein, although the latter is the most precise method. As a less precise method (which is advocated by this author only as a last resort), ISO 7731 also offers a broadband analysis that can be performed by obtaining the dBA level of the ambient noise, and if the signal exceeds this level by 15 dB, it is said to be audible in most circumstances. However, this does not take into account upward masking or other spectrally specific effects, and it may result in (unnecessarily) higher signal levels than computed by either spectral technique. ISO 7731 also includes recommendations for signal temporal characteristics with repetition rates of 0.5–4 Hz, unambiguous meaning, discriminability, and addition of redundant visual signals if the ambient noise exceeds 100 dBA.

The broadband S/N recommendation of 15 dB of ISO 7731 is generally in keeping with those of auditory researchers. For example, Sorkin (1987) suggests that signal levels 6–10 dB above masked threshold are adequate to ensure 100% detectability, whereas signals which are approximately 15 dB above their masked threshold will elicit rapid operator response. He also suggests that signals more than 30 dB above the masked threshold could result in an unwanted startle response and that no signal should exceed 115 dB. [This suggested the upper limit on signal level is consistent with OSHA hearing conservation requirements (OSHA, 1983), which prohibits exposure to continuous noise levels greater than 115 dBA.] These recommendations are in line with those of other authors (Deatherage, 1972; Wilkins and Martin, 1982).

Masked thresholds estimated via ISO 7731 are not necessarily exact, nor are they intended to be. The "clearly audible" design estimates represent conservative estimates for a large segment of the population representing a wide range of hearing levels for nonspecific noise environments and signals. Further information on the design of auditory warnings and alarms, including relevant technical standards and guidelines, appears in Robinson and Casali (2003). Before embarking on the design of any auditory signal that is associated with a safety issue, the designer should first determine if there are any standards or regulations that have bearing. In this area of acoustics, the coverage of consensus standards is fairly broad and in depth.

## 7.3 Analysis of Speech Intelligibility in Noise

Many of the concepts presented above that relate to the masking of nonspeech signals by noise apply equally well to the masking of speech, so they will not be repeated in this section. However, for the spoken message, the concern is not simply audibility or detection but, rather, intelligibility. The listener must understand *what* was said, not simply know that something was said. Furthermore, speech is a very complex broadband signal whose components are not only differentially susceptible to noise but also highly dependent on vocal effort, the gender of the speaker, and the content and context of the message. In addition, other factors must be considered, such as the effects of HPD use by the speaker and/or listener, hearing loss of the listener, or speech signal degradation occurring in a communications system.

### 7.3.1 Speech-to-Noise Ratio Influence

Similar to the case with nonverbal signals, the signed difference between the speech level and the background noise level is referred to as the *speech-to-noise* (S/N) *ratio*. The speech level referred to is usually the long-term rms level measured in decibels. When background noise levels are between 35 and 110 dB, an S/N ratio of 12 dB is usually adequate to reach a normal-hearing person's threshold of intelligibility (Sanders and McCormick, 1993); however, it is quite impossible for anyone to sustain the vocal efforts required in the higher noise levels without electronic amplification (i.e., a public address system). The *threshold of intelligibility* is defined as the level at which the listener is just able to obtain without perceptible effort the meaning of almost every sentence and phrase of continuous speech (Hawkins and Stevens, 1950, p. 11); essentially, this is 100% intelligibility. Intelligibility decreases as S/N decreases, reaching 70–75% (as measured using phonetically balanced words) at an S/N of 5 dB, 45–50% at an S/N of 0 dB, and 25–30% at an S/N of −5 dB (Acton, 1970).

At least in low to moderate noise levels, people seem to modulate their vocal effort automatically, using the *Lombard reflex*, to maintain S/N ratios in increasing background noise so that they can communicate with other people. However, there is an upper limit to this ability, and speech levels cannot be maintained at more than 90 dB for long periods (Kryter, 1994). Since a relatively high S/N ratio (12 dB or so) is necessary for reliable speech communications in noise, it should be obvious that in high noise levels (greater than about 75–80 dB), unaided speech cannot be relied upon except for short durations over short distances. Furthermore, since speech levels for females tend to be about 2–7 dB less than for males, depending on vocal effort, the female voice is at a disadvantage in high levels of background noise.

An additional factor which impinges on the amplitude modulation of one's own voice is that of the *occlusion effect* (Stenfelt and Reinfeldt, 2005; Casali, 2010a), which results when the ear canal is occluded, as with an earplug for hearing protection or with a custom-molded,

in-the-ear hearing aid. The occlusion effect results from an enhancement of internal bodily conduction of sound that is caused by occlusion of the ear canal and its resultant attenuation of the air conduction pathway, as compared to that which occurs with the open ear where both air conduction and bone conduction feedback are present (Stenfelt and Reinfeldt, 2005). Compared to one's own voice levels in the canal of an open ear, measurably higher SPLs result within an occluded canal; therefore, the auditory feedback to the occluded person is that his or her own voice sounds louder than with a normally open canal, and this affects amplitude modulation during vocal utterances. Therefore, the speech output is generally lower in level when the ear canal is occluded, the opposite of the desired effect when the speaker is in noise, as is usually the case with hearing protection. The effect is maximized as the entrapped volume of the ear canal is at its largest, such as with a shallowly inserted earplug, and in addition to making one's own voice sound louder, it often renders the voice as sounding like it has more bass and resonance. Also, sounds of bodily origin, such as breathing and footfalls, are heard as unnaturally loud (Casali, 2010a). This can cause particular problems for soldiers who may be wearing shallow-insertion, passive earplugs and who are trying to whisper and otherwise maintain covert operations.

### 7.3.2 Speech Bandwidth Influence

The speech bandwidth extends from 200 to 8000 Hz, with male voices generally having more energy than female voices at the low frequencies (Kryter, 1974); however, the region between 600 and 4000 Hz is most critical to intelligibility (Sanders and McCormick, 1993). This also happens to be the frequency range at which most auditory alarms are presented, providing an opportunity for the direct masking of speech by an alarm or warning. Therefore, speech communications in the vicinity of an activated alarm can be difficult.

Consonant sounds, which are generally higher than vowel sounds in the 600–4000 Hz bandwidth, are also more critical than vowels to intelligibility. This fact renders speech differentially susceptible to masking by band-limited noise, depending on the level of the noise. At low levels, bands of noise in the mid- to high-frequency ranges mask consonant sounds directly, thus impairing speech intelligibility more than would low-frequency sounds presented at similar levels. However, at high levels, low-frequency bands of noise can also adversely affect intelligibility due to upward spread of masking into the critical speech bandwidth.

When electronic transmission/amplification systems are used to overcome problems associated with speech intelligibility, it is important to understand that the systems themselves may exacerbate the problem if they are not designed properly. Most industrial telecommunications systems [i.e., intercoms, telephones, personal assistant (PA) systems] do not transmit the full speech bandwidth, nor do they reproduce the entire dynamic range of the human voice. To reduce costs and simplify the electronics, such systems often filter the signal and pass (transmit) only a portion of the speech bandwidth (e.g., the telephone passband is generally 300–3600

Hz). If the frequencies above 4000 Hz or the frequencies below 600 Hz are filtered out (not transmitted), there is little negative impact on speech intelligibility, even though the voice may not appear as natural or as pleasing as when its full bandwidth is available. However, if the frequencies between 1000 and 3000 Hz are filtered out of the signal, intelligibility is severely impaired (French and Steinberg, 1947).

In addition to filtering the speech signal, it is possible to clip the speech peaks so that the full dynamic range of a speaker's voice is not transmitted to a listener. This clipping may be intentional on the part of the designer to reduce the cost of the system or it may be an artifact of the amplitude distortion caused by an overloaded amplifier. Either way, the effects on intelligibility are the same. Since the speech peaks contain primarily vowel sounds and intelligibility relies predominantly on the recognition of consonants, there is little loss in intelligibility due strictly to peak clipping. However, if the clipping is caused by distortion within the amplifier, there may be ancillary distortion of the speech signal in other ways that could affect intelligibility adversely.

### 7.3.3 Acoustic Environment Influence

The acoustic environment (room volume, distances, barriers, reverberation, etc.) can also have a dramatic effect on speech intelligibility. This is a complex subject in sound propagation and architectural acoustics, and a detailed treatment is beyond the scope of this chapter, but more information may be found in Kryter (1974, 1994) and Harris (1991). One fairly obvious point is that as the distance between the listener and the speech source (person or loudspeaker) increases, the ability to understand the speech can be affected adversely if the S/N ratio decreases sufficiently. In the same vein, barriers in the source–receiver path can create shadow zones in which the S/N ratio is insufficient for reliable intelligibility. Finally, speech intelligibility decreases linearly as reverberation time increases. Reverberation time ($RT_{60}$) is defined for a given space as the time (in seconds) required for a steady sound to decay by 60 dB from its original value after being shut off. Each 1-s increase in reverberation time will result in a loss of approximately 5% in intelligibility (Fletcher, 1953). Thus, rooms with long reverberation times, producing an echo effect, will not provide good conditions for speech reception.

### 7.3.4 Speech Intelligibility Analysis Method Based on the Preferred Speech Interference Level

There are a number of techniques to analyze or in some cases to predict accurately the intelligibility of a speech communications system based on empirical measurements of an incident noise and, in some cases, additional measurements of the system's speech output, be it amplified or live unamplified voice. A variety of techniques are covered in Sanders and McCormick (1993) and Kryter (1994). However, one of the better known techniques, the preferred speech interference level (PSIL), which involves only measurements of the noise and is straightforward to administer (although limited in its predictive ability), warrants discussion here.
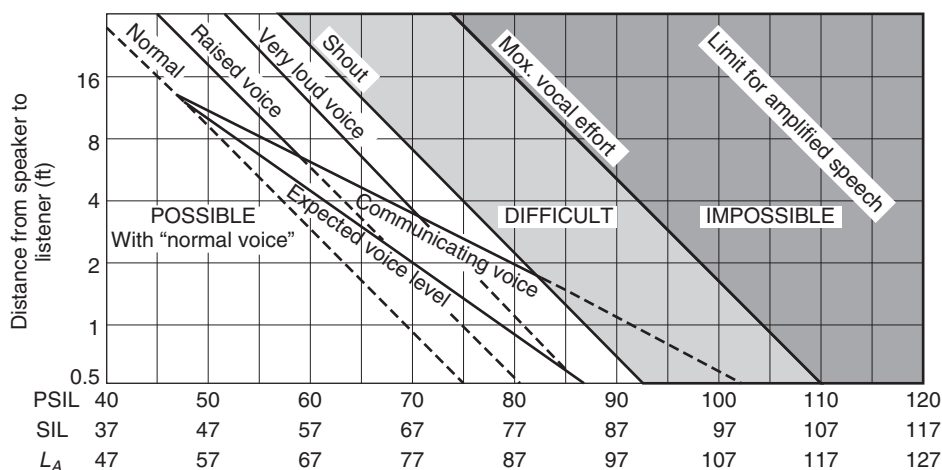
**Figure 11** Relationship among PSIL, speech difficulty, vocal effort, and speaker–listener separation. (Adapted with permission from Sanders and McCormick, 1993.)

The PSIL is the arithmetic average of the noise levels measured in three octave bands centered at 500, 1000, and 2000 Hz (Peterson and Gross, 1978). It is most useful when the spectrum of the background noise is relatively flat and intended only as an indication of whether or not there is likely to be a communications problem, not as a predictor of intelligibility. If the background noise is not flat, is predominated by or contains strong tonal components, or fluctuates a great deal, the utility of the PSIL is lessened. As an example of PSIL application, the hypothetical octave-band noise spectrum presented earlier in column 4 of Table 1 can be used. The PSIL for this spectrum is $(80.7 + 81.5 + 87.9)/3 = 83$. With this information, Figure 11 can be consulted to determine how difficult verbal communication is likely to be in this noise. At a PSIL of 83, verbal communications will be "difficult" at any speaker–listener distance greater than about 18 in. Even at closer distances, a "raised" or "very loud" voice must be used. If octave-band levels are not available, the A-weighted sound level may also provide rough guidance concerning the speech-interfering effects of background noise, also shown in Figure 11. In summary, the PSIL is a useful, simple tool for estimating the degree of difficulty that can be expected when verbal communications are attempted in a steady, flat background noise.

### 7.3.5 Speech Intelligibility Analysis Method Based on the Speech Intelligibility Index (SII) and Extended SII

In contrast to the PSIL, a more precise analytical prediction of the interfering effects of noise on speech communications may be conducted using the speech intelligibility index (SII) technique defined in ANSI S3.5-1997(R2007) (ANSI, 2007b). Essentially, this well-known standardized technique utilizes a weighted sum of the S/N ratios in specified frequency bands to compute an SII score ranging between 0.0 and 1.0, with higher scores indicative of greater predicted speech intelligibility. While the end result is an SII score on a simple scale or 0.0–1.0, the process of measurement and calculation is complex by comparison to the PSIL. However, the SII is more accurate than the PSIL, broader in its coverage, and can account for many additional factors, such as speaker vocal effort, room reverberation, monaural versus binaural listening, hearing loss, varying message content, hearing protector effects, communications system gain, and the existence of external masking noise.

Four calculation methods are available with the SII: the critical band method (most accurate), the 1/3-octave-band method, the equally contributing critical band method, and the octave-band method (least accurate) (ANSI, 2007b). At a minimum, the calculations require knowledge of the spectrum level of the speech and noise as well as the listeners' hearing thresholds. Where speech spectrum level(s) are unavailable or unknown, the standard offers guidance in their estimation. Although quite flexible in the number and types of conditions to which it can be applied, application of the standard is limited to natural speech, otologically normal listeners with no linguistic or cognitive deficiencies, and situations that do not include sharply filtered bands of speech or noise. Software programs for calculation of the SII may be obtained at http://www.sii.to/html/programs.html.

The SII "score" actually represents the proportion of the speech cues that would be available to the listener for "average speech" under the noise/speech conditions for which the calculations were performed. Hence, intelligibility is predicted to be greatest when the SII = 1.0, indicating that all of the speech cues are reaching the listener, and poorest when the SII = 0.0, indicating that none of the speech cues are reaching the listener. The general steps used in calculating the SII and estimating intelligibility are beyond the scope of this chapter, but

they may be found in the standard itself, ANSI S3.5-1997(R2007) (ANSI, 2007b), or in paraphrased terms with examples in Robinson and Casali (2003).

A limitation of the SII model is that it employs a long term speech and noise spectrum and thus was designed and validated for masking noises that are stationary, that is, invariant over time. However, in fluctuating noises, speech intelligibility for the normal hearer may be different, and often better, than that in stationary noises because the listener can benefit from the relatively quiet periods in the noise. In an effort to extend the SII model to accommodate nonstationary noises, Rhebergen and others (e.g., Rhebergen et al., 2006) developed and have recently been involved in validating the Extended SII. The basic principle is that the system's speech output and its noise environment are partitioned into segments or time frames over the time course of the speech presentation. Within each frame, the SII is calculated to provide an instantaneous value; next, the instantaneous SII values are averaged to produce an SII for that particular speech-in-noise condition (Rhebergen et al., 2006). Extant data for these metrics indicate that the extended SII provides accurate predictions for a variety of time-variant noise conditions; however, more validation and refinements were underway as of the publication date of this chapter. This method demonstrates promise for prediction of speech reception thresholds in fluctuating noise (which is a common masking situation), thus adding practical value to the utility of the SII that is standardized in ANSI S3.5-1997(R2007) (ANSI, 2007b).

### 7.3.6 Speech Intelligibility Experimental Test Methods

In lieu of analytical techniques such as the PSIL and the SII, both of which require spectral measurements, an alternative (or complementary) approach is to conduct an experiment to measure intelligibility for a given set of conditions with a group of human listeners. For this purpose, there exists a standard, ANSI S3.2-1989(R2009), which provides not only guidance for conducting such tests but also three alternative sets of standard speech stimuli (ANSI, 2009). The standard is intended for designers and manufacturers of communications systems and provides valuable insight into the subject of speech intelligibility and how various factors associated with the speaker, transmission path/environment, and listener can affect it. The standard accommodates empirical testing of intelligibility in the following situations: indoors or outdoors, speaking face to face or in vicinity, telephonic systems, public address systems, radio systems, and complex systems that include air, wire, wireless, fiber optics, and water transmission paths that are applied in certain military, remote, or emergency systems. Thus, ANSI S3.2-1989(R2009) can be of benefit to the human factors designers/evaluators of many types of communications systems when empirical measurement of speech intelligibility performance is necessary for evaluation, acceptance, or proving efforts.

Although space does not permit a detailed description of the procedures, the ANSI S3.2 standard's strategy involves presenting speech stimuli to a listener in an environment that replicates the conditions of concern and measuring how much of the speech message is understood. The speech stimuli may be produced by a trained talker speaking directly to the listener while in the same environment or via an intercom system. Alternatively, the materials may be recorded and presented electronically. Use of recorded stimuli and/or electronic presentation of the stimuli offers the greatest control over the speech levels presented to the listener.

### 7.4 Other Considerations for Signal Detectability and Speech Intelligibility

#### 7.4.1 Distance Effects

It cannot be overemphasized that the noise and signal levels referred to in the analysis techniques above refer to the levels measured *at the listener's location*. Measurement made at some central location or the specified output levels of the alarm or warning devices are not representative of the levels present at a given workstation and cannot be used for masked threshold calculations. In a free-field, isotropic environment, the sound level of an alarm or warning will decrease in inverse relationship to the distance from the source, in accordance with the formula

$$p_1/p_2 = d_2/d_1 \qquad (27)$$

where $p_1$ and $p_2$ are the sound pressures of the signal at distances $d_1$ and $d_2$, respectively, in micropascals or dynes per square centimeters and $d_1$ and $d_2$ are, respectively, distance 1 (near point) and distance 2 (far point) at which the signal is measured, in linear distance units, or, alternatively, where the drop between distance 1 and 2 in the SPL of the signal in decibels is given by

$$\text{SPL}_{\text{drop}} = 20 \log_{10}(d_2/d_1) \qquad (28)$$

These formulas provide accurate results in outdoor environments where there are no barriers, such as trees, or highly reflecting planes, such as paved parking lots. Indoors, the formulas will typically overestimate the drop in signal level, where reflective surfaces reinforce the signal as it propagates.

#### 7.4.2 Barrier Effects

Furthermore, buildings or other large structures in the source–receiver path can create "shadow zones" in which little or no sound is audible. It is for these reasons that the U.S. Department of Defense (1981) recommends that frequencies below 1000 Hz be used for outdoor alarms since low frequencies are less susceptible to atmospheric absorption and diffract more readily around barriers. Similar problems can be encountered indoors as well. Problems associated with the general decrease in SPL with increasing distance as well as shadow zones created by walls, partitions, screens, and machinery/vehicles must be considered. Since different materials reflect and absorb sound depending on its frequency, not only do the sound levels change from position to position, but the spectra of both the noise and signals/speech can change as well. Finally,

since most interior spaces reverberate to some degree, the designer should also be concerned with phase differences between reflected sounds, which can result in superposition effects of enhancement or cancellation of the signals and speech from location to location. It is for all these reasons that it is necessary to know the SPL at the *listener's location* when considering masked thresholds.

### 7.4.3  Hearing Protection Device Effects

HPDs are often blamed for exacerbating the effects of noise on the audibility of speech and signals, although, at least for people with normal hearing, protectors may actually facilitate hearing in some noisy situations. Overall, the research evidence on normal hearers generally suggests that *conventional* passive HPDs have little or no degrading effect on the wearer's under-standing of external speech and signals in ambient noise levels above about 80 dBA and may even yield some improvements, with a crossover between disadvantage and advantage between 80 and 90 dBA. However, HPDs do often cause increased misunderstanding and poorer detection (compared to unprotected conditions) in lower sound levels, where HPDs are not typically needed for hearing defense anyway but may be applied for reduction of annoyance (Casali and Gerges, 2006). In intermittent noise, HPDs may be worn during quiet periods so that when a loud noise occurs, the wearer will be protected. However, during those quiet periods, the conventional passive HPDs typically reduce hearing acuity. In certain of these cases, the family of *amplitude-sensitive augmented HPDs* may be benefi-cial, including those that provide, during quiet periods, minimal or moderate passive attenuation via acoustic valving systems (or, alternatively, more amplification of external sounds via electronically-modulated sound transmission through the HPD) but then also provide increased passive attenuation (or less amplification) as the incident noise increases. However, the real performance effects of these and other augmented HPDs are very situation-specific, and the interested reader is pointed to the reviews in Casali (2010a,b).

Noise- and age-induced hearing losses generally occur in the high-frequency regions first, and for those so impaired, the effects of HPDs on speech perception and signal detection are not clear-cut. Due to their already elevated thresholds for mid- to high-frequency speech sounds being raised further by the protector, hearing-impaired persons are usually disadvantaged in their hearing by conventional HPDs. Although there is no consensus across studies, certain reviews have con-cluded that sufficiently hearing-impaired persons will usually experience additional reductions in communica-tions abilities with conventional HPDs worn in noise. In some instances, HPDs with electronic hearing-assistive circuits, sometimes called *sound-transmission or sound restoration HPDs*, can be offered to hearing-impaired persons to determine if their hearing, especially in quiet to moderate noise levels below about 85 dBA, may be improved with such devices while still receiving a measure of protection. However, as noted above, the realized benefits of such devices are very dependent

upon the particular signal-in-noise situation as well as the individual's particular hearing loss (Casali, 2010a,b).

Conventional passive HPDs cannot differentiate or selectively pass speech or nonverbal signal (or speech) energy versus noise energy at a given frequency. Therefore, conventional HPDs do not improve the S/N ratio in a given frequency band, which is the most important factor for achieving reliable signal detection or intelligibility. Conventional HPDs attenuate high-frequency sound more than low-frequency sound, thereby attenuating the power of consonant sounds that are important for word discrimination as well as most warning signals, both of which lie in the higher frequency range, while also allowing low-frequency noise through. Thus, the HPD may enable an associated upward spread of masking to occur if the penetrating noise levels are high enough. Certain augmented HPD technologies help to overcome the weaknesses of conventional HPDs as to low-frequency attenuation in particular; these include the aforementioned *active noise reduction* (ANR) devices, which via electronic phase-derived cancellation of noises below about 1000 Hz improve the low-frequency attenuation of passive HPDs. Concomitant benefits of ANR-based HPDs can include the reduction of upward spread of masking of low-frequency noise into the speech and warning signal bandwidths, as well as reduction of noise annoyance in certain environments that are dominated by low frequencies, such as jet aircraft cockpits and passenger cabins (Casali et al., 2004; Casali, 2010b).

In any case, by far the most commonly applied HPDs are relatively simple, conventional products for which the paramount objective is the passive attenuation of noise and thus the prevention of noise-induced hearing loss. However, as noted throughout this chapter, there are many workplace, military, and leisure situations wherein noise may be a hazard to the ears, but there also exists a critical need to hear external signals and/or speech and, in general, to maintain one's situation awareness. It is for these reasons that more design team efforts, combining the expertise of human factors engineers with those of acousticians and audiologists, need to be aimed directly at improving hearing pro-tector capabilities for signal detection, identification, localization, and communications along with providing ample protective effectiveness against noise hazards.

### 7.4.4  Hearing-Aided Users

People with a hearing loss sufficient to require the use of hearing aids are already at a disadvantage when attempting to hear auditory alarms, warnings, or speech, and this disadvantage is exacerbated when noise levels are high. Activation of hearing aids in high levels of noise so as to improve hearing of speech or signals can increase the risk of additional damage to hearing due to amplification of the ambient noise (Humes and Bess, 1981). But shutting off the hearing aids increases the chance that the signals will be missed, and since it has been shown that vented hearing aid inserts do not function well as hearing protectors (Berger, 2003b), there is still a risk of further hearing damage by doing so. Recommendations for accommodating hearing-aided

users in the workplace appear in Robinson and Casali (2003).

## 7.5 Summary of Guidance for Reducing Effects of Noise on Signals and Speech

The following principles regarding masking effects on nonverbal signals and speech are offered as a summary for general guidance:

1. Due to *direct masking*, the greatest increase in masked threshold occurs for nonverbal signal frequencies that are equal or near the predominant frequencies of the masking noise. Therefore, warning signals should not utilize tonal frequencies equivalent to those of the masker. Preferably, the signal should contain energy in the most sensitive range of human hearing, approximately 1000–4000 Hz, unless the noise energy is intense at these frequencies.

2. If the signal and masker are tonal in nature, the primary masking effect is at the fundamental frequency of the masker and at its harmonics. For instance, if a masking noise has primary frequency content at 1000 Hz, this frequency and its harmonics (2000, 3000, 4000, etc.) should be avoided as signal frequencies.

3. The greater the SPL of the masker, the more the increase in masked threshold of the signal. A general rule of thumb is that the S/N ratio at the listener's ear should at a minimum be about 15 dB above the masked threshold for reliable signal detection. However, in noise levels above about 80 dBA, the signal levels required to maintain a S/N ratio of 15 dB above masked threshold may increase the hearing exposure risk, especially if signal presentation occurs frequently. Therefore, if lower S/N values become necessary, it is best to design contrasting signals which are unlike the masker in frequency and have modulated or alternating frequencies to grab attention.

4. Warning signals should not exceed the masked threshold by more than 30 dB to avoid verbal communications interference and operator annoyance (Sorkin, 1987).

5. As the SPL of the masker increases, the primary change in the masking effect is that it spreads upward in frequency, often causing signal frequencies which are higher than the masker to be missed (i.e., *upward masking*). Since most warning signal guidelines recommend that midrange and high-frequency signals (about 1000–4000 Hz) be used for detectability, it is important to consider that the masking effects of noise dominated by lower frequencies can spread upward and cause interference in this range. Therefore, if the noise has its most significant energy in this range, a low-frequency signal, say 500 Hz, may be necessary. However, as shown in Figure 3*a*, it must be kept in mind that the ear is not as sensitive to low frequencies, so the signal level must be set carefully to ensure reliable audibility.

6. Masking effects can also spread downward in frequency, causing signal frequencies below those of the masker to be raised in threshold (i.e., *remote masking*). The effect is most prominent at signal frequencies that are subharmonics of the masker. With typical industrial noise sources, remote masking is generally less of a problem than direct or upward masking.

7. When a signal must be localized, it is advantageous to include signal energy content below 1000 Hz and above 3000 Hz to maximize one's ability to locate the signal, taking advantage of both interaural time and interaural level differences, respectively.

8. In extremely loud environments of about 110 dB and above, nonauditory signal channels such as visual and vibrotactile should be considered as alternatives to auditory displays. They should also be used for redundancy in some lower level noises where the auditory signal may be overlooked or it blends in as the background noise varies and also where people who have hearing loss must attend to the signal.

9. Speech intelligibility in noise depends on a combination of complex factors and, as such, predictions based on simple S/N ratios should not be relied on. However, in very general terms, S/N ratios of 15 dB or higher should result in intelligibility performance above about 80% words correct for normal-hearing persons in broadband noise (Acton, 1970). Above speech levels of about 85 dBA, there is some decline in intelligibility even if the S/N ratio is held constant (Pollack, 1958). In very high noise levels, it is impractical and may pose additional hearing hazard risk to amplify the voice to maintain the high S/N ratios necessary for good intelligibility performance. The S/N ratio required for reliable intelligibility may be reduced via the use of certain techniques, such as reduction of speaker-to-listener distances, use of smaller vocabularies, provision of contextual cues in the message, use of the phonetic alphabet, and use of noise-attenuating headphones and noise-canceling microphones in electronic systems.

10. If hearing protection is necessary in a noisy environment, and if the detection, identification, discrimination, and/or localization of auditory signals is also necessary, and/or speech communications is needed, various types of augmented hearing protectors, as opposed to conventional passive devices, should be considered, albeit with great care as to their selection (Casali 2010a,b). Degradation of communications and situation awareness, while situation- and individual-specific, can create safety hazards for the individual who is occluded with an HPD.

11. Electronic speech communications systems should reproduce speech frequencies accurately in the range 500–5000 Hz, which encompasses the most sensitive range of hearing and includes

the speech sounds important for message under-standability. More specifically, because much of the information required for word discrimination lies in the consonants, which are in the higher end of the frequency range and of low power (while the power of the vowels is in the peaks of the speech waveform), the use of electronic peak clipping and reamplification of the waveform may improve intelligibility because the power of the consonants is thereby boosted relative to the vowels. Furthermore, to maintain intelligibility it is critical that frequencies in the region 1000–4000 Hz be faithfully reproduced in electronic communication systems. Filtering out of frequencies outside this range will not appreciably affect word intelligibility but will influence the quality of the speech.

12. Actual human speech typically results in higher intelligibility in noise than that of computer-generated speech, and there are also differences among synthesizers as to their intelligibility. Especially for critical message displays and annunciators, live, recorded, or digitized human speech may be preferable to synthesized speech (Morrison and Casali, 1994), and if synthesized speech is used, the selection of synthesizer must be made carefully.

## REFERENCES

Acton, W. I. (1970), "Speech Intelligibility in a Background Noise and Noise-Induced Hearing Loss," *Ergonomics*, Vol. 13, No. 5, pp. 546–554.

Ahroon, W. A. (2007), "Biomedical Design Criteria for Auditory Protection and Performance," paper presented at the Biomedical Design Criteria for Auditory Protection and Performance—AIBS Review Panel, U.S. Army Aeromedical Research Laboratory, Fort Rucker, AL, June 12–13.

American National Standards Institute (ANSI) (1974), "Method for the Measurement of Real-Ear Protection of Hearing Protectors and Physical Attenuation of Earmuffs," ANSI S3.19-1974, ANSI, New York.

American National Standards Institute (ANSI) (2006), "Specification for Sound Level Meters," ANSI S1.4-1983(R2006), ANSI, New York.

American National Standards Institute (ANSI) (2007a), "Specification for Integrating-Averaging Sound Level Meters," ANSI S1.43-1997(R2007), ANSI, New York.

American National Standards Institute (ANSI) (2007b), "Methods for the Calculation of the Speech Intelligibility Index," ANSI S3.5-1997 (R2007), ANSI, New York.

American National Standards Institute (ANSI) (2008), "Methods for Measuring the Real-Ear Attenuation of Hearing Protectors," ANSI S12.6-1997 (R2008), ANSI, New York.

American National Standards Institute (ANSI) (2009), "American National Standard Methods for Measuring the Intelligibility of Speech over Communications Systems," ANSI S3.2-1989(R2009), ANSI, New York.

Berger, E. H. (2003a), "Noise Control and Hearing Conservation: Why Do It?" in *The Noise Manual*, rev. 5th ed., E. H. Berger, L. H. Royster, J. D. Royster, D. P. Driscoll, and M. Layne, Eds., American Industrial Hygiene Association, Fairfax, VA, pp. 2–17.

Berger, E. H. (2003b), "Hearing Protection Devices," in *The Noise Manual*, rev. 5th ed., E. H. Berger, L. H. Royster, J. D. Royster, D. P. Driscoll, and M. Layne, Eds., American Industrial Hygiene Association, Fairfax, VA, pp. 379–454.

Berger, E. H., Royster, L. H., Royster, J. D., Driscoll, D. P., and Layne. M., Eds. (2003), *The Noise Manual*, rev. 5th ed., American Industrial Hygiene Association, Fairfax, VA.

Casali, J. G. (1999), "Litigating Community Noise Annoyance: A Human Factors Perspective," in *Proceedings of the 1999 Human Factors and Ergonomics Society 42nd Annual Conference*, Houston, TX, September 27-October 1, pp. 612–616.

Casali, J. G. (2010a), "Passive Augmentations in Hearing Protection Technology Circa 2010 including Flat-Attenuation, Passive Level-Dependent, Passive Wave Resonance, Passive Adjustable Attenuation, and Adjustable-Fit Devices: Review of Design, Testing, and Research." *International Journal of Acoustics and Vibration*, Vol. 15, No. 4, pp. 187–195.

Casali, J. G. (2010b), "Powered Electronic Augmentations in Hearing Protection Technology circa 2010 including Active Noise Reduction, Electronically-Modulated Sound Transmission, and Tactical Communications Devices: Review of Design, Testing, and Research." *International Journal of Acoustics and Vibration*, Vol. 15, No. 4, pp. 168–186.

Casali, J. G., and Alali, K. (2009), "Vehicle Backup Alarm Localization (or Not): Effects of Passive and Electronic Hearing Protectors, Ambient Noise Level, and Backup Alarm Spectral Content," in *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting,* San Antonio, TX, October 19-23, pp. 1617–1621.

Casali, J. G., and Gerges, S. (2006). "Protection and Enhancement of Hearing in Noise," in *Reviews of Human Factors and Ergonomics*, Vol. 2, R. C. Williges, Ed., Human Factors and Ergonomics Society, Santa Monica, CA, Chapter 7, pp. 195–240.

Casali, J. G., and Robinson, G. S. (1999), "Noise in Industry: Auditory Effects, Measurement, Regulations, and Management," in *Handbook of Occupational Ergonomics*, W. Karwowski and W. Marras, Eds., CRC Press, Boca Raton, FL, pp. 1661–1692.

Casali, J. G., Robinson, G. S., Dabney, E. C., and Gauger, D. (2004), "Effect of Electronic ANR and Conventional Hearing Protectors on Vehicle Backup Alarm Detection in Noise," *Human Factors*, Vol. 46, No. 1, pp. 1–10.

Crocker, M., Ed. (1998), *Handbook of Acoustics*, Wiley, New York.

Deatherage, B. H. (1972), "Auditory and Other Sensory Forms of Information Presentation," in *Human Engineering Guide to Equipment Design*, H. P. Van Cott and R. G. Kincade, Eds., Wiley, New York, pp. 123–160.

Defense Occupational and Environmental Health Readiness System Data Repository (DOEHRS-DR) (2007), U.S. Army Center for Health Promotion and Preventative Medicine, Aberdeen Proving Ground, MD.

Driscoll, D. P., and Royster, L. H. (2003), "Noise Control Engineering," in *The Noise Manual*, rev. 5th ed., E. H. Berger, L. H. Royster, J. D. Royster, D. P. Driscoll, and M. Layne, Eds., American Industrial Hygiene Association, Fairfax, VA, pp. 279–378.

Driscoll, D. P., Stewart, N. D., and Anderson, R. R. (2003), "Community Noise," in *The Noise Manual*, rev. 5th ed., E. H. Berger, L. H. Royster, J. D. Royster, D. P. Driscoll, and M. Layne, Eds., American Industrial Hygiene Association, Fairfax, VA, pp. 601–637.

Earshen, J. J. (1986), "Sound Measurement: Instrumentation and Noise Descriptors," in *Noise and Hearing Conservation Manual*, E. H. Berger, W. D. Ward, J. C. Morrill, and L. H. Royster, Eds., American Industrial Hygiene Association, Akron, OH, pp. 38–95.

Egan, J. P., and Hake, H. W. (1950), "On the Masking Pattern of a Simple Auditory Stimulus," *Journal of the Acoustical Society of America*, Vol. 22, No. 5, pp. 622–630.

Fidell, S. M., and Pearsons, K. S. (1997), "Community Response to Environmental Noise," in *Encyclopedia of Acoustics*, M. Crocker, Ed., Wiley, New York, pp. 1083–1091.

Fletcher, H. (1940), "Auditory Patterns," *Reviews of Modern Physics*, Vol. 12, pp. 47–65.

Fletcher, H. (1953), "The Masking of Pure Tones and of Speech by White Noise," in *Speech and Hearing in Communication*, Van Nostrand Reinhold, Princeton, N.J.

French, N., and Steinberg, J. (1947), "Factors Governing the Intelligibility of Speech Sounds," *Journal of the Acoustical Society of America*, Vol. 19, pp. 90–119.

Gerges, S., and Casali, J. G. (2007), "Ear Protectors," in *Handbook of Noise and Vibration Control*, M. Crocker (Ed.), Wiley, New York, Chapter 31, pp. 364–376.

Harris, C. M., Ed. (1991), *Handbook of Acoustical Measurements and Noise Control*, McGraw-Hill, New York.

Hawkins, J. E., and Stevens, S. S. (1950), "The Masking of Pure Tones and of Speech by White Noise," *Journal of the Acoustical Society of America*, Vol. 22, No. 1, pp. 6–13.

Humes, L. E., and Bess, F. H. (1981), "Tutorial on the Potential Deterioration in Hearing Due to Hearing Aid Usage," *Journal of Speech and Hearing Research*, Vol. 24, No. 1, pp. 3–15.

International Organization for Standardization (ISO) (2003), "Danger Signals for Public and Work Areas—Auditory Danger Signals," ISO 7731-2003(E), ISO, Geneva.

Kryter, K. D. (1974), "Speech Communication," in *Human Engineering Guide to Equipment Design*, H. P. Van Cott and R. G. Kincade, Eds., Wiley, New York, pp. 161–226.

Kryter, K. D. (1994), *The Handbook of Hearing and the Effects of Noise*, Academic, New York.

Melnick, W. (1991), "Hearing Loss from Noise Exposure," in *Handbook of Acoustical Measurements and Noise Control*, C. M. Harris, Ed., McGraw-Hill, New York, pp. 18.1–18.19.

Mine Safety and Health Administration (MSHA) (1999), "Health Standards for Occupational Noise Exposure; Final Rule," 30 CFR Part 62, 64, MSHA, Code of Federal Regulations, *Federal Register*, Washington, DC.

Morrison, H. B., and Casali, J. G. (1994), "Intelligibility of Synthesized Voice Messages in Commercial Truck Cab Noise for Normal-Hearing and Hearing-Impaired Listeners," in *Proceedings of the 1994 Human Factors and Ergonomics Society 38th Annual Conference*, Nashville, TN, October 24-28, pp. 801–805.

National Institute for Occupational Safety and Health (NIOSH) (1975), "List of Personal Hearing Protectors and Attenuation Data," Publication 76-120, NIOSH-HEW, Washington, DC, pp. 21–37.

National Institute for Occupational Safety and Health (NIOSH) (2001), *Work-Related Hearing Loss*, NIOSH, Washington, DC.

National Institutes of Health (NIH) Consensus Development Panel (1990), "Noise and Hearing Loss," *Journal of the American Medical Association*, Vol. 263, No. 23, pp. 3185–3190.

Occupational Safety and Health Administration (OSHA) (1971a), "Occupational Noise Exposure (General Industry); Final Rule," 29CFR1910.95, OSHA, Code of Federal Regulations, *Federal Register*, Washington, DC.

Occupational Safety and Health Administration (OSHA) (1971b), "Occupational Noise Exposure (Construction Industry); Final Rule," 29CFR1926.101, OSHA, Code of Federal Regulations, *Federal Register*, Washington, DC.

Occupational Safety and Health Administration (OSHA) (1980), "Noise Control: A Guide for Workers and Employers," OSHA 3048, OSHA, U.S. Department of Labor, Washington, DC.

Occupational Safety and Health Administration (OSHA) (1983), "Occupational Noise Exposure; Hearing Conservation Amendment; Final Rule," 29CFR1910.95, OSHA, Code of Federal Regulations, *Federal Register*, Washington, DC.

Ostergaard, P. (2003), "Physics of Sound and Vibration," in *The Noise Manual*, rev 5th ed., E. H. Berger, L. H. Royster, J. D. Royster, D. P. Driscoll, and M. Layne, Eds., American Industrial Hygiene Association, Fairfax, VA, pp. 19–39.

Park, M. Y., and Casali, J. G. (1991), "A Controlled Investigation of In-Field Attenuation Performance of Selected Insert, Earmuff, and Canal Cap Hearing Protectors," *Human Factors*, Vol. 33, No. 6, pp. 693–714.

Patterson, R. D. (1982), "Guidelines for Auditory Warning Systems on Civil Aircraft," Paper 82017, Civil Aviation Authority, Airworthiness Division, Cheltenham, Gloucestershire, England.

Peterson, A., and Gross, E., Jr. (1978), *Handbook of Noise Measurement*, 8th ed., General Radio Co., Concord, MA.

Peterson, A. P. G. (1979), "Noise Measurements: Instruments," in *Handbook of Noise Control*, C. M. Harris, Ed., McGraw-Hill, New York, pp. 5-1–5-19.

Pollack, I. (1958), "Speech Intelligibility at High Noise Levels: Effects of Short-Term Exposure," *Journal of the Acoustical Society of America*, Vol. 30, pp. 282–285.

Poulton, E. (1978), "A New Look at the Effects of Noise: A Rejoinder," *Psychological Bulletin*, Vol. 85, pp. 1068–1079.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2006), "Extended Speech Intelligibility Index for the Prediction of the Speech Reception Threshold in Fluctuating Noise," *Journal of the Acoustical Society of America*, Vol. 120, No. 6, pp. 3988–3997.

Robinson, G. S., and Casali, J. G. (2003), "Speech Communications and Signal Detection in Noise," in *The Noise Manual*, rev. 5th ed., E. H. Berger, L. H. Royster, J. D. Royster, D. P. Driscoll, and M. Layne, Eds., American Industrial Hygiene Association, Fairfax, VA, pp. 567–600.

Rossing, T. (1990), *The Science of Sound*, Addison-Wesley, Reading, MA.

Sanders, M. S., and McCormick, E. J. (1993), *Human Factors in Engineering and Design*, 7th ed., McGraw-Hill, New York.

Saunders, G. H., and Griest, S. (2009), "Hearing Loss in Veterans and the Need for Hearing Loss Prevention Programs," *Noise and Health*, Vol. 11, No. 42, pp. 14–21.

Sorkin, R. D. (1987), "Design of Auditory and Tactile Displays," in *Handbook of Human Factors*, G. Salvendy, Ed., McGraw-Hill, New York, pp. 549–576.

Stenfelt, S., and Reinfeldt, S. (2005), "Human Body Conduction Sensitivity in a Sound Field," Paper No. 20,022, in *Proceedings of the 2005 International Congress and Exhibition of Noise Control Engineering*, Rio de Janeiro, Brazil, August 7–10.

Stevens, S. S. (1936), "A Scale for the Measurement of a Psychological Magnitude: Loudness," *Psychological Review*, Vol. 43, pp. 405–416.

Stevens, S. S. (1972), "Perceived Level of Noise by Mark VII and Decibels (E)," *Journal of the Acoustical Society of America*, Vol. 51, No. 2, Pt. 2, pp. 575–601.

Taylor, W., Pearson, J., Mair, A., and Burns, W. (1964), "Study of Noise and Hearing in Jute Weavers," *Journal of the Acoustical Society of America*, Vol. 38, pp. 113–120.

U.S. Department of Defense (DoD) (1981), *Human Engineering Design Criteria for Military Systems, Equipment and Facilities*, MIL-STD-1472C, DoD, Washington, DC.

U.S. Environmental Protection Agency (EPA) (1979), "Noise Labeling Requirements for Hearing Protectors," 40CFR211, *Federal Register*, Vol. 44, No. 190, pp. 56130–56147.

U.S. Environmental Protection Agency (EPA) (1981), "Noise in America: The Extent of the Noise Problem," Report 550/9-81-101, EPA, Washington, DC.

U.S. Environmental Protection Agency (EPA) (2009), "Product Noise Labeling Hearing Protection Devices; Proposed Rule," 40 CFR Part 211, *Federal Register*, EPA Docket No. EPA-HQ-OAR-2003-0024; FRL-8934-9, EPA, Washington, DC, August 5.

Wegel, R. L., and Lane, C. E. (1924), "The Auditory Masking of One Pure Tone by Another and Its Probable Relation to the Dynamics of the Inner Ear," *Physiological Review*, Vol. 23, p. 266.

Wilkins, P. A., and Martin, A. M. (1982), "The Effects of Hearing Protection on the Perception of Warning Sounds," in *Personal Hearing Protection in Industry*, P. W. Alberti, Ed., Raven, New York, pp. 339–369.

Wilkins, P. A., and Martin, A. M. (1985), "The Role of Acoustical Characteristics in the Perception of Warning Sounds and the Effects of Wearing Hearing Protection," *Journal of Sound and Vibration*, Vol. 100, No. 2, pp. 181–190.

Zwicker, E. (1960), "En Verfahren zur Berechnung der Lautstärke," *Acustica*, Vol. 10, pp. 304–308.

# CHAPTER 24

# ILLUMINATION

**Peter R. Boyce**
**Consultant**
**Canterbury, United Kingdom**

## 1 INTRODUCTION

Illumination is the act of placing light on an object. By providing illumination, stimuli for the human visual system are produced and the sense of sight is allowed to function. With light we can see, without light we cannot see. This chapter is devoted to describing how to measure and produce illumination, the effects of different lighting conditions on visual performance and visual comfort, the photobiological and psychological effects of illumination, and the risks inherent in exposure to light.

## 2 MEASUREMENT OF ILLUMINATION

### 2.1 Photometric Quantities

Light is a part of the electromagnetic spectrum, lying between the wavelength limits 380–780 nm. What separates this wavelength region from the rest is that radiation in this region is absorbed by the photoreceptors of the human visual system, which initiates the process of seeing.

The most fundamental measure of the electromagnetic radiation emitted by a source is its radiant flux. This is a measure of the rate of flow of energy emitted and is measured in watts. The most fundamental quantity used to measure light is luminous flux. Luminous flux is radiant flux multiplied by the relative spectral sensitivity of the human visual system over the wavelength range 380–780 nm.

The relative spectral sensitivity of the human visual system is based on the perception of brightness associated with each wavelength. In fact, there are two different relative spectral sensitivities, sanctified by international agreement arranged through the Commission Internationale de l'Eclairage (CIE, 1983, 1990). There are two relative spectral sensitivities because the human visual system has two classes of photoreceptor: cones, which operate primarily when light is plentiful, and rods, which operate when light is very limited. These two photoreceptor types have different spectral sensitivities: the day photoreceptor, the cones, characterized by the CIE standard photopic observer, and the night photoreceptor, the rods, characterized by the CIE standard scotopic observer (Figure 1).

Luminous flux is used to quantify the total light output of a light source in all directions. While this is important, for lighting practice it is also important to be able to quantify the luminous flux emitted in a given direction. The measure that quantifies this concept is luminous intensity. Luminous intensity is the luminous flux emitted per unit solid angle in a specified direction.
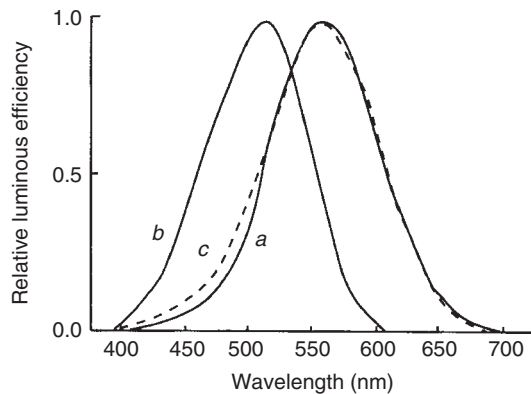


**Figure 1** Relative luminous efficiency functions for (a) the CIE standard photopic observer and (b) the CIE standard scotopic observer. The CIE standard photopic observer is based on a 2-degree field of view. Also shown (c) is the relative luminous efficiency function for a 10-degree field of view in photopic conditions.

The unit of measurement is the candela, which is equivalent to one lumen per steradian. Luminous intensity is used to quantify the distribution of light from a luminaire.

Both luminous flux and luminous intensity have area measures associated with them. The luminous flux falling on unit area of a surface is called the illuminance. The unit of measurement of illuminance is the lumens per square meter, or lux. The luminous intensity emitted per unit projected area in a given direction is the luminance. The unit of measurement of luminance is the candela per square meter. The illuminance incident on a surface is the most widely used electric lighting design criterion. The luminance of a surface is a correlate of its brightness. Table 1 summarizes these photometric quantities and the relationship between illuminance and luminance.

Unfortunately for consistency, photometry has a long history that has generated a number of different units of measurement for illuminance and luminance. Table 2 lists some of the alternative units, together with the multiplying factors necessary to convert from the alternative unit to the System Internationale (SI) units of lumens per square meter for illuminance and candelas per square meter for luminance. The SI units will be used throughout this chapter.

Table 3 shows some illuminances and luminances typical of commonly occurring situations.

### 2.2 Colorimetric Quantities

The photometric quantities described above do not take into account the wavelength combination, that is, the color of the light being measured. There are two approaches to characterizing color, the color order system and the CIE colorimetry system.

#### 2.2.1 Color Order Systems

A color order system is a physical, three-dimensional representation of color space. It is three dimensional because colors have three separate subjective attributes; hue, brightness, and strength. Hue tells us whether the color is primarily red or yellow or green or blue. Brightness tells us to what extent the color transmits or reflects light. Strength tells us whether the color is strong or weak.

There are several different color order systems used in different parts of the world (Wyszecki and Stiles, 1982). Probably the most widely used is the Munsell Book of Color available from the Munsell Color Company. Figure 2 shows the three-dimensional color space of the Munsell system. The position of any color is identified by an alphanumeric code made up of three terms: hue, value, and chroma (e.g., a strong red is given the alphanumeric 7.5R/4/12). Hue, value, and chroma are related to the three attributes of color: hue, brightness, and strength, respectively. Building materials, such as paints, plastic, and ceramics, are commonly classified in terms of a color order system.

#### 2.2.2 CIE Colorimetric System

Sometimes, it is necessary to quantify the color of a light or a surface before either exists. To meet this

**Table 1 Photometric Quantities**

| Quantity | Definition | Units |
|---|---|---|
| Luminous flux | That quantity of radiant flux which expresses its capacity to produce visual sensation | lumen (lm) |
| Luminous intensity | The luminous flux emitted in a very narrow cone containing the given direction divided by the solid angle of the cone (i.e., luminous flux/unit solid angle) | candela (cd) |
| Illuminance | The luminous flux/unit area at a point on a surface | lumen meter$^{-2}$ (lm m$^{-2}$) |
| Luminance | The luminous flux emitted in a given direction divided by the product of the projected area of the source element perpendicular to the direction and the solid angle containing that direction, i.e. luminous flux/unit solid angle/unit area | candela meter$^{-2}$ (cd m$^{-2}$) |
| Reflectance | The ratio of the luminous flux reflected from a surface to the luminous flux incident on it: | |
| For a matte surface | $luminance = \dfrac{illuminance \times reflectance}{\pi}$ | |
| Luminance factor | The ratio of the luminance of a reflecting surface, viewed in a given direction to that of a perfect white uniform diffusing surface identically illuminated: | |
| For a nonmatte surface for a specific viewing direction and lighting geometry | $luminance = \dfrac{illuminance \times luminance\ factor}{\pi}$ | |

**Table 2 Common Photometric Units of Measurement for Illuminance and Luminance and Factors Necessary to Change Them to SI Units**

| Quantity | Unit | Dimensions | Multiplying Factor to Convert to SI Unit |
|---|---|---|---|
| Illuminance (SI unit = lumen meter$^{-2}$) | lux | lumen meter$^{-2}$ | 1.00 |
| | meter candle | lumen meter$^{-2}$ | 1.00 |
| | phot | lumen centimeter$^{-2}$ | 10,000.00 |
| | foot candle | lumen foot$^{-2}$ | 10.76 |
| Luminance (SI unit = candela meter$^{-2}$) | nit | candela meter$^{-2}$ | 1.00 |
| | stilb | candela centimeter$^{-2}$ | 10,000.00 |
| | — | candela inch$^{-2}$ | 1,550.00 |
| | — | candela foot$^{-2}$ | 10.76 |
| | apostilb[a] | lumen meter$^{-2}$ | 0.32 |
| | blondel[a] | lumen meter$^{-2}$ | 0.32 |
| | lambert[a] | lumen centimeter$^{-2}$ | 3,183.00 |
| | foot-lambert[a] | lumen foot$^{-2}$ | 3.43 |

[a]These four items are based on an alternative definition of luminance. This definition is that if the surface can be considered as perfectly matte, its luminance in any direction is the product of the illuminance on the surface and its reflectance. Thus, the luminance is described in lumens per unit area. This definition is deprecated in the SI system.

**Table 3 Typical Illuminance and Luminance Values**

| Situation | Illuminance on Horizontal Surface (lm/m$^2$) | Typical surface | Luminance (cd/m$^2$) |
|---|---|---|---|
| Clear sky in summer in northern temperate zones | 150,000 | Grass | 2,900 |
| Overcast sky in summer in northern temperate zones | 16,000 | Grass | 300 |
| Textile inspection | 1,500 | Light grey cloth | 140 |
| Office work | 500 | White paper | 120 |
| Heavy engineering | 300 | Steel | 20 |
| Residential road lighting | 10 | Asphalt road surface | 0.2 |
| Moonlight | 0.1 | Asphalt road surface | 0.002 |
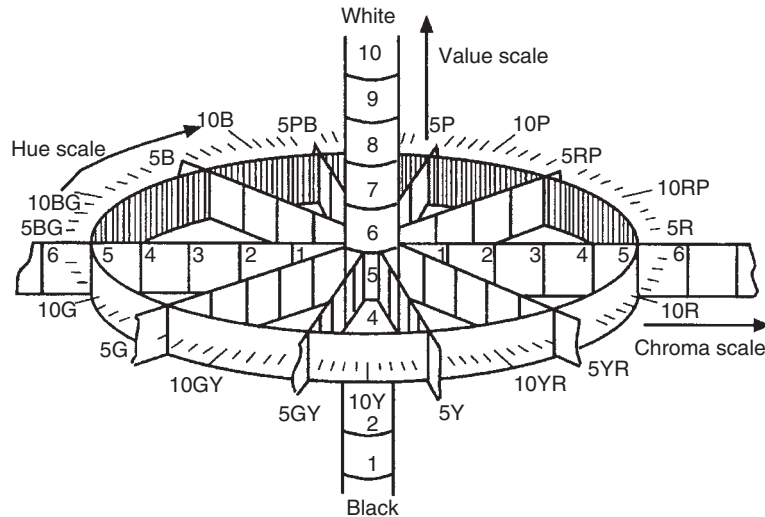
**Figure 2** Organization of Munsell color order system. The hue letters are B = blue, PB = purple/blue, P = purple, RP = red/purple, R = red, YR = yellow/red, Y = yellow, GY = green/yellow, G = green, BG = blue/green.

need and to provide a more accurate characterization of color, the CIE has developed a system of colorimetry ranging from the simple to the complex (CIE, 1978, 1995, 2004a,b). The most fundamental characteristic of light is the spectral power distribution reaching the eye. It is this spectral power distribution that largely determines the color seen, although the perception of color is also influenced by the surroundings (Purves and Beau Lotto, 2003). Unfortunately, the implications of comparisons between spectral power distributions are difficult to comprehend. The CIE has developed two three-dimensional color spaces, both based on mathematical manipulations applied to spectral distributions (Robertson, 1977; CIE, 1978). These color spaces, Lab and Luv, provide a convenient means of quantifying color, the Lab space used mainly for object colors and the Luv space for self-luminous colors. If two colors have the same coordinates in one of these color spaces, under the same observing conditions they will appear the same. The distance between two colors in the color space is related to how easily they can be distinguished.

An earlier CIE color space, the 1964 Uniform Color Space, is used in the calculation of the CIE general color-rendering index, a single number index which is widely applied to light sources to indicate how accurately they render colors relative to some standard (CIE, 1995). Specifically, the positions in color space of 8 test colors, under a reference light source and under the light source of interest, are calculated. The separation between the two positions of each test color are calculated, the separations for all the test colors are summed and scaled to give a value of 100 when there is no separation for any of the test colors, i.e. for perfect color rendering.

It should be noted that the CIE general color-rendering index is a very crude metric. Different light sources have different reference light sources and the

summation means that light sources that render the test colors differently can have the same color-rendering index. Much more sophisticated are the color appearance models now available (Hunt, 1991; CIE, 2004b), but their existence has had little impact on lighting practice. Rather, a two-dimensional color surface is still widely used to characterize the color appearance of light sources and to define the acceptable color characteristics of light signals (CIE, 1994). This is the CIE 1931 chromaticity diagram shown in Figure 3. Essentially it is a slice through the color space at a fixed luminance. The curved boundary of the chromaticity diagram consists of the colors produced by single wavelengths. The equal-energy point in the center of the diagram corresponds to a colorless surface. The further the coordinates of a color are from the equal-energy point and the closer they are to the boundary, the greater the strength of the color. Figure 3 also shows several areas in which a signal light needs to fall if it is to be perceived as the specified color. The color appearance of nominally white light sources is conventionally described by their correlated color temperature. This is the temperature of the full radiator that is closest to the coordinates of the light source on the CIE 1931 chromaticity diagram (Wyszecki and Stiles 1982). A useful summary of these colorimetry systems is given in the tenth edition of the *Lighting Handbook* of the Illuminating Engineering Society of North America (IESNA, 2011).

## 2.3 Instrumentation

The instrumentation for measuring photometric and colorimetric quantities can be divided into laboratory and field equipment. Laboratory equipment tends to be large and/or sophisticated and hence expensive. Field equipment is small and portable. The luminous flux from a light source, the luminous intensity distribution
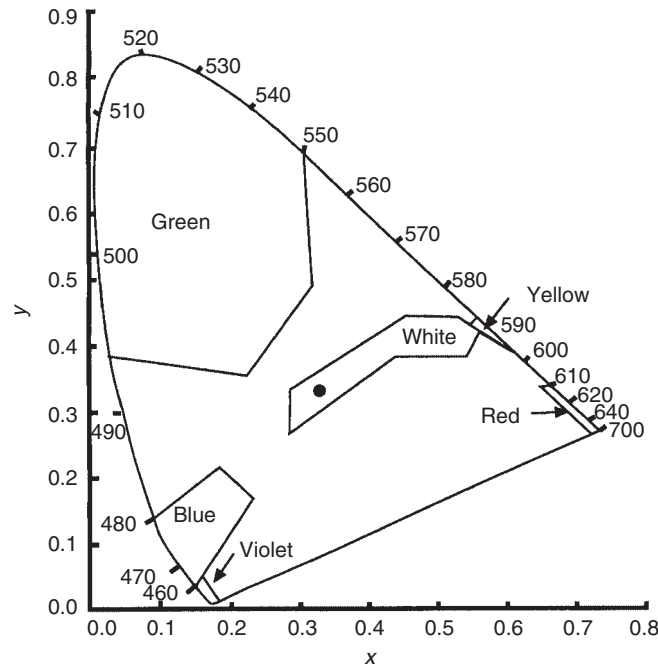
**Figure 3** CIE 1931 chromaticity diagram. The boundary curve is the spectrum locus with the wavelengths (nm) marked. The filled circle is the equal-energy point. The enclosed areas indicate the chromaticity coordinates of light signals that will be identified as the specified colors.

of a luminaire, and light source color properties are conventionally measured in the laboratory.

The two most widely used field instruments are the illuminance meter and the luminance meter. Illuminance meters have three important characteristics: sensitivity, color correction, and cosine correction. Sensitivity refers to the range of illuminances covered, the range desired being dependent on whether the instrument is to be used to measure daylight, interior lighting, or night-time exterior lighting. Color correction means that the illuminance meter has a spectral sensitivity matching the CIE standard photopic observer. Cosine correction means that the illuminance meter's response to light striking it from directions other than the normal follows a cosine law.

The luminance meter is designed to measure the average luminance over a specified area. The luminance meter has an optical system that focuses an image on a detector. Looking through the optical system allows the operator to identify the area being measured and usually displays the luminance of the area. The important characteristics of a luminance meter are its spectral response, its sensitivity, and the quality of its optical system. Again, a good luminance meter has a spectral response matching the CIE standard photopic observer. The sensitivity needed depends on the conditions under which it will be used. The quality of its optical system can be measured by its sensitivity to light from outside the measurement area (CIE, 1987).

Procedures for using illuminance or luminance meters in the field and for light measurements in the laboratory are described and referenced in the guidance published by national bodies [IESNA, 2011; Chartered Institution of Building Services Engineers (CIBSE), 2009; Society of Light and Lighting (SLL), 2009]. It should be noted that virtually all commercial instrumentation used to measure illuminance and luminance uses the CIE standard photopic observer as the basis of the instrument's spectral sensitivity, even when the instrument is designed to be used in mesopic and scotopic conditions.

Recently, another approach has been developed for rapidly acquiring the distribution of luminances over a large area. This approach uses multiple images captured by a digital camera and is called high-dynamic-range (HDR) imaging (Inanici, 2006). At the moment, HDR imaging is mainly being used for capturing luminance distributions that are subject to large and sudden changes, for example, sky luminances.

## 3 PRODUCTION OF ILLUMINATION

Illumination is produced naturally by the sun and artificially by electric light sources.

### 3.1 Daylight, Sunlight, and Skylight

Natural light is light received on Earth from the sun, either directly or after reflection from the moon. The prime characteristic of natural light is its variability. Natural light varies in magnitude, spectral content, and distribution with different meteorological conditions, at

different times of day and year, and at different latitudes. Moonlight is of little interest as a source of illumination, but daylight is used, and strongly desired, for the lighting of buildings. Daylight can be divided into two components, sunlight and skylight. Sunlight is light received at Earth's surface, directly from the sun. Sunlight produces strong, sharp-edged shadows. Skylight is light from the sun received at Earth's surface after scattering in the atmosphere. Skylight produces only weak, diffuse shadows. The balance between sunlight and skylight is determined by the nature of the atmosphere and the distance that the light passes through it. The greater is the amount of water vapor and the longer the distance, the higher is the proportion of skylight.

The illuminances on Earth's surface produced by daylight can cover a large range, from 150,000 lx on a sunny summer day to 1000 lx on a heavily overcast day in winter. Several models exist for predicting the daylight incident on a plane at different locations for different atmospheric conditions (Robbins, 1986). These models can be used to predict the contribution of daylight to the lighting of interiors. Alternatively, there are now available sets of measured illuminance or irradiance data for many different sites around the world. These make it possible to do climate-based modeling of daylight availability in interiors and the impact of daylight on the energy use of buildings (Mardaljevic et al., 2009).

The spectral composition of daylight also varies with the nature of the atmosphere and the path length through it. The correlated color temperature of daylight can vary from 4000 K for an overcast day to 40,000 K for a clear blue sky. For calculating the appearance of objects under natural light, the CIE recommends the use of one of three different spectral distributions corresponding to correlated color temperatures of 5503, 6504, and 7504 K (Wyszecki and Stiles, 1982).

## 3.2 Electric Light Sources

The lighting industry makes several thousand different types of electric lamps. Those used for providing illumination can be divided into three classes: incandescent, discharge, and solid state. Incandescent lamps produce light by heating a filament. Discharge lamps produce light by an electric discharge in a gas. Solid-state lamps produce light by the passage of an electric current through a semiconductor. Incandescent lamps can operate directly from mains electricity. Discharge lamps all require control gear between the lamp and the electricity supply, because different electrical conditions are required to initiate the discharge and to sustain it. Solid-state lamps require devices, called drivers, to limit the current through the semiconductor.

Electric light sources can be characterized on several different dimensions:

- *Luminous Efficacy*. The ratio of luminous flux produced to power supplied (lumens per watt). If the lamp needs control gear, the watts supplied should include the power demand of the control gear.
- *Correlated Color Temperature (CCT)*. A measure of the color appearance of the light produced, measured in degrees Kelvin (see Section 2.2.2).

- *CIE General Color-Rendering Index (CRI)*. A measure of the ability to render colors accurately (see Section 2.2.2).
- *Lamp Life*. The number of burning hours until either lamp failure or a stated percentage reduction in light output occurs. Lamp life can vary widely with switching cycle.
- *Run-Up Time*. The time from switch-on to full light output.
- *Restrike Time*. The time delay between the lamp being switched off before it will reignite.

Table 4 summarizes these characteristics for two incandescent lamp types, seven discharge lamp types, and one solid-state type that are widely used for illumination and gives the most common applications for each lamp type. The values in Table 4 should be treated as indicative only. Details about the characteristics of any specific lamp should always be obtained from the manufacturer.

## 3.3 Control of Light Distribution

Being able to produce light is only part of what is necessary to produce illumination. The other part is to control the distribution of light from the light source. For daylight, this is done by means of window shape, placement, and glass transmittance (Robbins, 1986). For electric light sources, it is done by placing the light source in a luminaire. The luminaire provides electrical and mechanical support for the light source and controls the light distribution. The light distribution is controlled by using reflection, refraction, or diffusion, individually or in combination (Simons and Bean, 2000). One factor in the choice of which method of light control to adopt in a luminaire is the balance desired between the reduction in the luminance of the light source and the precision required in light distribution. Highly specular reflectors can provide precise control of light distribution but do little to reduce source luminance. Conversely, diffusers make precise control of light distribution impossible but do reduce the luminance of the luminaire. Refractors are an intermediate case. The light distribution provided by a specific luminaire is quantified by the luminous intensity distribution. All reputable luminaire manufacturers provide luminous intensity distributions for their luminaires.

## 3.4 Control of Light Output

The control of daylight admitted through a window is achieved by mechanical structures, such as light shelves, or by adjustable blinds (Littlefair, 1990). Whenever the sun or a very bright sky is likely to be directly visible through a widow, some form of blind will be required. Blinds can take various forms, horizontal, Venetian, vertical, and roller being the most common. Blinds can also be manually operated or motorized, either under manual control or under photocell control. Probably the most important feature to consider when selecting a blind is the extent to which it preserves a view of the outside. Roller blinds that can be drawn down to a position where the sun and/or sky is hidden but the lower part of the widow is still open are an attractive option.

**Table 4  Properties of Some Electric Light Sources Widely Used for Illumination**

| Source | Luminous Efficacy (lm/W) | CCT (K) | CRI | Lamp Life (h) | Run-up Time (min) | Restrike Time (min) | Application |
|---|---|---|---|---|---|---|---|
| Incandescent | | | | | | | |
|   Tungsten | 8–14 | 2500–2700 | 100 | 1000 | Instant | Instant | Residential |
|   Tungsten–halogen | 15–25 | 2700–3200 | 100 | 1500–5000 | Instant | Instant | Residential, retail, display |
| Discharge | | | | | | | |
|   Fluorescent | 20–96 | 2700–17,000 | 50–98 | 8000–19,000 | 0.5 | Instant | Commercial |
|   Compact fluorescent | 20–70 | 2700–6500 | 80–90 | 5000–15,000 | 0.5–1.5 | Instant | Commercial, retail, residential |
|   Mercury vapor | 33–57 | 3200–3900 | 40–50 | 8000–10,000 | 4 | 3–10 | Older industrial and road |
|   Metal halide | 60–98 | 3000–6000 | 60–93 | 2000–10,000 | 1–8 | 5–20 | Industrial, commercial, retail and road |
|   Low-pressure sodium | 70–180 | N/A | N/A | 15,000–20,000 | 10–20 | 1 | Road |
|   High-pressure sodium | 53–142 | 1900–2150 | 19–65 | 10,000–20,000 | 3–7 | 0–1 | Industrial, road |
|   Induction | 47–80 | 2550–4000 | 80 | 60,000 | Instant | Instant | Road |
| Solid state | | | | | | | |
|   White light-emitting diode (LED) using phosphor | 21–33 | 3000–4000 | 70–83 | 40,000 | Instant | Instant | Residential, retail |

Roller blinds made of a mesh material can preserve a view through the whole window while reducing the luminance of the view out. Such blinds are an attractive option where the problem is an overbright sky but will be of limited value when a direct view of the sun is the problem. The same applies to low-transmission glass.

For electric light sources, control of light output is provided by switching or dimming systems. Switching systems can vary from the conventional manual switch to sophisticated daylight control systems that dim lamps near windows when there is sufficient daylight. Time switches are used to switch off all or parts of a lighting installation at the end of the working day. Occupancy sensors are used to switch off lighting when there is nobody in the space. Such switching systems can reduce electricity waste, but they will be irritating if they switch lighting off when it is required and they may shorten lamp life if switching occurs frequently. The factors to be considered when selecting a switching system are whether to rely on a manual or an automatic system and, if it is automatic, how to match the switching to the activities in the space. If your interest is primarily in reducing electricity consumption, a good principle is to use automatic switch-off and manual switch-on. This principle uses human inertia for the benefit of reducing energy consumption.

As for dimming systems, these all reduce light output and energy consumption, but a different system is required for each lamp type. The factors to consider when evaluating a dimming system are the range over which dimming can be achieved without flicker or the lamp extinguishing, the extent to which the color properties of the lamp change as the light output is reduced, and any effect dimming has on lamp life and energy consumption. There are large individual differences in preferred illuminances so whether or not giving people individual control of dimming will reduce energy consumption depends on the maximum illuminance provided (Boyce et al., 2006b).

## 4  FUNCTIONAL CHARACTERISTICS OF HUMAN VISUAL SYSTEM

### 4.1  Visual System Structure

Illumination is important to humans because it alters the stimuli to the visual system and the operating state of the visual system itself. Therefore, an understanding of the capabilities of the visual system and how they vary with illumination is important to an understanding of the effects of illumination. The visual system is composed of the eye and brain working together. Light entering the eye is brought to focus on the retina by the combined optical power of the air/cornea surface and the lens of the eye. The retina is really an extension of the brain, consisting of two different types of photoreceptors and numerous nerve interconnections. At the photoreceptors, the incident photons of light are absorbed and converted to electrical signals. The nerve interconnections take these signals and carry out some basic image processing. The processed image is transmitted up the optic nerve of each eye to the optic chiasma, where nerve fibers from the two eyes are combined and transmitted to the left
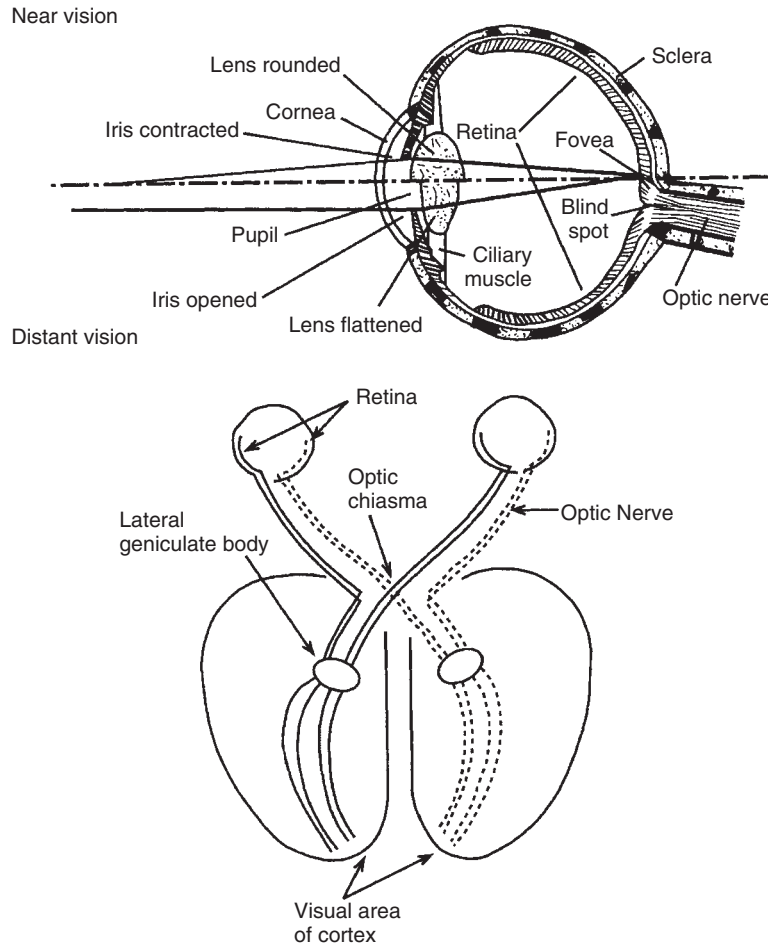
**Figure 4**   Section through the eye adjusted for near and distant vision and a schematic diagram of the binocular nerve pathways of the visual system.

and right parts of the visual cortex. It is in the visual cortex that the signals from the eye are interpreted in terms of past experience (Figure 4).

Many of the capabilities of the visual system can be understood from the organization of the retina. The two types of visual photoreceptors, called rods and cones from their anatomical appearance, have different wavelength sensitivities and different absolute sensitivities to light and are distributed differently across the retina.

Rods are the more sensitive of the two and effectively provide a night retina. Cones are less sensitive to light and operate during daytime. In fact, there are three types of cones, each with a different spectral sensitivity. These cones are commonly called long-, middle-, and short-wavelength cones, from their regions of maximum spectral sensitivity. These three cone types combine together to give the perception of color. Figure 5 shows the distribution of rods and cones across the retina. Cones are concentrated in a small central area of the

retina called the fovea that lies where the visual axis of the eye meets the retina, although there are cones distributed evenly across the rest of the retina. Rods are absent from the fovea, reaching their maximum concentration about $20°$ from the fovea. This variation in concentration of rods and cones with deviation from the fovea is amplified by the number of photoreceptors connected to each optic nerve fiber. In the fovea, the ratio of photoreceptors to optic nerve fibers is close to 1 but increases rapidly as the deviation from the fovea increases. The net effect of this structure is to provide different functions for the fovea and the periphery. The fovea is the part of the retina which provides fine discrimination of detail. The rest of the retina is primarily devoted to detecting changes in the visual environment that require the attention of the fovea.

## 4.2  Wavelength Sensitivity

The rod and cone photoreceptors have different absolute spectral sensitivities (Figure 6). The spectral response of
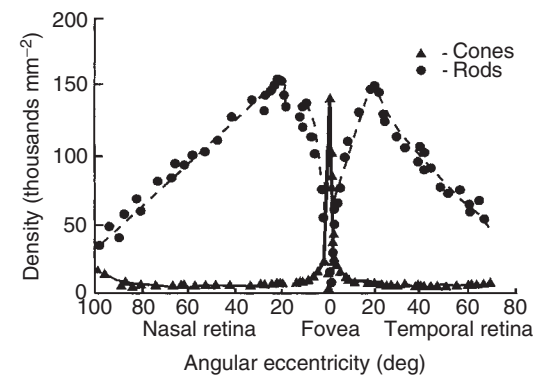
**Figure 5** Density of rod and cone photoreceptors across the retina on a horizontal meridian. (After Osterburg, G., *Acta Ophthalmologica*, Vol. 13, Suppl. 6, 1935.)



**Figure 6** Log relative spectral sensitivity of rod and cone photoreceptors plotted against wavelength. (After Wald, G., *Science,* Vol. 101, p. 653, 1945.)

the cones lies between 380 and 780 nm with the peak sensitivity occurring at 555 nm. The spectral response of the rods lies between 380 and 780 nm with the peak at 507 nm. The peak sensitivity of the rods is much greater than that of the cones. These spectral sensitivities form the basis of the CIE standard observers and hence the photometric quantities discussed in Section 2.1. By adjusting the spectral emission of a light source to lie within the most sensitive part of the spectral response of the visual system, lamp manufacturers are able to vary the number of lumens emitted for each watt of power applied.

## 4.3 Adaptation

The visual system can operate over a range of about 12 log units of luminance, from a luminance of $10^{-6}$ to $10^6$ cd/m$^2$, from starlight to bright sunlight. But it cannot cover this range simultaneously. At any instant in time, the visual system can cover a range of 2 or 3 log units of luminance. Luminances above this limited range are seen as glaringly bright, those below as undifferentiated black. The capabilities of the visual system depend on where in the complete range of luminances it is adapted. Three different functional ranges of luminance are conventionally identified: the photopic, mesopic, and scotopic. Table 5 summarizes the visual system capabilities in each of these functional ranges.

The visual system continuously adjusts its state of adaptation through three mechanisms: neural, mechanical, and photochemical. These three mechanisms differ in their speed and range of adjustment. The neural mechanism, which is based in the retina, operates in milliseconds and covers a range of two to three log units in luminance. The mechanical mechanism involves the expansion and contraction of the iris. The consequent changes in pupil size take about a second but cover less than one log unit in luminance. The photochemical mechanism covers the whole range of luminance but is slow, the changes taking minutes. The exact time will depend on the starting and finishing luminances. If both starting and finishing luminances for the adaptation are greater than 3 cd/m$^2$, only cones are involved. As the time constant for cones is of the order of 2–3 min, adaptation takes only a few minutes. When the starting luminance is in the operating range of the cones and the finishing luminance is within the operating range of the rods, a two-stage adaptation process occurs involving both cones and rods. As rods have a time constant around 7–8 min, the adaptation time is much longer. Complete adaptation from a high photopic luminance to darkness can take up to an hour.

**Table 5 Functional Ranges of Visual System Capabilities**

| Operating State Range | Luminance (cd/m$^2$) | Active Photoreceptors | Peak Wavelength Sensitivity | Characteristics |
|---|---|---|---|---|
| Photopic | > 3 | Cones | 555 nm | Good color vision Good resolution |
| Scotopic | < 0.001 | Rods | 507 nm | No color vision Poor resolution Fovea ''blind'' |
| Mesopic | > 0.001 and <3 | Cones and rods | 555 nm in fovea, between 555 nm and 507 nm elsewhere | Reduced color vision, reduced resolution relative to photopic |

Interior lighting is almost always sufficient for the visual system to be operating in the photopic region. Exterior lighting on roads and in urban areas is usually sufficient to keep the visual system operating in the low-photopic or high-mesopic regions. It is in very rural areas, at sea, or underground, where there is neither exterior lighting nor moonlight, that the visual system reaches the scotopic state. The speed of adaptation is important where a large and sudden change in the luminance occurs. Examples of situations where this happens are the entrance to road tunnels during daytime (Bourdy et al., 1987) and the onset of emergency lighting during a power failure (Boyce, 1985). These problems are overcome either by installing a gradual reduction in luminance, which allows more time for adaptation to occur, or by setting a minimum luminance within the neural adaptation range.

### 4.4   Color Vision

When photopically adapted, the visual system can discriminate many thousands of colors. This ability to discriminate colors reduces as the adaptation luminance decreases through the mesopic region and vanishes in the scotopic vision. This is because color vision is mediated by the cone photoreceptors. Different light sources have different spectral emissions and hence render colors differently. To ensure good color discrimination, it is necessary to use a light source that has a high CIE general color-rendering index and that produces sufficient light to ensure the visual system is operating in the photopic state.

### 4.5   Receptive Field Size and Eccentricity

The retina is organized in such a way that increasing numbers of photoreceptors are connected to each optic nerve fiber as the deviation from the fovea increases.

   This feature of the visual system is important when detection of a stimulus is necessary and it can occur anywhere in the visual field. The visual system will normally operate by first detecting the stimulus off-axis, that is, in the peripheral visual field, and then turning the eye so that the stimulus is brought onto the fovea for detailed examination. In order to identify a stimulus off-axis, the stimulus should be clearly different from its background, in luminance or color, and should change in space or time, that is, it should either move or flicker. A flickering light is commonly used to draw drivers' attention to important signs placed beside or above the road.

### 4.6   Meaningful Stimulus Parameters

Any stimulus to the visual system can be described by five parameters: its visual size, luminance contrast, chromatic contrast, retinal image quality, and retinal illumination. These parameters are important in determining the extent to which the visual system can detect and identify the stimulus.

### 4.6.1   Visual Size

The visual size of a stimulus describes how big the stimulus is. The larger a stimulus is, the easier it is to detect.

There are several different ways to express the size of a stimulus presented to the visual system, but all of them are angular measures. The visual size of a stimulus for detection is best given by the solid angle the stimulus subtends at the eye. The solid angle is given by the quotient of the areal extent of the object and the square of the distance from which it is viewed. The larger the solid angle is, the easier the stimulus is to detect.

   The visual size for resolution is usually given as the angle the critical dimension of the stimulus subtends at the eye. What the critical dimension is depends on the stimulus. For two points, the critical dimension is the distance between the two points. For two lines, it is the separation between the two lines. For a Landolt ring, it is the gap width. The larger is the visual size of detail in a stimulus, the easier it is to resolve the detail.

   For complex stimuli, the measure used to express their dimensions is the spatial frequency distribution. Spatial frequency is the reciprocal of the angular subtense of a critical detail, in cycles per degree. Complex stimuli have many spatial frequencies and hence a spatial frequency distribution. The match between the spatial frequency distribution of the stimulus and the contrast sensitivity function of the visual system (see Section 5.2) determines if the stimulus will be seen and what detail will be resolved.

   Lighting can change the visual size of three-dimensional stimuli by casting shadows that extend or diminish the apparent visual size of the stimulus.

### 4.6.2   Luminance Contrast

The luminance contrast of a stimulus quantifies its luminance relative to its background. The higher the luminance contrast is, the easier it is to detect the stimulus. There are two different forms of luminance contrast. For stimuli that are seen against a uniform background, luminance contrast is defined as

$$C = |L_t - L_b|/L_b$$

where

$\quad C$ = luminance contrast
$\quad L_b$ = Luminance of the background
$\quad L_t$ = Luminance of the detail

   This formula gives luminance contrasts that range from 0 to 1 for stimuli that have details darker than the background and from 0 to infinity for stimuli that have details brighter than the background. It is widely used for the former, for example, printed text.

   For stimuli which have a periodic pattern, for example, a grating, the luminance contrast or modulation is given by

$$C = (L_{max} - L_{min})/(L_{max} + L_{min})$$

where

$\quad C$ = Luminance contrast
$\quad L_{max}$ = maximum luminance
$\quad L_{min}$ = minimum luminance

   This formula gives luminance contrast that ranges from 0 to 1.

Lighting can change the luminance contrast of a stimulus by producing disability glare in the eye or veiling reflections from the stimulus or by changing the incident spectral radiation when colored stimuli are involved.

### 4.6.3  Chromatic Contrast

Luminance contrast uses the total amount of light emitted from a stimulus and ignores the wavelengths of the emitted light. It is the wavelengths emitted from the stimulus that largely determine its color. It is possible to have a stimulus with zero luminance contrast that can still be detected because it differs from its background in color, that is, it has chromatic contrast. There is no widely accepted measure of chromatic contrast, although various suggestions have been made (Tansley and Boynton, 1978). Fortunately, chromatic contrast only becomes important for detection when luminance contrast has reached a low level, typically less than 0.2 (O'Donell and Colombo, 2008).

Lighting can alter chromatic contrast by using light sources with different spectral emission characteristics.

### 4.6.4  Retinal Image Quality

As with all image-processing systems, the visual system works best when it is presented with a clear, sharp image. The sharpness of the stimulus can be quantified by the spatial frequency distribution of the stimulus; a sharp image will have high spatial frequency components present, a blurred image will not.

The sharpness of the retinal image is determined by the stimulus itself, the extent to which medium through which it is transmitted scatters light, and the ability of the visual system to focus the image on the retina. Lighting can do little to alter any of these factors, although it has been shown that light sources that are rich in the short wavelengths produce smaller pupil sizes and these tend to improve visual acuity slightly (Berman et al., 2006). The suggested explanation is that the smaller pupil sizes produce greater depth of field and hence better retinal image quality (Berman et al., 1993).

### 4.6.5  Retinal Illumination

The retinal illumination determines the state of adaptation of the visual system and therefore alters its capabilities. The retinal illumination is determined by the luminance in the visual field modified by the pupil size. Retinal illumination is measured in trolands, a quantity formed from the product of the luminance of the visual field and the pupil size (Wyszecki and Stiles, 1982). Illuminances and surface reflectances determine the luminances of the visual field. Luminances and light spectrum determine pupil size.

## 5   EFFECTS ON THRESHOLD VISUAL PERFORMANCE

Qualitatively, threshold visual performance is the performance of a visual task close to the limits of what is possible. Quantitatively, it is the performance of a task at a level such that it can be correctly carried out on 50% of the occasions it is undertaken. Threshold visual performance is affected by many different variables. For example, visual acuity is affected by the form of the target used, the luminance contrast of the target, the duration for which it is presented, where in the visual field it appears, and the luminance of the surround relative to the luminance of the immediate background. In this discussion of threshold visual performance, attention will be limited to the effects of variables that are controlled by the lighting system, that is, the adaptation luminance and the spectral content of the light. Information on the influence of other variables can be obtained from Boff and Lincoln (1988). In the data presented it will be assumed that the observer is fully adapted to the prevailing luminance, the image of the target is on the fovea, the target is presented for an unlimited time, and that observer is correctly refracted.

### 5.1   Visual Acuity

Visual acuity is the limit in the ability to resolve detail. Visual acuity has been frequently measured using gratings or Landolt C's. Visual acuity can be quantified as the angle subtended at the eye by the size of detail that can be correctly detected on 50% of the occasions it is presented.

No matter what target is used visual acuity improves, that is, the size of detail that can be resolved decreases as adaptation luminance increases. Figure 7 shows that as adaptation luminance increases from scotopic to photopic conditions, visual acuity increases, asymptotically approaching a maximum at high luminances. Table 3 gives some luminances typically found in interior and exterior lighting installations. Given a value for the adaptation luminance, Figure 7 can be used to determine if detail of a given size can be resolved or not. A useful rule of thumb is that the detail needs to be four times bigger than the visual acuity limit if it is to be resolved sufficiently quickly to avoid affecting visual performance (Bailey et al., 1993) .

As for light spectrum, for a lamp producing white light, then, at the same luminance, a spectrum that produces a smaller pupil size will enhance visual acuity slightly (Berman et al., 2006).
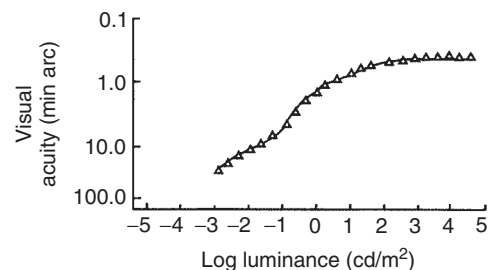


**Figure 7**   Effect of adaptation luminance on gap size of Landolt C target which can just be resolved. (After Shlaer, S., *Journal of General Physiology*, Vol. 21, p. 165, 1937.)
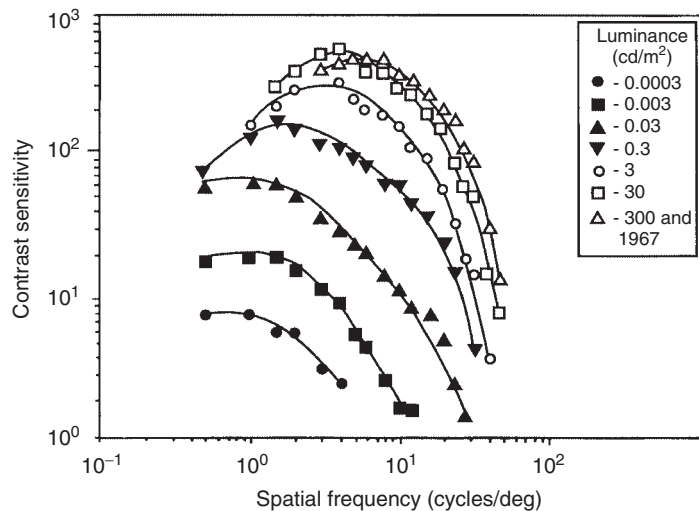
**Figure 8** Effect of adaptation luminance on contrast sensitivity function. Contrast sensitivity plotted against spatial frequency (cycles/degree) for different adaptation luminances. (After Savoy, R. L., and MaCann, J. J., *Journal of the Optical Society of America*, Vol. 65, p. 343, 1975.)

## 5.2 Contrast Sensitivity Function

Contrast sensitivity is the reciprocal of the luminance contrast that can be detected on 50% of the occasions it is presented. Contrast sensitivity is usually measured using a sinusoidal grating target. The contrast sensitivity function is contrast sensitivity plotted against the spatial frequency of the sinusoidal target.

Figure 8 shows the effect of adaptation luminance on the contrast sensitivity function. It shows that as the adaptation luminance increases from scotopic to photopic conditions, the contrast sensitivity increases for all spatial frequencies; the spatial frequency at which the peak contrast sensitivity occurs increases and the highest spatial frequency that can be detected also increases. Figure 8 can be used to determine if a given target will be visible by breaking the target into its spatial frequency components and determining if any of the components are within the limit set by the contrast sensitivity function (Sekular and Blake, 1994). The target will only be visible if at least one of its components falls within this limit, although it should be noted that the appearance of the target will be different depending on which component or components are visible. As a rule of thumb, for a target to be easily seen, it is necessary for the luminance contrast to be at least twice the contrast threshold.

As for the light spectrum, the results of Berman et al. (2006) imply that the contrast sensitivity function is somewhat influenced by different white-light spectra.

## 5.3 Temporal Sensitivity Function

The temporal sensitivity function shows percentage modulation amplitude plotted against the frequency of the modulation. Figure 9 shows the effect of adaptation luminance on the temporal sensitivity function. It shows
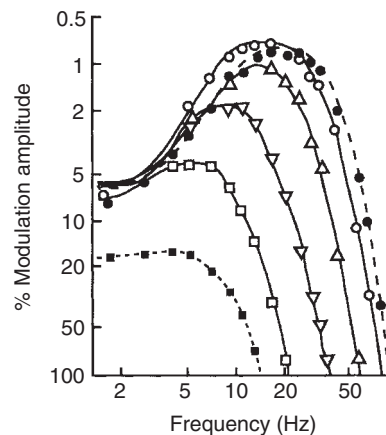


**Figure 9** Effect of adaptation luminance on temporal sensitivity function. Percentage modulation amplitude plotted against frequency (Hz) for different levels of retinal illumination. The retinal illuminations are: filled square = 0.06 trolands; open square = 0.65 trolands; open, inverted triangle = 7.1 trolands; open, upright triangle = 77 trolands; open circle = 850 trolands; filled circle = 9300 trolands. (After Kelly, D. H., *Journal of the Optical Society of America*, Vol. 51, p. 422, 1961.)

that as the adaptation luminance increases from mesopic to photopic conditions, the temporal sensitivity increases for all frequencies; the frequency at which the peak temporal sensitivity occurs increases, and the highest frequency that can be detected also increases. Figure 9 can be used to determine if a given temporal variation will be visible by breaking the waveform representing

the light fluctuation into its frequency components and determining if any of the components are within the limit set by the temporal sensitivity function. The fluctuation will only be visible if at least one of its frequency components falls within this limit.

Temporal fluctuation in luminous flux (i.e., flicker) is undesirable in lighting installations. To eliminate flicker, it is necessary to increase the frequency and/or decrease the modulation sufficiently to take their combination outside the limits set by the temporal sensitivity function. In practice, this is easily done. Incandescent lamps have sufficient thermal inertia to ensure that, even though the frequency of the fluctuation is only twice the supply frequency (120 Hz for a 60-Hz electrical supply), the modulation is small so there is little chance of seeing flicker from such a lamp. Discharge and solid-state lamps do not have thermal inertia so their modulation can be high where there is no phosphor used to modify the spectrum of the light emitted. Where a phosphor

is used, the persistence of the phosphor will tend to reduce the modulation. To ensure that discharge and solid-state lamps do not produce visible flicker, it is best to use a control gear that operates at a high frequency or, alternatively, in the case of solid-state lamps, zero frequency, that is, from a direct-current supply.

## 5.4 Color Discrimination

The ability to discriminate between two colors of the same luminance depends on the difference in spectral power distribution of the light received at the eye. Figure 10 shows the MacAdam ellipses, the area around a number of chromaticities, each magnified 10 times, within which no discrimination of color can be made, even under side-by-side comparison conditions (Wyszecki and Stiles, 1982).

The effect of illuminance on the ability to discriminate between colors is limited in the photopic region,
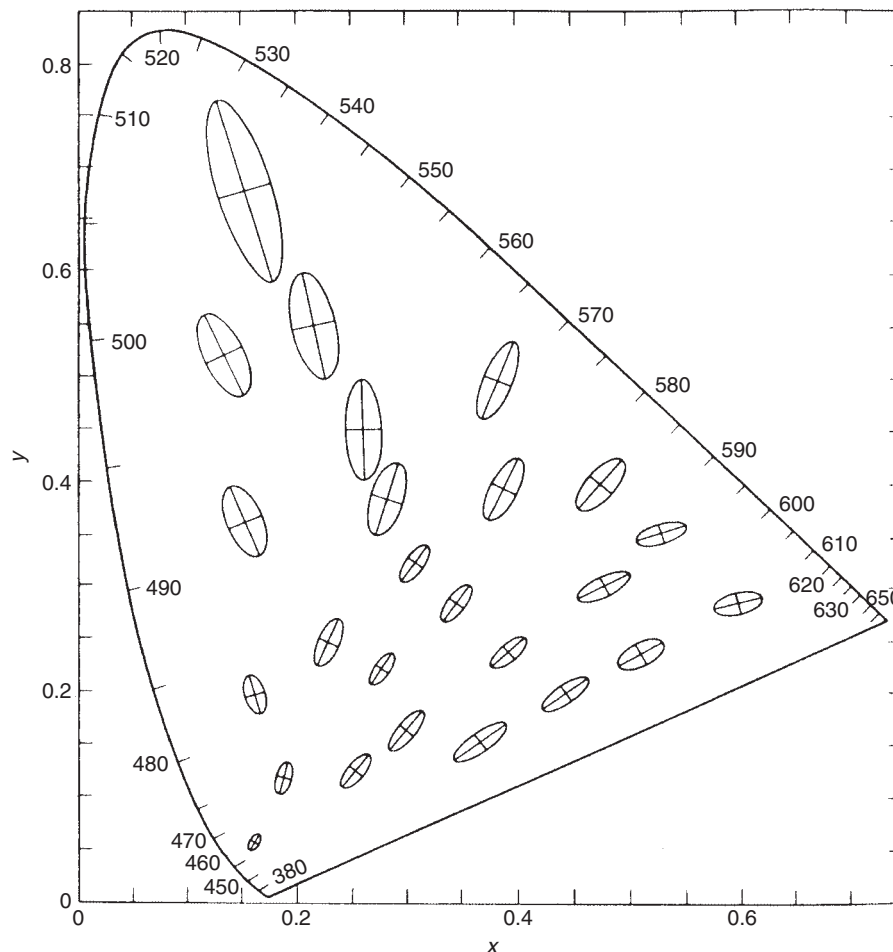


**Figure 10** MacAdam ellipses plotted on the CIE 1931 chromaticity diagram. The boundary of each ellipse represents 10 times the standard deviation of color matches made for the indicated chromaticity. (After MacAdam, D. L., *Journal of the Optical Society of America*, Vol. 32, p. 247, 1942.)

an illuminance of 300 lx being sufficient for good color judgment work (Cornu and Harlay, 1969). As the visual system enters the mesopic region, the ability to discriminate colors deteriorates and ultimately fails as the scotopic region is reached.

The effect of the light spectrum is much more important. The position of a color on the CIE 1931 chromaticity diagram is determined by the spectrum of the light and, if it is reflected from or transmitted through a surface, the spectral reflectance or transmittance of that surface. Therefore, by changing the light spectrum emitted by the lamp, it is possible to make colors easily discriminable or difficult to discriminate. The careful choice of light source is important wherever good color discrimination is important.

## 5.5 Interactions

The fact that there are many other variables besides adaptation luminance and light spectrum that influence threshold visual performance has been mentioned earlier. It is now necessary to introduce another complication, namely, interaction between the various components of visual system performance. As an example, consider the effect of luminance contrast on visual acuity. Visual acuity is conventionally measured using targets with a high luminance contrast. However, as the luminance contrast of the target is decreased, visual acuity also worsens. Similarly, the temporal sensitivity function as presented applies to a uniform luminance field. If the field has a pattern and hence a distribution of spatial frequencies, the temporal sensitivity function may be changed (Koenderink and Van Doorn, 1979).

Put crudely, what this means is that as visual performance gets closer to threshold, almost everything about the stimulus presented to the visual system becomes important. Further details on some of the interactions that occur are given in Boff and Lincoln (1988)

## 5.6 Approaches to Improving Threshold Visual Performance

Working close to threshold is not easy. In fact, it can be argued that the main function of anyone designing illumination is to provide conditions that avoid the need to use the visual system close to threshold. However, if this is required, then the following steps can be taken to improve threshold visual performance. Not all of the following steps will be possible in every situation and not all are appropriate for every problem. The discussion above should indicate which approach is likely to be most effective.

### Changing the Task:

- Increase the size of the detail in the task.
- Increase the luminance contrast of the detail in the task.
- Present the task so that it can be looked at directly, that is, with the fovea.
- Change the color of the target to make it more conspicuous
- Reduce the velocity of the task.
- Present the task for a longer time.

### Changing the Environment:

- Increase the adaptation luminance.
- Select a lamp with better color properties.
- Design the lighting so that it is free from disability glare and veiling reflections (see Section 7).

## 6 EFFECTS ON SUPRATHRESHOLD VISUAL PERFORMANCE

Suprathreshold visual performance is the performance of tasks that are easily visible because the stimuli they present to the visual system are well above those associated with threshold conditions. This raises the question as to why lighting conditions make a difference to task performance once what has to be seen is clearly visible. The answer is that although the stimuli are clearly visible, lighting influences the speed with which the visual information extracted from the stimuli can be processed. The aspect of lighting which determines this effect is the retinal illumination. The retinal illumination is determined by the luminance of the visual field that is viewed and hence by the illuminance on the surfaces which form that field.

### 6.1 Relative Visual Performance Model for On-Axis Detection

The relative visual performance (RVP) model of visual performance is an empirical model of the reaction time for the detection of different visual stimuli seen on the fovea for a range of adaptation luminances, luminance contrasts, and visual sizes (Rea and Ouellette, 1988, 1991). Figure 11 shows the form of the RVP model for four different visual size tasks, each surface being for a range of contrasts and retinal illuminances. The overall shape of the relative visual performance surface has been described as a plateau and an escarpment (Boyce and Rea, 1987). In essence what it shows is that the visual system is capable of a high level of visual performance over a wide range of visual sizes, luminance contrasts, and retinal illuminations (the plateau), but at some point either visual size or luminance contrast or retinal illumination will become insufficient and visual performance will rapidly collapse (the escarpment) toward threshold. The existence of a plateau of visual performance, or rather a near plateau because there is really a slight improvement in visual performance across the plateau, implies that for a wide range of visual conditions visual performance changes very little with changes in the lighting conditions.

The RVP model of suprathreshold visual performance provides a quantitative means of predicting the effects of changing either task size or contrast or the adaptation luminance on visual performance. It has been developed using rigorous methodology and has been validated against independently collected data (Eklund et al., 2001; Boyce, 2003). However, it is important to note that it should only be applied to a limited range of tasks. Specifically, it is most appropriate for tasks where task performance is dominated by the visual component (see Section 6.3), which do not require the use of off-axis vision to any extent; present stimuli to the visual system that can be completely characterized by
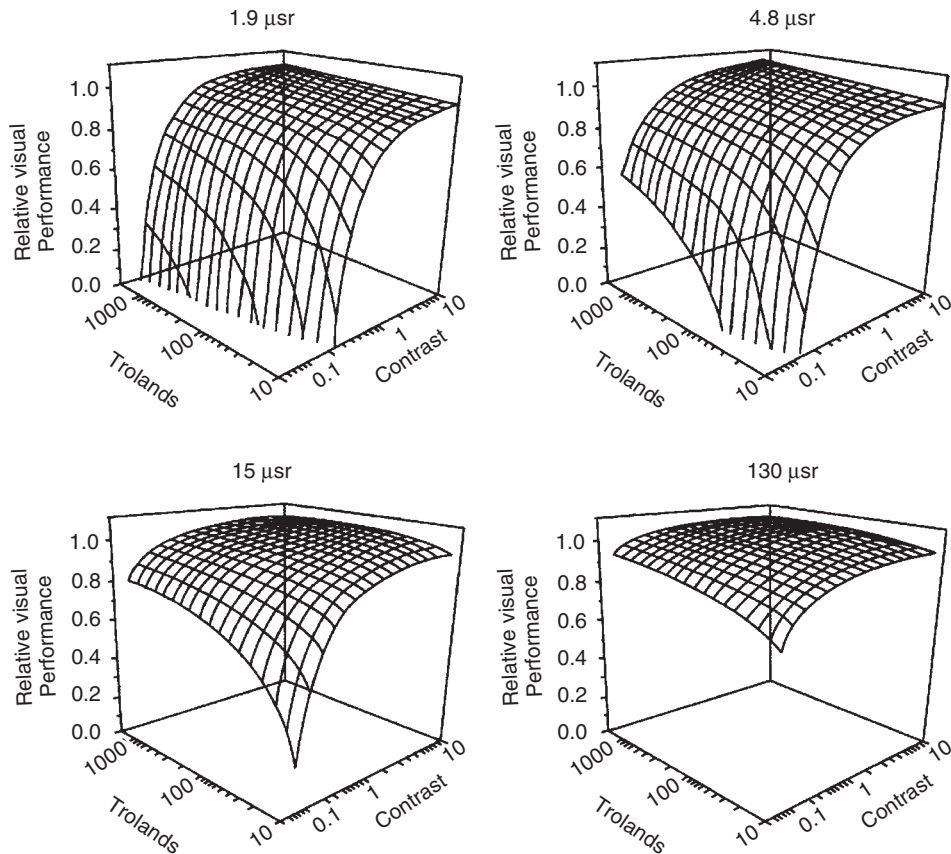
**Figure 11** Relative visual performance surfaces plotted against retinal illumination, in trolands, and luminance contrast for four stimuli subtending four different solid angles, measured in microsteradians. (After Rea, M. S., and Ouellette, M. J., *Lighting Research and Technology*, Vol. 23, p. 135, 1991.)

visual size, luminance contrast, and background luminance; and have values for these variables that fall within the ranges used to develop the model. Where the task involves chromatic contrast as well as luminance contrast, the RVP model is likely to be misleading and the light spectrum used for illumination will be important. Where the task is achromatic, the light spectrum is not likely to be important for suprathreshold visual performance unless performance is limited by the size of detail that needs to be seen (Boyce et al., 2003).

## 6.2 Visual Search

One class of tasks for which the RVP model is not applicable is comprised of those in which the object to be detected can appear anywhere in the visual field. These tasks involve visual search. Visual search is typically undertaken through a series of eye fixations, the fixation pattern being guided either by expectations about where the object to be seen is most likely to appear or by what part of the visual scene is most important. Typically, the object to be detected is first detected off-axis and then confirmed or resolved by an on-axis

fixation. The speed with which a visual search task is completed depends on the visibility of the object to be found, the presence of other objects in the search area, and the extent to which the object to be found is different from the other objects. The simplest visual search task is one in which the object to be found appears somewhere in an otherwise empty field, for example, paint defects on a car body. The most difficult visual search task is one where the object to be found is situated in a cluttered field and the clutter is very similar to the object to be found, for example, searching for a face in a crowd.

The lighting conditions necessary to achieve fast visual search are similar to those used to improve foveal threshold visual performance. By improving foveal threshold visual performance, the peripheral threshold visual performance is also improved so the object to be found is made more visible. The lighting required for fast visual search will have to be matched to the physical characteristics of the object to be found. For example, if the object is two dimensional and of matte reflectance located on a matte background, increasing the adaptation luminance is about the only option. However, if the object is three dimensional and has a specular reflectance

component, then light distribution can be used to increase the apparent size by casting shadows and the luminance contrast of the object by producing highlights on or around the object—changes which will be much more effective than simply increasing the adaptation luminance. Likewise, if the object is distinguished from its background primarily by color, the light spectrum used is an important consideration. It is this need to match the lighting conditions to the nature of the objects to be found which makes the design of lighting installations for visual inspection tasks so difficult and diverse (IESNA, 2011; Boyce, 2003).

The extent to which a lighting installation is effective in revealing an object can be estimated from the object's visibility lobe (Inditsky et al., 1982). The visibility lobe is the distribution of the probability of detecting the object within one eye fixation pause. This probability is a maximum when the object is viewed on-axis and decreases with increasing deviation from the fovea. The probability distribution is assumed to be radially symmetrical about the visual axis, resulting in circles around the fixation point, each circle having a given probability of detection within one fixation pause. For objects which appear on a uniform field, the visibility lobe is based on the detection of the object. For objects which appear among other similar objects, the visibility lobe is based on the discriminability of the object from the others surrounding it. Visual search will be fastest for objects which have the largest visibility lobe.

## 6.3 Visual Performance, Task Performance, and Productivity

Figure 12 shows the relationships between the stimuli to the visual system and their impact on visual performance, task performance, and productivity. The stimuli to the visual system, including the retinal illumination, determine the operation of the visual system and hence the level of visual performance achieved. This visual performance then contributes to task performance. It is important to point out that visual performance and task performance are not necessarily the same. Task performance is the performance of the complete task. Visual performance is the performance of the visual component of the task. Task performance is what is needed in order to measure productivity and to establish cost–benefit ratios. Visual performance is the only thing that changing the lighting conditions can affect directly.

Most apparently visual tasks have three components: visual, cognitive, and motor. The visual component refers to the process of extracting information relevant
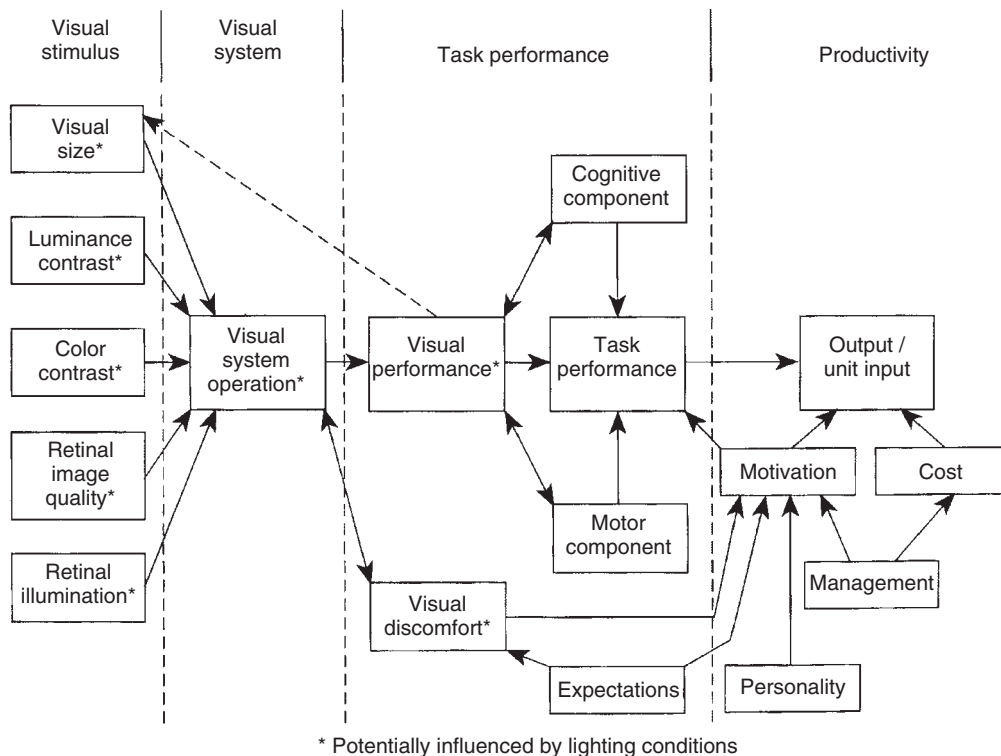


**Figure 12** Schematic of relationships between stimuli to the visual system and their impact on visual performance, task performance, and productivity. The arrows indicate the direction of the effects. The dotted arrow between visual performance and visual size indicates that, if visual performance is poor, a common response is to move closer to the stimulus to increase its visual size.

to the performance of the task using the sense of sight. The cognitive component is the process by which sensory stimuli are interpreted and the appropriate action determined. The motor component is the process by which the stimuli are manipulated to extract information and/or the actions decided upon are carried out. Every task is unique in its balance between visual, cognitive, and motor components and hence in the effect lighting conditions have on task performance. It is this uniqueness which makes it impossible to generalize from the effect of lighting on the performance of one task to the effect of lighting on the performance of another. The RVP model for on-axis tasks and the visual search models discussed above can be used to quantify the effects of lighting conditions on visual performance, but there is no general model to translate those results to task performance.

### 6.4  Approaches to Improving Suprathreshold Visual Performance

The main purpose of lighting installations is to ensure that people can perform the work they need to do quickly, easily, comfortably, and safely. To achieve this desirable aim, it is necessary to provide lighting which ensures people are working on the plateau of visual performance and not on the escarpment. The RVP model of visual performance provides a simple means of checking whether lighting is adequate for the visual performance of many on-axis tasks. The visibility lobe provides an approach to quantifying the effect of lighting conditions on visual search tasks. Alternatively, most countries have well-established recommendations for the illuminances to be provided for working interiors (IESNA, 2011; CIBSE, 2009). Most of these recommendations easily exceed what would be deduced as necessary from a consideration of visual performance alone.

Although the discussion above has focused on lighting conditions, it is important to recognize that improving suprathreshold visual performance can be achieved by changing the characteristics of the task as well as the lighting. The following list is divided into two parts: task changes and lighting changes. Not all of the following suggestions will be possible in every situation and not all are appropriate for every problem.

#### Changing the Task:

- Increase the size of the detail in the task.
- Increase the luminance contrast of the detail in the task.
- For off-axis tasks in a cluttered field, make the object to be detected clearly differ from the surrounding objects on as many different dimensions as possible, for example, size, contrast, color, and shape.
- Ensure the object presents a clear, sharp image on the retina.

#### Changing the Environment:
- Increase the adaptation luminance.
- Where the task involves color, select a lamp with better color properties.

- Design the lighting so that it is free from disability glare and veiling reflections (see Section 7).
- Design the lighting to increase the apparent size or luminance contrast of the object.

## 7  EFFECTS ON COMFORT

Lighting installations are rarely designed for visual performance alone. Visual comfort is almost always a consideration. The aspects of lighting which cause visual discomfort include those relevant to visual performance and extend beyond them. This is because the factors relevant to visual performance are generally restricted to the task and its immediate area, whereas the factors affecting visual discomfort can occur anywhere within the lit space.

### 7.1  Symptoms and Causes of Visual Discomfort

Visual discomfort can give rise to an extensive list of symptoms. Among the more common are red, sore, itchy, and watering eyes; headaches and migraine attacks; and aches and pains associated with poor posture. Visual discomfort is not the only possible source of these symptoms. All can have other causes. It is this vagueness which makes it essential to consider the nature of the visual environment before ascribing any of these symptoms to the lighting conditions.

Features of the visual environment that can cause visual discomfort are as follows:

*Visual Task Difficulty*. The visual system is designed to extract information from the visual environment. Any visual task that is close to threshold contains information that is difficult to extract. The usual reaction to a high level of visual difficulty is to bring the task closer to increase its visual size. As the task is brought closer, the accommodation mechanism of the eye has to adjust to keep the retinal image sharp. This adjustment can lead to muscle fatigue and hence symptoms of visual discomfort.

*Under- and Overstimulation*. The visual system is designed to extract information from the visual environment. Discomfort occurs either when there is no information to be extracted or when there is an excessive amount of repetitive information. Examples of no information occur when driving in fog or in a "whiteout" snowstorm. In both cases, the visual system is searching for information which is hidden but which may appear suddenly and require a rapid response. The stress experienced while driving in these conditions is a common experience. As for overstimulation, the important point is not the total amount of visual information, but rather the presence of large areas of the same spatial frequency. Wilkins (1993) has associated the presence of large areas of specific spatial frequencies in printed text with the occurrence of headaches, migraines, and reading difficulties.

*Distraction*. The visual system is designed to extract information from the visual environment. To do this, it has a large peripheral field which detects the presence of objects which are then examined using the small, high-resolution fovea. For this system to work, objects in the peripheral field that are bright, moving, or flickering have to be easily detected. If, upon examination, these bright, moving, or flickering objects prove to be of little interest, they become sources of distraction because their attention-gathering power is not diminished after one examination. Ignoring objects that automatically attract attention is stressful and can lead to symptoms of visual discomfort.

*Perceptual Confusion*. The visual system is designed to extract information from the visual environment. The visual environment consists of a pattern of luminances developed from the differences in reflectance of the surfaces in the field of view and the distribution of illuminance on those surfaces. Perceptual confusion occurs when there is a pattern of luminances present which is solely related to the illuminance distribution and conflicts with the pattern of luminance associated with the reflectances of the surfaces.

## 7.2 Lighting Conditions That Can Cause Discomfort

There are many different aspects of lighting that can cause discomfort. Insufficient light for the performance of a task has been discussed earlier and will not be discussed again. Rather, attention will be devoted to flicker, glare, shadows, and veiling reflections. It should be noted that whether or not these aspects of lighting cause discomfort will depend on the context. All can be used to positive effect in some contexts.

**Flicker**  A lighting installation that produces visible flicker will be almost universally disliked unless it is being used for entertainment. Large individual differences in the sensitivity to flicker imply that a clear safety margin is necessary. This can be achieved by high-frequency operation and/or the mixing of light from lamps powered from different phases of the electricity supply. The same approaches, which will result in a changed frequency and/or a reduced percentage modulation, can be used to diminish any stroboscopic illusions. The use of high-frequency control gear has been associated with a reduction in the prevalence of headaches under fluorescent lighting (Wilkins et al. 1989).

**Glare**  Glare occurs in two ways. First, it is possible to have too much light. Too much light produces a simple photophobic response in which the observer screws up his eyes, blinks, or looks away. Too much light is rare indoors but is common in full sunlight. Second, glare occurs when the range of luminances in a visual environment is too large. Glare of this sort can have two effects: a reduction in threshold visual performance and a feeling of discomfort. Glare which reduces threshold visual performance is called disability glare. It is due to light scattered in the eye reducing the luminance contrast of the retinal image on the fovea. The magnitude of disability glare can be estimated by calculating the equivalent veiling luminance (IESNA, 2011).

The effect of disability glare on the luminance contrast of the object being looked at can be determined by adding the equivalent veiling luminance to all elements in the formulas for luminance contrast (see Section 4.6.2). Disability glare is rare in interior lighting but is common on roads at night from oncoming headlights and during the day from the sun. Usually disability glare also causes discomfort, but it is possible to have disability glare without discomfort when the glare source is large in area. This can be seen by looking at a picture hung on a wall adjacent to a window. The picture will usually be much easier to see when the eye is shielded from the window.

As for discomfort glare, this, by definition, does not cause any shift in threshold visual performance but does cause discomfort. There are many different national systems for predicting the magnitude of discomfort glare produced by interior lighting installations (IESNA, 2011; CIBSE, 2009; CIE, 2002). All these systems are based on formulas that imply that discomfort glare increases as the luminance and solid angle of the glare source increase and decreases as the luminance of the background and the deviation from the glare source increase. Lighting equipment manufacturers use these formulas to produce tabular estimates of the level of discomfort glare produced by a regular array of their luminaires for a range of standard interiors. These tables provide all the precision necessary for estimating the average level of discomfort glare likely to occur in an interior, although the precision with which they predict an individual's sense of discomfort is low (Stone and Harker, 1973).

**Shadows**  Shadows are cast when light coming from a particular direction is intercepted by an opaque object. If the object is big enough, the effect is to reduce the illuminance over a large area. This is typically the problem in industrial lighting where large pieces of machinery cast shadows in adjacent areas. The effect of these shadows can be overcome either by increasing the proportion of interreflected light by using high reflectance surfaces or by providing local lighting in the shadowed area. If the object is smaller, the shadow can be cast over a meaningful area, which in turn can cause perceptual confusion, particularly if the shadow moves. An example of this is the shadow of a hand cast on a blueprint. This problem can be reduced by increasing the interreflected light in the space or by providing local lighting which can be adjusted in position.

Although shadows can cause visual discomfort, it should be noted that they are also an essential element in revealing the form of three-dimensional objects. Techniques of display lighting are based around the idea of creating highlights and shadows to change the perceived form of the object being displayed.

The number and nature of shadows produced by a lighting installation depends on the size and number of light sources and the extent to which light is

interreflected around the space. The strongest shadow is produced from a single point source in a black room. Weak shadows are produced when the light sources are large in area and the degree of interreflection is high.

***Veiling Reflections***  Veiling reflections occur when a source of high luminance, usually a luminaire or a window, is reflected from a specularly reflecting surface, such as a glossy printed page or a display screen. The luminance of the reflected image changes the luminance contrast of the printed text or the display. The extent to which this changes visual performance can be estimated using the RVP model, but the extent to which it causes discomfort is different. Bjorset and Fredericksen (1979) have shown that a 20% reduction in luminance contrast is the limit of what is acceptable, regardless of the luminance contrast without veiling reflections (Figure 13).

The two factors that determine the magnitude of veiling reflections are the specularity of the material being viewed and the geometry between the observer, the object, and any sources of high luminance. If the object is completely diffusely reflecting, no veiling reflections occur, but if it has a specular reflection component, veiling reflections can occur. The positions where they occur are those where the incident ray corresponding to the reflected ray which reaches the observer's eye from the object comes from a source of high luminance. This means that the strength and magnitude of veiling reflections can vary dramatically within a single lighting installation (Boyce and Slater, 1981).

Like shadows, veiling reflections can also be used positively, but when they are, they are conventionally called highlights. Display lighting of specularly reflecting objects is all about producing highlights to reveal the specular nature of the surface.

### 7.3  Comfort, Performance, and Expectations

While lighting conditions that make it difficult to achieve good visual performance will almost always
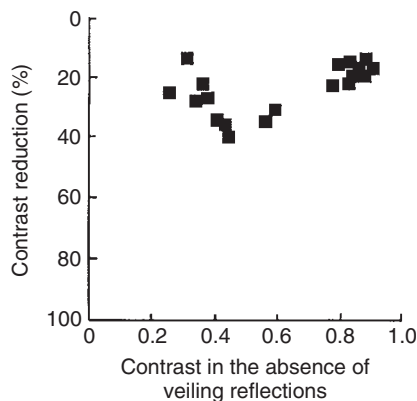
**Figure 13** Luminance contrast reduction considered acceptable by 90% of observers plotted against luminance contrast of materials when no veiling reflections occurred. (after Bjorset, H. H., and Fredericksen, E. A., *Proceedings of the 19th Session of the CIE,* 1979.)

be considered uncomfortable, lighting conditions that allow a high level of visual performance may also be considered uncomfortable. Figure 14 shows the mean detection speed for finding a number from many laid out at random on a table, and the percentage of people considering the lighting good. As might be expected, increasing the illuminance on the table increases mean detection speed and the percentage considering the lighting good. However, as the illuminance exceeds 2000 lx, the percentage considering the lighting good declines even though the mean detection speed continues to increase. This result indicates that if you wish to achieve a satisfactory lighting installation it is necessary to provide lighting which allows easy visual performance and avoids discomfort and that visual discomfort is more sensitive to lighting conditions than visual performance.

There is another aspect of visual comfort which distinguishes it from visual performance. Visual performance is determined solely by the capabilities of the visual system. Visual comfort is linked to peoples' expectations. Any lighting installation which does not meet expectations may be considered uncomfortable even though visual performance is adequate; and expectations can change over time. Figure 12 also demonstrates another potential impact of visual comfort. Lighting conditions which are considered uncomfortable may influence task performance by changing motivation even when they have no effect on the stimuli presented to the visual system and hence on visual performance.

### 7.4  Approaches to Improving Visual Comfort

In order to ensure visual comfort it is necessary to ensure that the lighting allows a good level of visual performance, does not cause distraction, and allows sufficient stimulation without perceptual confusion. This can be done by

- Identifying the visual tasks to be performed and then determining the characteristics of the lighting needed to allow a high level of visual performance of the tasks (see Sections 5 and 6)
- Eliminating flicker from the lighting by using appropriate control gear for discharge and solid state lamps. If this is not possible, reduce the modulation of the flicker by mixing light from sources operating on different phases of the electricity supply
- Reducing disability glare by careful selection, placing, and aiming of luminaires so as to reduce the luminous intensity of the luminaires close to the common lines of sight
- Reducing discomfort glare by careful selection and layout of luminaires. Use the appropriate national discomfort glare system to estimate the magnitude of discomfort glare. Using high reflectance surfaces in the space will help reduce discomfort glare by increasing the background luminance against which the luminaires are seen
- Considering the density and areal extent of any shadows which are likely to occur. If shadows
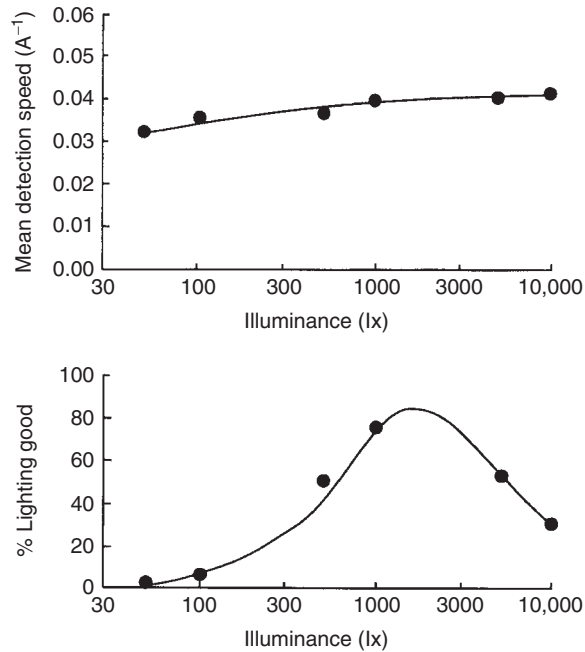
**Figure 14**  Mean detection speed for locating a specified number from amongst others at different illuminances, and the percentage of observers who consider the lighting good at each illuminance. (after Muck, E., and Bodmann, H. W., *Lichttechnik*, Vol. 13, p. 502, 1961.)

are undesirable and large area shadows are likely to occur, use high reflectance surfaces in the space to increase the amount of interreflected light and use more lower-wattage lamps to supply the desired illuminance. If shadows cannot be avoided because of the extent of obstruction in the space, be prepared to provide supplementary task lighting in the shadowed areas. If dense, small area shadows occur in the immediate work area, use adjustable task lighting to moderate their impact

- Considering the extent to which veiling reflections (or highlights) are desirable. If they are undesirable, veiling reflections can be reduced by
  - Reducing the specular reflectance of the surface being viewed
  - Changing the geometry between the viewer, the surface being viewed, and the offending zone
  - Reducing the luminance of the offending zone
  - Increasing the amount of inter-reflected light in the space

## 8  INDIVIDUAL DIFFERENCES

Differences between individuals in visual capabilities are common and are usually dealt with by providing lighting which is more than adequate for visual performance and visual comfort. However, there are three sources of individual differences which are both common and consistent enough in direction to deserve special consideration. They are the effects of age, partial sight, and defective color vision.

### 8.1  Changes with Age

As the visual system ages, a number of changes in its structure and capabilities occur. Usually, the first to occur is an increase in the near point, i.e., the shortest distance at which a clear, sharp retinal image can be achieved. This increase occurs due to an increase in the rigidity of the lens with age. This change, called presbyopia, is why the majority of people over fifty have to wear glasses or contact lenses to read.

While the increasing rigidity of the lens, and other forms of focusing difficulty, can be compensated by adjusting the optical power of the eye's optical system with lenses, the other changes that occur in the eye cannot. As the visual system ages, the amount of light reaching the retina is reduced, more of the light entering the eye is scattered, and the color of the light is altered by preferential absorption of the short, visible wavelengths. The rate at which these changes occur accelerates after about sixty. The consequences of these changes with age are reduced visual acuity, reduced contrast sensitivity, reduced color discrimination, increased time taken to adapt to large and sudden changes in adaptation luminance, and increased sensitivity to glare (Boyce, 2003).

Lighting can be used to compensate for these changes, to some extent. Older people benefit from higher illuminances than are needed by young people

(Smith and Rea, 1979), but simply providing more light may not be enough. The light has to be provided in such a way that both disability and discomfort glare are carefully controlled and veiling reflections are avoided. Where elderly people are likely to be moving from a well lit area to a dark area, such as from a supermarket to a parking lot, a transition zone with a gradually reducing illuminance is desirable. Such a transition zone allows their visual system more time to make the necessary changes in adaptation.

## 8.2 Helping People with Low Vision

Low vision is a state of vision that falls between normal vision and total blindness. The World Health Organization has a system for classifying vision from normal to total blindness. Low vision occurs when the visual acuity in the better eye is less than 6/18 or the visual field is less than 10 degrees. A visual acuity of 6/18 means that the individual can just resolve details at 6 m which people with normal vision can resolve at 18 m.

While some people are born with low vision, the majority of people with low vision are elderly. Kahn and Moorhead (1973) found that among those with low vision, 20 percent reached this state between birth and 40 years, 21 percent between 41 and 60 years and 59 percent after 60 years of age. Surveys in the United States and the United Kingdom suggest that the percentages of the total population who are classified as having low vision are in the range 0.5 to 1 percent. This percentage increases markedly in less developed countries (Tielsch, 2000).

The three most common causes of low vision are cataract, macular degeneration, and glaucoma. These causes involve different parts of the eye and have different implications for how lighting might be used to help.

Cataract is an opacity developing in the lens. The effect of cataract is to absorb and scatter more light as the light passes through the lens. This change results in reduced visual acuity and reduced contrast sensitivity over the entire visual field and greater sensitivity to glare. The extent to which more light can help a person with cataract depends on the balance between absorption and scattering. More light will help overcome the increased absorption but if scattering is high, the consequent deterioration in the luminance contrast of the retinal image will reduce visual capabilities. There is really little alternative to testing the effectiveness of additional light on an individual basis. What is true for everyone with cataract is that they will be very sensitive to glare from luminaires and windows. Careful selection of luminaires and window treatments to limit glare is desirable. The use of dark backgrounds against which objects are to be seen will also help.

Macular degeneration occurs when the macular of the retina, which is just above the fovea, becomes opaque due to bleeding or atrophy. An opacity immediately in front of the fovea implies a serious reduction in visual acuity and in contrast sensitivity at high spatial frequencies. It also implies that the ability to discriminate colors will be reduced. Typically, these changes make reading difficult if not impossible.

However, peripheral vision is unaffected so the ability to find ones way around is unchanged. Providing more light, usually by way of a task light, will help people in the early stage of macular deregulation to read, although as the deterioration progresses additional light will be less effective. Increasing the size of the retinal image by magnification or by getting closer is helpful at all stages.

Glaucoma is shown by a progressive narrowing of the visual field. Glaucoma is due to an increase in intraocular pressure which damages the retina and the anterior optic nerve. Glaucoma will continue until complete blindness occurs unless the intraocular pressure is reduced. As glaucoma develops it leads to a reduction in visual field size, reduced contrast sensitivity, poor night vision, and slowed transient adaptation but the resolution of detail seen on-axis is unaffected until the final stage. Lighting has limited value in helping people in the early stages of glaucoma, because where damage has occurred the retina has been destroyed. However, consideration should be given to providing enough light for exterior lighting at night to enable the fovea to operate.

While the extent to which providing more light is helpful will depend on the specific cause of low vision, there is one approach that is generally useful. This approach is to simplify the visual environment and to make its salient details more visible. As an example, consider the problem of how to set a table so that a person with low vision can eat with confidence. The plate containing the food and the associated cutlery can be made more visible by using a contrasting tablecloth, e.g., a dark tablecloth with a white plate and cutlery. The food on the plate can be made easier to identify by using an overlarge plate so that individual food items can be separated from each other. The whole scene can be simplified can be using solid colors rather than patterns. This same approach of simplification and enhanced visibility can be applied to whole rooms, for example, by painting a door frame in a contrasting color to the door so that the door is easily identified.

## 8.3 Consequences of Defective Color Vision

About 8 percent of males and 0.5 percent of females have some form of defective color vision (McIntyre, 2002). For most activities this causes few problems, either because the exact identification of color is unnecessary or because there are other cues by which the necessary information can be obtained. Where defective color vision does become a problem is where color is the sole means used to identify significant information as, for example, in some forms of electrical wiring. People with defective color vision will have difficulty with such activities (Steward and Cole, 1989).

Where self-luminous colors are used as signals, care should be taken to restrict the range of colored lights used to those which can be distinguished by people with the most common forms of color defect. For example, the CIE has recommended areas on the CIE 1931 chromaticity diagram within which red, green, yellow, blue, and white signal lights should lie. These areas are designed so that the red signal will be named as red and the green as green by people with the most common forms of defective color vision (CIE, 1994).

## 9   OTHER EFFECTS OF LIGHT ENTERING THE EYE

Although making things visible is the most obvious effect of light entering the eye, there are two other ways in which light can affect us. The first is through the circadian system. The second is through the psychological impact of what is visible (CIE, 2009).

### 9.1   Human Circadian System

Light entering the eye does more than stimulate the visual system. It also influences the human circadian system and hence our biological rhythms. The human circadian system has three components: an internal oscillator, a number of external oscillators that entrain the internal oscillator, and a messenger hormone, melatonin, that carries the internal "time" information to all parts of the body through the bloodstream (Djik et al., 1995). The light – dark cycle is one of the most potent of the external stimuli for entrainment. By varying the amount of light exposure and when it is presented, it is possible to shift the phase of the circadian system clock, either forward or backward, as required. In addition, it is possible to have an immediate alerting effect by exposure to light during circadian night (Boyce, 2003). The amount of light required to produce phase shifts or an immediate alerting effect is within the range of current lighting practice. However, the spectral sensitivity of the circadian system is not the same as the visual system, the peak sensitivity being about 480 nm (Brainard et al., 2001; Thapan et al., 2001). This is because a different photoreceptor is used by the circadian system (Berson et al., 2002), although there is evidence that there is some interaction between the circadian photoreceptor and the visual photoreceptors (Bullough et al., 2008). This means that the effectiveness of light sources for stimulating the circadian system cannot be evaluated using the CIE photometry system (see Section 2).

The growing understanding of the importance of the light dark cycle for circadian rhythms has significance for human health and well-being. One application where exposure to light has been of interest is for shiftwork. The short term problems of shift work are fatigue, produced by poor quality sleep, and maintaining alertness during work. Long term, there is evidence that shift workers have a higher risk of cardiovascular disease, gastrointestinal ailments, and emotional and social problems. The short term problems are believed to occur because of a mismatch between the demands of the work and the state of the worker's circadian rhythm. Put plainly, the workers are expected to work when their physiology is telling them to sleep and sleep when their physiology is telling them to be awake. Light is useful in alleviating this problem because it can more rapidly shift the human circadian rhythm so that it better matches the functional requirements, but to do this requires control of light exposure for the complete twenty-four hours (Eastman et al. 1994). As for the immediate alerting affect, improvements in alertness and cognitive performance have been found following exposure to high light levels during night shift work, together with physiological changes indicative of the state of the circadian rhythm (French et al., 1990; Boyce et al., 1997).

Another human health problem that has been shown to be sensitive to light exposure is seasonal affective disorder (SAD). People with this condition typically experience decreased energy and stamina, depression, feelings of despair and a greater need for sleep during the winter months. Light therapy, in which the patient is exposed to a high illuminance for a set period each day, has been shown to alleviate these symptoms in many patients (Lam and Levitt, 1999).

The uses of light to alleviate the problems of shift work and to treat seasonal depression are just the most advanced examples of the influence of light on human well-being. Other applications of light therapy include the treatment of sleep disorders, more general, non-seasonal depression and jet lag as well as the alleviation of the fractured sleep/wake cycles of Alzheimer's patients. Also of interest is what damaging side-effects exposure to light during circadian night might have (Brainard et al., 1999; Figueiro et al., 2006). Until a clearer understanding of the positive and negative impacts of exposure to light during circadian night is achieved, it would be wise to treat attempts to use light exposure to manipulate such a fundamental part of our physiology as the circadian system with caution (Boyce, 2006).

### 9.2   Positive and Negative Affect

Psychology is a vast field and the psychology of lighting is only a small part of it. The area relevant to lighting practice that has been most consistently studied is that of perception. Studies have been undertaken in abstract situations and have lead to quantitative relations being proposed between simple sensations such as brightness and photometric measurements such as luminance (Boyce, 2003). Other studies have been undertaken in rooms with complete lighting installations and have lead to an understanding of the link between the perception of gloom and such photometric characteristics of the room as reflectance and illuminance distributions (Shepherd et al., 1989). Yet others have tried to establish if lighting generates cues by which people interpret a room, for example, does lighting the walls enhance the perception of spaciousness (Flynn et al., 1973).

While such studies have certainly influenced lighting design they cannot be said to constitute a coherent body of knowledge. Further, they cannot form a basis for lighting practice until the impacts of specific perceptions are understood. To understand the consequences of perception of lighting it is necessary to take a broader view. This view centers around positive affect. Positive affect, defined as pleasant feelings induced by commonplace events or circumstances, has been found to influence cognition and social behavior (Isen and Baron, 1991). Specifically, positive affect has been shown to increase efficiency in making some type of decisions, and to promote innovation and creative problem solving. It also changes the choices people make and the judgments they deliver. For example, it has been shown to alter peoples' preference for resolving conflict by collaboration rather

than avoidance and also to change their opinions of the tasks they perform.

Given these usually desirable outcomes of positive affect, it is necessary to ask what can generate positive affect. The answer is both small and wide. Small, because the stimuli which have been shown to generate positive affect are low-level stimuli, ranging from receiving a small but unexpected gift from a manufacturer's representative to being given positive feedback about task performance. Wide, because positive affect can be influenced by the physical environment, the organizational structure, and the organizational culture. Lighting is clearly a part of the physical environment and has been shown to influence positive affect (Baron et al., 1992; McCloughan et al., 1999) but it is only one of many factors that can do that.

As would be expected, it is also possible to generate negative affect. There is considerable information on the influence of frustration or anger on aggression and on the relationship between anxiety and performance (Baron, 1977). It seems reasonable to propose that lighting conditions that cause visual discomfort could generate negative affect.

Positive and negative affect provide plausible routes whereby the perception of the visual environment might influence the efficiency and effectiveness of organizations. As such, they represent a very different approach to identifying what is the most appropriate form of lighting for organizations to the visibility based recommendations used in lighting practice today. The possibility that improving the quality of lighting beyond that required for good visibility without discomfort would lead to enhanced organizational performance is a topic of current interest (CIE, 1998a; Boyce et al., 2006a). Somewhat encouraging for this belief is the finding that lighting perceived to be of better quality has been shown to be reliably linked to more positive feelings of health and well-being (Veitch et al., 2008).

## 10   TISSUE DAMAGE

The part of the electromagnetic spectrum from 100 nm to 1 mm is called optical radiation. This part of the electromagnetic spectrum covers ultraviolet (100–400 nm), visible (400-760 nm) and infrared radiation (760 nm–1 mm). Sunlight and electric light sources all emit optical radiation. In sufficient quantities, optical radiation can cause damage to the eye and the skin. Details are given in CIE (2006).

### 10.1   Mechanisms for Damage to Eye and Skin

There are two mechanisms for tissue damage to occur; photochemical and thermal. They are not mutually exclusive, both can occur for the same incident optical radiation but one will have a lower damage threshold than the other. Photochemical damage is related to the energy absorbed by the tissue within the repair or replacement time of the cells of the tissue. Thermal damage is determined by the magnitude and duration of the temperature rise.

The factors that determine the likelihood of tissue damage are the spectral irradiance incident on the tissue, the spectral sensitivity of the tissue, the time for which the radiation is incident and, for thermal damage, the area over which the irradiance occurs. Spectral irradiance will be determined by the spectral radiant intensity of the source of optical radiation; the spectral reflectance and/or the spectral transmittance of materials from which the optical radiation is reflected or through which it is transmitted; and the distance from the source of optical radiation. Area is important for thermal tissue damage because the potential for dissipating heat gain is greater for a small area than a large area.

The visual system provides an automatic protection from tissue damage in the eye, for all but the highest levels of visible radiation. This is the involuntary aversion response produced when viewing bright light. The response is to blink and look away, thereby reducing the duration of exposure. Of course, this involuntary response only works for sources that have a high visible radiation component, such as the sun. Sources that produce large amounts of ultraviolet and infrared radiation with little visible radiation are particularly dangerous because they do not trigger the aversion response.

### 10.2   Acute and Chronic Damage to Eye and Skin

Tissue damage can be classified according to the duration of exposure it takes to produce the damage. Acute forms of damage are detectable immediately or at least within a few hours of exposure. Chronic forms of damage only become apparent after many years.

Ultraviolet radiation incident on the skin produces an immediate pigment darkening, followed a few hours later by erythema (reddening of the skin) and, ultimately, by a tan, produced by an increase in the number, size and pigmentation of melanin granules. Excessive ultraviolet radiation incident on the eye can produce, a few hours later, an inflammation of the cornea called photokeratitis. This typically lasts a few days followed by recovery. As for chronic damage, prolonged exposure to ultraviolet radiation has been shown to be associated with various forms of skin cancer and cataract.

Visible radiation incident on the skin will produce erythema but not tanning, and, in sufficient quantity, skin burns. Visible radiation incident on the eye reaches the retina. This irradiance represents both an acute photochemical and an acute thermal hazard to the eye. Photochemical damage to the retina is associated with short wavelength light (blue light). The thermal damage covers retinal burns. As for chronic damage, it may be that prolonged and repeated exposure to light is involved in the retinal aging process.

Infrared radiation incident on the skin again initially produces skin reddening and, at a high enough irradiance, burns. Infrared radiation incident on the eye will cause heating of various elements of the eye, depending on the spectral content of the irradiance and the spectral transmittance of the various components of the eye. Infrared radiation from 760 nm to 1400 nm will reach

the retina and can cause retinal burns. Longer wavelengths will be absorbed by other components of the eye. Prolonged heating of the lens is believed to be involved in the incidence of cataract.

## 10.3   Damage Potential of Different Light Sources

The light source with the greatest potential for tissue damage is the sun. The sun produces copious amounts of ultraviolet, visible, and infrared radiation. Voluntary staring at the sun is a common cause of retinal burns. Voluntary exposure of the skin to the sun commonly produces sunburn. However, there exist some electric light sources which can be hazardous, some being intended for lighting and others being used as a source of optical radiation for industrial processes.

The extent to which a light source represents a hazard can be evaluated by applying the recommendations of the American Conference of Governmental Industrial Hygienists (ACGIH, 2009), using recommended procedures (IESNA, 2000, 2005, 2007; CIE, 1998b, 2006). These recommendations take several different forms, ranging from maximum permissible exposure times to irradiance limits. Application of these standards to various electric light sources indicates that such sources, as conventionally used for interior lighting, rarely represent a hazard (McKinlay et al., 1988; Bergman et al., 1995; Kohmoto, 1999).

## 10.4   Approaches to Limiting Damage

The approach to minimizing the damage caused by optical radiation is to limit the irradiance and/or the time of exposure. Whether any such action is necessary can be determined by applying the ACGIH recommendations to the situation.

For sources of optical radiation used for lighting, if the threshold limiting values are exceeded, it will often be possible to use a different light source that is less hazardous. If this is not possible then it is necessary to filter the source to eliminate some of the hazardous wavelengths or to use some form of eye or skin protection to attenuate the optical radiation or to limit the exposure time.

For sources of optical radiation used in industrial processes, the source should be installed in an enclosure, with an interlock so that opening the enclosure extinguishes the source. If this is not possible, then appropriate forms of eye and skin protection are required.

## 11   EPILOGUE

Illumination has been a subject of study for more than ninety years. The result has been a growing understanding of how lighting conditions and the visual system interact to facilitate visual performance and diminish visual discomfort. This knowledge has formed the framework around which many national illuminating engineering organizations have built recommendations for lighting practice (IESNA 2011; CIBSE, 2009). These recommendations provide a firm basis for designing everyday lighting installations, provided the recommendations are applied with thought and not by rote.

There are three current areas of study with considerable potential to change lighting practice:

- The value of better lighting quality for the efficiency of organizations.

  If it can be shown that better quality lighting has a reliable economic impact on organizational efficiency, the economics of lighting will be dramatically changed.

- The effect of light spectrum in mesopic conditions.

  In mesopic conditions light sources which stimulate the rod photoreceptors more produce a perception of greater brightness and allow superior off-axis performance (Rea et al, 2009; Fotios and Cheal, 2009). Such findings are likely to change both the light sources and the design recommendations for exterior lighting

- The non-visual effects of light.

  If lighting can be shown to influence the health and capabilities of people in everyday situations above and beyond allowing them to see, then the whole basis of lighting recommendations may need to be changed.

## REFERENCES

American Conference of Governmental Industrial Hygienists (ACGIH) (2009), "TLVs and BEIs Threshold Limit Values for Chemical Substances and Physical Agents, Biological Exposure Indices," ACGIH, Cincinnati, OH.

Bailey, I., Clear, R., and Berman, S. (1993), "Size as a Determinant of Reading Speed," *Journal of the Illuminating Engineering Society*, Vol. 22, pp. 102–117.

Baron, R. A., (1977), *Human Aggression*, Plenum, New York.

Baron, R. A., Rea, M. S., and Daniels S. G. (1992), "Effects of Indoor Lighting (Illuminance and Spectral Distribution) on the Performance of Cognitive Tasks and Interpersonal Behaviors: The Potential Mediating Role of Positive Affect," *Motivation and Emotion*, Vol. 16, pp. 1–33.

Bergman, R. S., Parham, T. G., and McGowan, T. K. (1995), "UV Emission from General Lighting Lamps," *Journal of the Illuminating Engineering Society*, Vol. 24, pp. 13–24.

Berman, S. M., Fein, G., Jewett, D. L., and Ashford, F. (1993), "Luminance-Controlled Pupil Size Affects Landolt C Task Performance," *Journal of the Illuminating Engineering Society*, Vol. 22, pp. 150–165.

Berman, S. M., Navvab, M., Martin, M. J., Sheedy, J., and Tithof, W. (2006), "A Comparison of Traditional and High Colour Temperature Lighting on the Near Acuity of Elementary School Children" *Lighting Research and Technology*, Vol. 38, pp. 41–52.

Berson, D. M., Dunn, F. A., and Takai, M. (2002), "Phototransduction by Retinal Ganglion Cells That Set the Circadian Clock," *Science*, Vol. 295, pp. 1070–1073.

Bjorset, H. H., and Frederiksen, E. (1979), "A Proposal for Recommendations for the Limitation of the Contrast Reduction in Office Lighting," in *Proceedings of the 19th Session of the CIE*, Kyoto, Japan, Commission Internationale de l'Eclairage, Vienna, pp. 310–314.

Boff, K. R., and Lincoln J. E. (1988), "Engineering Data Compendium: Human Perception and Performance," Harry

G. Armstrong Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, OH.

Bourdy, C., Chiron, A., Cottin, C., and Monor, A. (1987), "Visibility at a Tunnel Entrance: Effect of Temporal Adaptation," *Lighting Research and Technology*, Vol. 19, pp. 35–44.

Boyce, P. R. (1985), "Movement under Emergency Lighting: The Effect of Illuminance," *Lighting Research and Technology*, Vol. 17, pp. 51–71.

Boyce, P. R. (2003), *Human Factors in Lighting*, Taylor and Francis, London.

Boyce, P. R., (2006), "Lemmings, Light and Health" *Light and Engineering*, Vol. 14, pp. 24–31.

Boyce, P. R., and Rea, M. S. (1987), "Plateau and Escarpment: The Shape of Visual Performance," in *Proceedings of the 21st Session of the CIE*, Venice, Italy, Commission Internationale de l'Eclairage, Vienna, pp. 82–85.

Boyce, P. R., and Slater, A. I. (1981), "The Application of Contrast Rendering Factor to Office Lighting Design," *Lighting Research and Technology*, Vol. 13, pp. 65–79.

Boyce, P. R., Beckstead, J. W., Eklund, N. H., Strobel, R. W., and Rea, M. S. (1997), "Lighting the Graveyard Shift: The Influence of a Daylight-Simulating Skylight on the Task Performance and Mood of Night Shift Workers," *Lighting Research and Technology*, Vol. 29, pp. 105–142.

Boyce, P. R., Akashi, Y., Hunter, C. M., and Bullough, J. D. (2003), "The Impact of Spectral Power Distribution on the Performance of an Achromatic Visual Task," *Lighting Research and Technology*, Vol. 35, pp. 141–161.

Boyce, P. R., Veitch, J. A., Newsham, G. R., Jones, C. C., Heerwagen, J., Myer, M. A., and Hunter, C. M. (2006a), "Lighting Quality and Office Work: Two Field Simulation Studies," *Lighting Research and Technology*, Vol. 38, pp. 191–223.

Boyce, P. R., Veitch, J. A., Newsham, G. R., Jones, C. C., Heerwagen, J., Myer, M. A., and Hunter, C. M. (2006b), "Switching and Dimming Behaviour in Offices," *Lighting Research and Technology*, Vol. 38, pp. 358–378.

Brainard, G. C., Kavet, R., and Kheifets, L. I. (1999), "The Relationship between Electromagnetic Field and Light Exposures to Melatonin and Breast Cancer: A Review of the Relevant Literature," *Journal of Pineal Research*, Vol. 26, pp. 65–100.

Brainard. G. C., Hanifin, J. P., Greeson, J. M., Byrne, B., Glickman, G., Gerner, E., and Rollag, M. D. (2001), "Action Spectrum for Melatonin Regulation in Humans: Evidence for a Novel Circadian Photoreceptor," *Journal of Neuroscience*, Vol. 21, pp. 6405–6412.

Bullough, J. D., Bierman, A., Figueiro, M. G., and Rea, M. S. (2008) "On Melatonin Suppression from Polychromatic and Narrow Band Light" *Chronobiology International*, Vol. 25, pp. 653–656.

Chartered Institution of Building Services Engineers (CIBSE) (2009), *CIBSE Code for Lighting*, CIBSE, London.

Commission Internationale de l'Eclairage (CIE) (1978), "Recommendations on Uniform Color Spaces, Color-Difference Equations, Psychometric Color Terms," Supplement 2 to Publication 15, CIE, Vienna.

Commission Internationale de l'Eclairage (CIE) (1983), "The Basis of Physical Photometry," CIE Publication 18.2, CIE, Vienna.

Commission Internationale de l'Eclairage (CIE) (1987), "Methods of Characterizing Illuminance Meters and Luminance Meters: Performance, Characteristics and Specification," CIE Standard 69, CIE, Vienna.

Commission Internationale de l'Eclairage (CIE) (1990), "Spectral Luminous Efficiency Function for Photopic Vision," CIE Publication 86 CIE 1988 $2°$, CIE, Vienna.

Commission Internationale de l'Eclairage (CIE) (1994), "Review of the Official Recommendations of the CIE for the Colours of Signal Lights," CIE Technical Report 107, CIE, Vienna.

Commission Internationale de l'Eclairage (CIE) (1995), "Method of Measuring and Specifying Color Rendering Properties of Light Sources," CIE Publication 13.3, CIE, Vienna.

Commission Internationale de l'Eclairage (CIE) (1998a), "Proceedings of the First CIE Symposium on Lighting Quality," CIE Publication x015-1998, CIE, Vienna.

Commission Internationale de l'Eclairage (CIE) (1998b), "Measurements of Optical Radiation Hazards," CIE Publication x016-1998, CIE, Vienna.

Commission Internationale de l'Eclairage (CIE) (2002), "CIE Collection on Glare," CIE Publication 146-2002, CIE, Vienna.

Commission Internationale de l'Eclairage (CIE) (2004a), "Colorimetry," CIE Publication 15-2004, CIE, Vienna.

Commission Internationale de l'Eclairage (CIE) (2004b), "A Colour Appearance Model for Colour Management Systems: CIECAM02," CIE Publication 159-2004, CIE, Vienna.

Commission Internationale de l'Eclairage (CIE) (2006), "Photobiological Safety of Lamps and Lamp Systems," CIE Standard S009, CIE, Vienna.

Commission Internationale de l'Eclairage (CIE) (2009), "Ocular Lighting Effects on Human Physiology and Behavior," CIE Publication 158-2009, CIE, Vienna.

Cornu, L., and Harlay, F. (1969), "Modifications de la Discrimination Chromatique en Fonction de l'Eclairement," *Vision Research*, Vol. 9, pp. 1273–1280.

Dijk, D-J., Boulos, Z., Eastman, C. I., Lewy, A. J., Campbell, S. S., and Terman, M. (1995), "Light Treatment for Sleep Disorders: Consensus Report, II Basic Properties of Circadian Physiology and Sleep Regulation," *Journal of Biological Rhythms*, Vol. 10, pp. 113–125.

Eastman, C. I., Stewart, K. T., Mahoney, M. P., Liu, L., and Fogg, L. F. (1994), "Dark Goggles and Bright Light Improve Circadian Rhythm Adaptation to Night Shift Work," *Sleep*, Vol. 17, pp. 535–543.

Eklund, N. H., Boyce, P. R., and Simpson, S. N. (2001), "Lighting and Sustained Performance: Modeling Data-Entry Task Performance," *Journal of the Illuminating Engineering Society*, Vol. 30, pp. 126–141.

Figueiro, M. G., Rea, M. S., and Bullough, J. D. (2006), "Does Architectural Lighting Contribute to Breast Cancer," *Journal of Carcinogenesis*, Vol. 5, pp. 20–32.

Flynn, J. E., Spencer, T. J., Martyniuck, O., and Hendrick, C. (1973), "Interim Study of Procedures for Investigating the Effect of Light on Impression and Behavior," *Journal of the Illuminating Engineering Society*, Vol. 3, pp. 87–94.

Fotios, S., and Cheal, C. (2009), "Obstacle Detection: A Pilot Study Investigating the Effect of Lamp Type, Illuminance and Age," *Lighting Research and Technology*, Vol. 41, pp. 321–342.

French, J., Hannon, P., and Brainard, G. C. (1990), "Effects of Bright Illuminance on Body Temperature and Human Performance," *Annual Review of Chronopharmacology*, Vol. 7, pp. 37–40.

Hunt, R. W. G. (1991), "Revised Colour Appearance Model for Related and Unrelated Colours," *Color Research and Application*, Vol. 16, pp. 146–165.

Illuminating Engineering Society of North America (IESNA), (2000) *ANSI/IESNA RP-27.2-00, Recommended Practice for Photobiological Safety for Lamps and Lamp Systems – Measurement Techniques*, New York: IESNA.

Illuminating Engineering Society of North America (IESNA), (2005) *ANSI/IESNA RP-27.1-05, Recommended Practice for Photobiological Safety for Lamps and Lamp Systems – General Requirements*, New York: IESNA.

Illuminating Engineering Society of North America (IESNA), (2007) *ANSI/IESNA RP-27.2-07, Recommended Practice for Photobiological Safety for Lamps and Lamp Systems – Risk Group Classification and Labeling*, New York: IESNA.

Illuminating Engineering Society of North America (IESNA) (2011), *Lighting Handbook*, 10th ed., IESNA, New York.

Inanici, M. N. (2006), "Evaluation of High Dynamic Range Photography as a Luminance Data Acquisition System," *Lighting Research and Technology*, Vol. 38, pp. 123–134.

Inditsky, B., Bodmann, H. W., and Fleck, H. J. (1982), "Elements of Visual Performance, Contrast Metric–Visibility Lobe–Eye Movements," *Lighting Research and Technology*, Vol. 14, pp. 218–231.

Isen, A. M., and Baron, R. A. (1991), "Positive Affect as a Factor in Organizational Behavior," *Research in Organizational Behavior*, Vol. 13, pp. 1–53.

Kahn, H. A., and Moorhead, H. B. (1973), "Statistics on Blindness in the Model Reporting Area 1969–70," Publication No. (NIH) 73-427, U.S. Department of Health, Education and Welfare, Washington, DC.

Koenderink, J. J., and van Doorn, A. J. (1979), "Spatiotemporal Contrast Detection Threshold Surface Is Bimodal," *Optics Letters*, Vol. 4, pp. 32–34.

Kohmoto, K. (1999), "Evaluation of Actual Light Sources with Proposed Photobiological Lamp Safety Standard and Its Applicability to Guide on Lighted Environment," in *Proceedings of the CIE*, 24th Session, Warsaw, Commission Internationale de l'Eclairage, Vienna, pp. 305–309.

Lam, R. W., and Levitt, A. J. (1999), *Canadian Consensus Guidelines for the Treatment of Seasonal Affective Disorder*, Clinical and Academic Publishing, Vancouver, British Columbia.

Littlefair, P. J. (1990), "Innovative Daylighting: Review of Systems and Evaluation Methods," *Lighting Research and Technology*, Vol. 22, pp. 1–17.

Mardaljevic, J., Heschong, L., and Lee, E. (2009), "Daylight Metrics and Energy Saving," *Lighting Research and Technology*, Vol. 41, pp. 261–283.

McCloughan, C. L. B., Aspinall, P. A., and Webb, R. S. (1999), "The Impact of Lighting on Mood," *Lighting Research and Technology*, Vol. 31, pp. 81–88.

McIntyre, D. A. (2002), *Colour Blindness: Causes and Effects*, Dalton, Chester, United Kingdom.

McKinlay, A. F., Harlen, F., and Whillock, M. J. (1988), *Hazards of Optical Radiation*, Adam Hilger, Bristol, United Kingdom.

O'Donell, B., and Colombo, E. (2008), "Simple Reaction Times to Chromatic Stimuli: Luminance and Chromatic Contrast," *Lighting Research and Technology*, Vol. 40, pp. 359–370.

Purves, D., and Beau Lotto R. (2003), *Why We See What We Do: An Empirical Theory of Vision*, Sinauer Associates, Sunderland, MA.

Rea, M. S., and Ouellette, M. J. (1988), "Visual Performance Using Reaction Times," *Lighting Research and Technology*, Vol. 20, pp. 139–153.

Rea, M. S., and Ouellette, M. J. (1991), "Relative Visual Performance: A Basis for Application," *Lighting Research and Technology*, Vol. 23, pp. 135–144.

Rea, M. S., Bullough, J. D., and Akashi, Y. (2009) "Several Views of Metal Halide and High-Pressure Sodium Lighting for Outdoor Applications," *Lighting Research and Technology*, Vol. 41, pp. 297–320.

Robbins, C. L. (1986), *Daylighting, Design and Analysis*, Van Norstrand Reinhold Company, New York.

Robertson, A. R. (1977), "The CIE 1976 Color-Difference Formulae," *Color Research and Application*, Vol. 2, pp. 7–11.

Sekular, R., and Blake, R. (1994), *Perception*, McGraw-Hill, New York.

Shepherd, A. J., Julian, W. G., and Purcell, A. T. (1989), "Gloom as a Psychophysical Phenomenon," *Lighting Research and Technology*, Vol. 21, pp. 89–97.

Simons, R. H., and Bean, A. R. (2000), *Lighting Engineering*, Butterworth-Heinemann, London.

Smith, S. W., and Rea, M. S. (1979), "Relationships between Office Task Performance and Ratings of Feelings and Task Evaluations under Different Light Sources and Levels," in *Proceedings of the 19th Session of the CIE*, Kyoto, Japan, Commission Internationale de l'Eclairage, Vienna, pp. 207–211.

Society of Light and Lighting (2009), *The SLL Lighting Handbook*, Chartered Institution of Building Services Engineers, London.

Steward, J. M., and Cole, B. L. (1989), "What Do Colour Defectives Say about Everyday Tasks," *Optometry and Vision Science*, Vol. 66, pp. 288–295.

Stone, P. T., and Harker, S. P. D. (1973), "Individual and Group Differences in Discomfort Glare Responses," *Lighting Research and Technology*, Vol. 5, pp. 41–49.

Tansley, B. W., and Boynton, R. M. (1978), "Chromatic Border Perception: The Role of Red- and Green-Sensitive Cones," *Vision Research*, Vol. 18, pp. 683–697.

Thapan, K., Arendt, J., and Skene, D. J. (2001), "An Action Spectrum for Melatonin Suppression: Evidence for a Novel Non-Rod, Non-Cone Photoreceptor System in Humans," *Journal of Physiology*, Vol. 535, pp. 261–267.

Tielsch, J. M. (2000), "The Epidemiololgy of Vision Impairment," in *The Lighthouse Handbook on Vision Impairment and Vision Rehabilitation*, B. Silverstone, M. A. Lang, B. P. Rosenthal, and E. E. Faye, Eds., Oxford University Press, New York.

Veitch, J. A., Newsham, G. R., Boyce, P. R., and Jones C. C. (2008), "Lighting Appraisal, Well-Being and Performance in Open Plan Offices: A Linked Mechanisms Approach," *Lighting Research and Technology*, Vol. 40, pp. 133–151.

Wilkins A. (1993), "Reading and Visual Discomfort," in *Visual Process in Reading and Reading Disabilities*, D. M. Willows, R. S. Kruk, and E. Corcos, Eds., Lawrence Erlbaum Associates, Hillsdale, NJ.

Wilkins, A. J., Nimmo-Smith, I., Slater, A. I., and Bedocs, L. (1989), "Fluorescent Lighting, Headaches and Eyestrain," *Lighting Research and Technology*, Vol. 21, pp. 11–18.

Wyszecki, G., and Stiles, W. S. (1982), *Color Science: Concepts and Methods, Quantitative Data and Formulae*, Wiley, New York.

# DESIGN FOR HEALTH, SAFETY, AND COMFORT

# CHAPTER 25

# OCCUPATIONAL HEALTH AND SAFETY MANAGEMENT

**Mark R. Lehto**
**Purdue University**
**West Lafayette, Indiana**

**Bradley T. Cook**
**Miller Engineering, Inc.**
**Ann Arbor, Michigan**

## 1 ABOUT THIS CHAPTER*

This chapter covers the topic of occupational safety and health management from the perspective of an ergonomics practitioner. The chapter begins by presenting a brief history of workplace safety and health issues, including the evolution of legal responsibilities of the employer. Attention then shifts to important elements of occupational safety and health. Topics addressed here include methods of classifying sources and types of occupational injury and illness, causes of accidents, and strategies for preventing or controlling accidents. Next, contemporary methods for managing occupational safety and health are discussed. These methods are primarily instituted through a workplace safety program, which carries out activities such as ensuring compliance with safety and health standards, housekeeping, accident and illness reporting and monitoring, identification and management of workplace hazards, and administering methods of hazard control. Specific occupational hazards are

---

* This chapter is a revised and expanded version of Chapter 17, "Macroergonomics of Occupational Safety and Health," in *Introduction to Human Factors and Ergonomics for Engineers*, by M. Lehto and J. Buck, Taylor and Francis, 2008.

then covered, including dangers from falls, mechanical injury, pressure hazards, electrical hazards, fire and other heat sources, toxic materials, noise hazards, and other vibration problems. The chapter concludes with a discussion of hazard communication.

## 2  INTRODUCTION

Occupational health and safety is an interdisciplinary field that focuses on preventing occupational illnesses and injuries. Government agencies, universities, insurance companies, trade associations, professional organizations, manufacturers, and service organizations all play important roles in reducing the burden of occupational illnesses and injuries to society.

Prevention efforts have paid off dramatically in industrialized countries. The annual number of accidental work deaths per 100,000 workers in the United States, as of 2008, has dropped to about one-eighth of what it was 60 years ago. Nevertheless, the accident toll continues to be high. Single events such as the BP (British Petroleum) oil rig fire and explosion in 2010 can cause multiple deaths and injuries and disrupt entire sectors of the economy, with costs mounting in tens of billions of dollars. Overall, in the United States alone, the National Safety Council (NSC, 2010) estimates that there were about 4300 accidental work deaths and 3,200,000 disabling injuries in 2008, with an associated cost of $183 billion. Major components of this cost estimate include insurance administration costs, wage and productivity losses, medical expenses, and uninsured costs. Worldwide, the direct costs of accidents are estimated to exceed several hundred billion dollars. If indirect costs are considered, such as lost productivity, uninsured damage to facilities, equipment, products, and materials, and the cost of social welfare programs for injured workers and their dependents, the bill raises to several trillion dollars annually. Consideration of pain and suffering raises the cost to an even greater level.

It also should be emphasized that, despite the fact that industrial accident rates are decreasing in most industrialized countries, economic estimates of the annual costs of accidents show the opposite trend. For example, the Liberty Mutual (2009) workplace safety index (WSI) shows an increase of 42.8% in direct U.S. worker compensation costs for the most disabling workplace injuries over the 10-year period from 1998 to 2007. Worker injuries due to overexertion, slipping, tripping, or falling alone accounted for nearly two-thirds of these costs.

The root causes of these and many other types of worker injuries and illnesses can often be addressed through better workplace ergonomics as part of an occupational safety and health program. The application of ergonomic principles to reduce workplace injuries and illnesses began years ago and continues to be a critical element of safety and health management. However, better ergonomics is only part of the solution. Many other important strategies are available that can be used by management to keep risks at an acceptable level.

An important key to success is that, instead of simply reacting to accidents, injuries, or illness, management should be proactively taking steps to prevent them from occurring in the first place.

## 3  SOME HISTORICAL BACKGROUND

Before the nineteenth century, the responsibility for occupational safety and health was placed on the individual employee. The attitude of most companies was, "Let the worker beware!"

With the advent of the industrial revolution in the nineteenth century, workplace injuries became more prevalent. Dangerous machinery caused many accidents. As manufacturing moved to larger plants, there were increasing numbers of workers per plant and an increased need for supervision. There was little theory available to guide management in reducing workplace dangers. This lack of management knowledge and an unfortunate indifference to social concerns caused many safety problems during this era.

During the mid-nineteenth century, the labor movement began as a response to worker concerns about issues such as wages, hours, and working conditions. Unions began to address safety issues. However, industrial safety did not measurably improve. Unions had limited bargaining power as well as limited knowledge about how to effectively reduce workplace hazards. Furthermore, employers did not have a strong legal incentive to improve workplace safety. During this period the law was focused on the "master–servant rule" whereby workers looked to common law to win redress. The employer was not legally obligated to provide a reasonably safe and healthful environment. Employees' actions could void employer liability if:

1.  The risks were apparent and the employee continued to work.
2.  There was some negligence by the employee.
3.  Fellow workers contributed to the injury.

Public acceptance of this situation eventually began to waiver. During the late nineteenth and early twentieth centuries there were numerous changes in the law. One was the federal Employers' Liability Act of 1908, which provided for railroad employees to receive damages for injuries due to negligence of the company or its managers, even when there was some contributory negligence on the part of the injured employee. However, the burden remained for the employee to prove the company was negligent through litigation. During this time, there were also increasing efforts by unions, companies, and trade associations to promote industrial safety. Later came "Workman's Compensation" laws enacted by individual states. These laws provided for the compensation of all injured employees at fixed monetary levels, for the most part eliminating the question of fault and the need for litigation. Most states used a casualty insurance company or a self-insurance

fund to compensate the employee for an accident under these laws. Also during this time, there were some federal government departmental efforts by the Bureau of Mines, the Bureau of Labor, and the National Bureau of Standards to improve worker safety.

Progress since the late 1930s has been impressive. During the early part of that decade, U.S. federal government actions forced management to accept unions as a bargaining agent of employees. Unions became more widespread, and safety conditions were one aspect of the contracts negotiated between employers and unions. During World War II (1939–1946), there were widespread labor shortages, which resulted in increased effort by employers to recruit and retain personnel, including a greater attention to safety for workers.

The Environmental Protection Act of 1969 secured many sources of employee protection against chemical and radiation sources in industry. In 1970 the Occupational Safety and Health Act (Williams-Steiger Act) was passed which placed an obligation on employers to provide a workplace free from recognized hazards. This act also established the Occupational Safety and Health Administration (OSHA), a federal agency whose objective is to develop and enforce workplace safety and health regulations. As shown in Table 1, a large number of laws were added over this period that directly affect worker safety and employers' responsibility toward workers. For most hazards, the enactment of laws and regulations has evolved over time with growing recognition and understanding of their nature. This tendency is illustrated in Table 2, showing how laws, regulations, and standards limiting exposure to lead have evolved over time in the United States over the past 100 years.

## 3.1 Safety And Health Achievements and Future Promises

A number of notable successes and a few failures have been observed since the passage of the Occupational Safety and Health Act of 1970 and the resulting establishment of OSHA and its sister agency, the National Institute for Occupational Safety and Health (NIOSH). One of the most notable successes is the huge reduction in the incidence of occupational poisoning. Previously, many miners were poisoned with phosphorus and many painters and printers were poisoned with lead. There has also been a decline in silicosis, silio-tuberculosis, black lung, and other associated lung diseases that principally occurred in mining but also in the manufacture of asbestos products. Some success has occurred in reducing accidents from physical fatigue and in machine safeguarding; however, several forms of guarding have created losses in productivity. Also some very unsafe factories have been closed down, but those are special cases rather than general cases of success.

Several failures have also been noted. These failures include continuing incidence of: chronic bronchitis, occupational dermatitis (which is the most common industrial occupational disability), diseases from vibration, chemical poisonings other than lead and phosphorus, lesions to bones and joints, chronic vascular

**Table 1  Timeline of Selected Laws and Acts Related to Occupational Safety in United States**

| Year | Legislation |
| --- | --- |
| 1938 | Fair Labor Standards Act |
| 1938 | Federal Product Safety and Health Legislation: Food, Drug, and Cosmetic Act |
| 1953 | Flammable Fabrics Act |
| 1958 | Longshore Safety Act |
| 1959 | Radiation Hazards Act |
| 1960 | Federal Hazardous Substances Act |
| 1966 | Federal Metal and Nonmetallic Mine Safety Act |
| 1966 | National Traffic and Motor Vehicle Safety Act |
| 1966 | Highway Safety Act |
| 1967 | Flammable Fabrics Act Amendment |
| 1968 | Radiation Control for Health and Safety Act |
| 1968 | Fire Research and Safety Act |
| 1969 | Contract Work Hours and Safety Standards Act |
| 1969 | Federal Coal Mine Health and Safety Act |
| 1969 | Environmental Protection Act |
| 1970 | Williams-Steiger Occupational Safety and Health Act |
| 1970 | Poison Prevention Packaging Act |
| 1970 | Lead-Based Paint Poison Prevention Act |
| 1970 | Other Federal Safety and Health Legislation: Clean Air Amendments |
| 1975 | Hazardous Materials Transportation Act |
| 1976 | Toxic Substances Control Act |

impairment, neuroses, and mental disorders. Several of these failures are the focus of current research.

Other ongoing changes in industry are influencing safety and health. One is that during the past two decades U. S. industry has become more capital intensive through automation. One of the positive side effects of automation is that fewer workers are exposed to some of the more hazardous occupational environments associated with tasks such as painting and welding. However, there is a downside to more capital-intensive factories and that is the requirement for greater rates of production and equipment utilization. These higher speeds are stressful mentally and apt to induce physical accidents. They are also inducements for management to cut corners.

Another phenomenon that is affecting safety and health is the enactment of equal-employment opportunity laws by the federal government prohibiting discrimination in employment based on gender, age, race, origin of birth, and religion. Employee screening, such as strength testing, conducted to ensure workers are able to perform certain jobs safely, may disproportionately eliminate older and female applicants. The Equal Employment Opportunity Commission (EEOC) requires evidence that such selection criteria are in fact necessary. Obtaining such evidence can be difficult, so many

**Table 2  Timeline of Lead-Related Regulations, Standards, and Laws in the United States**

| | |
|---|---|
| 1880–1930 | Lead, especially from paint, is suspected by health care professionals of causing deaths to children with previous symptoms of seizure, drop wrist, drop foot, etc. |
| 1910 | Marion Rhodes, a congressman from Missouri, wants to extend labeling provision of the Food and Drug Act to cans of lead paint. His bill is rejected. |
| 1930 | White House Conference on Child Health alerts participants about lead paint on toys. |
| 1930s | Paint manufacturers advise consumers of availability of lead-free paints for toys and cribs. |
| 1933 | Massachusetts — ''Revised Rules, Regulations and Recommendations Pertaining to Structural Painting'' states that ''toys, cribs, furniture and other objects with which infants may come in contact should not be painted with lead colors.'' |
| 1935 | Baltimore Health Department becomes first in nation to offer blood test for lead as a diagnostic test. |
| 1949 | Maryland state legislature passes Toxic Finishes Law, making it ''unlawful for any person to manufacture, sell or offer to sell, any toy or plaything, including children's furniture, decorated or covered with paint or any other material containing lead or other substance of a poisonous nature, from contact with which children may be injuriously affected.'' The law was repealed a year later because of ambiguities of its definitions (''injuriously''), lack of enforceability, and failure to define acceptable levels of lead for paint. |
| 1951 | Baltimore passes ordinance banning the use of paints containing lead pigments, June 29, 1951 |
| 1954 | New York City Health Department approves following label on October 29: ''Contains lead. Harmful if eaten. Do not apply on toys, furniture, or interior surfaces which might be chewed by children.'' |
| 1955 | ASA standard Z66.1 is adopted. The standard limits lead in interior paint to 1%. |
| 1958 | Baltimore mayor signs ordinance mandating warning label on paint cans: ''WARNING — Contains Lead. Harmful if Eaten. Do not apply on any interior surfaces of a dwelling, or of a place used for the care of children, or on window sills, toys, cribs, or other furniture.'' |
| 1963 | Congress passes Clean Air Act. |
| 1970 | Public Law 91-695 is enacted. It provided federal funds for mass screening, treatment, education, and research on how to lessen lead hazard. Also mandated that federally funded public housing meet the 1% paint standard. |
| 1972 | U.S. Food and Drug Administration (FDA) bans paint with an excess of 0.5 percent lead from interstate commerce under provisions of the Federal Hazardous Substances Act. |
| 1974 | Congress asks the Consumer Product Safety Commission (CPSC) to assess danger of multiple layers of paint and to determine a ''safe'' amount of lead in paint. CPSC arrives at 0.5%. |
| 1977 | CPSC adopts 0.06 recommendation of the AAP and National Academy of Sciences (NAS). |
| 1978 | National consensus is reached that lowers lead content of all household paints — interior and exterior — from 1 to 0.06%. |
| 1978 | Tetraethyl lead is removed from gasoline. |

firms prefer to reduce job requirements that require abilities that tend to differ among people by age, gender, and other bases. The Americans with Disabilities Act of 1990 further states that industry must provide access to jobs for qualified individuals with disabilities through the design of processes and workspaces to accommodate their needs when it is not economically infeasible to do so. This law puts the burden of proof for economic infeasibility on the company.

## 4  FUNDAMENTAL ELEMENTS OF OCCUPATIONAL SAFETY AND HEALTH

Much of the improvement in occupational safety and health that has been attained over the past one hundred years can be attributed to two complementary, sometimes overlapping, approaches. The first approach focuses on preventing accidents and mitigating their effects by designing safer systems and taking steps to ensure they operate as intended. The second focuses on controlling the exposure of workers to harmful substances, energy sources, or other environmental or work-related stressors that cause illnesses or cumulative trauma. The two approaches reflect the traditional dichotomy between safety and health functions still present in some organizations but have much in common. In recent years, the trend has been toward combined approaches where safety and health professionals (see Box 1) work together to identify, predict, control, and correct safety and health problems (Goetsch, 2008).

The successful application of either approach requires knowledge and understanding of:

1. Types of hazards encountered by workers and how often they result in injuries, illness, or death
2. Causes of accidents and other forms of exposure to hazards
3. Strategies for preventing or controlling accidents and exposure to hazards and their effectiveness.

<table>
<tr><td colspan="2" style="border:1px solid">

**Box 1: Safety and Health Professionals**

Many safety and health professionals are certified in particular technical specialties. The Board of Certified Safety Professionals administers the Certified Safety Professional (CSP) while the Board of Certification in Professional Ergonomics offers the Certified Professional Ergonomist (CPE) and Certified Human Factors Professional (CHFP). Certification in the practice of industrial hygiene is available from the American Board of Industrial Hygiene. Also, all of the individual states in the United States license engineers as registered professional engineers (PE). Although specific licensing in safety is not available, one of the principles of licensure is safety. The National Society of Professional Engineers has as one its fundamental canons that engineers, in the fulfillment of their professional duties, shall hold paramount the safety, health, and welfare of the public. Those working in the safety area have several professional organizations that they may join, including the American Society of Safety Engineers (ASSE) and the Human Factors and Ergonomics Society (HFES).
</td></tr>
</table>

**Table 3 Occupations with High Incidence of Injuries in the United States 2008**

| Occupations | Number of Injuries And Illness | Incident Rate Per 100,000 Full-Time Workers |
|---|---|---|
| Laborers and freight, stock, and material movers | 79,590 | 440 |
| Heavy and tractor-trailer truck drivers | 57,700 | 362 |
| Nursing aids, orderlies, and attendants | 44,610 | 449 |
| Construction laborers | 31,310 | 383 |
| Retail salespersons | 28,900 | 90 |
| Janitors and cleaners | 28,110 | 243 |
| Light or delivery service truck drivers | 28,040 | 324 |
| General maintenance and repair workers | 20,800 | 213 |
| Registered nurses | 19,070 | 114 |
| Maids and housekeeping cleaners | 18,650 | 278 |
| Carpenters | 18,160 | 236 |
| Stock clerks and order fillers | 18,020 | 130 |

*Source*: Bureau of Labor Statistics.

As expanded upon in the following sections, classifications of occupational injuries and illnesses published by organizations such as the Bureau of Labor Statistics (BLS), NIOSH, and Centers for Disease Control and Prevention (CDC) in the United States provide a good starting point for developing understanding and perspective of particular hazards by identifying patterns of injury and illnesses, risk factors, and levels of exposure. Theories and models of accident causation build upon this perspective by identifying why accidents occur and suggesting methods of hazard control.

## 4.1 Classifications of Occupational Injuries and Illnesses

Organizations such as the BLS and NSC in the United States routinely collect and disseminate information regarding the incidence of occupational injuries and illnesses for particular occupations or industries. The rationale for following this approach is that some occupations have higher rates of certain injuries and disorders than others. Identifying such trends can guide prevention efforts by government regulators, management, and others by focusing attention on the industries and occupations with elevated incident rates. Example statistics of this type collected by the BLS are given in Table 3 for the occupations in the United States with the highest rates of injuries between 2005 and 2008.

More detailed classifications focus on particular elements of hazards, accidents, and injuries. One of the better known coding schemes of this type is the OIICS used by the BLS in the United States to code characteristics of injuries, illnesses, and fatalities. This coding scheme is used by the BLS in their annual Survey of

Occupational Injuries and Illnesses (SOII) and Census of Fatal Occupational Injuries (CFOI) programs. In the OIICS, occupational injuries or illnesses are classified in terms of the nature of injury or illness, body parts affected, primary and secondary sources of injury or illness, and the event or exposure type using a hierarchical coding scheme specified in the OIICS coding manual (http://www.bls.gov/iif/oshoiics.htm). The manual also provides selection rules and coding instructions. At the top level of the coding hierarchy, each major category (or section) is divided into one-digit codes (Table 4). The one-digit codes are then further divided into two-digit codes, and so on. For example, event code "313—Contact with overhead power lines" is a division of the broader code "31—Contact with electric current," which is a division of the even broader event code "3—Exposure to harmful substances or environments."

The source of injury or illness, with well over a thousand four-digit codes, is the largest single category of codes in the OIICS classification. The nature of injury or illness, with several hundred codes, is the second largest category, followed by event or exposure and part of body affected, both with around 300 codes. One of the advantages of the OIICS classification is that its hierarchical structure provides a systematic well-developed way of organizing the huge number of ways occupational injuries or illnesses can occur. For example, hundreds of chemicals or chemical products are systematically grouped into different subcategories within the source of injury or illness category. Patterns of occupational injury and illnesses can also be examined by examining combinations of categories,

**Table 4 Single-Digit Codes Used in Occupational Injury and Illness Classification System (OIICS) Developed by Bureau of Labor Statistics in United States**

Section, Division, and Title

Nature of Injury or Illness

0 Traumatic Injuries and Disorders
1 Systemic Diseases or Disorders
2 Infectious and Parasitic Diseases
3 Neoplasms, Tumors, and Cancer
4 Symptoms, Signs, and Ill-defined Conditions
5 Other Conditions or Disorders
8 Multiple Diseases, Conditions, or Disorders
9999 Nonclassifiable Systemic Diseases or Disorders

Part of Body Affected

0 Head
1 Neck, Including Throat
2 Trunk
3 Upper Extremities
4 Lower Extremities
5 Body Systems
8 Multiple Body Parts
9 Other Body Parts
9999 Nonclassifiable

Source of Injury or Illness

0 Chemicals and Chemical Products
1 Containers
2 Furniture and Fixtures
3 Machinery
4 Parts and Materials
5 Persons, Plants, Animals, and Minerals
6 Structures and Surfaces
7 Tools, Instruments, and Equipment
8 Vehicles
9 Other Sources
9999 Nonclassifiable

Event or Exposure

0 Contact with Objects and Equipment
1 Falls
2 Bodily Reaction and Exertion
3 Exposure to Harmful Substances or Environments
4 Transportation Accidents
5 Fires and Explosions
6 Assaults and Violent Acts
9 Other Events or Exposures
9999 Nonclassifiable

*Source*: http://www.bls.gov/iif/oshoiics.htm.

such as cross referencing the part of body affected for a particular source of injury.

Organizations such as the BLS, NIOSH, and the NSC[*] in the United States routinely follow this approach to describe the nature of injury or illness, body parts affected, source of injury, event or exposure type, and so on, for particular occupations or industries. Other examples include the Liberty Mutual WSI mentioned

---

[*] The NSC publishes summaries of this type each year in Injury Facts.

**Table 5 Top 10 Causes of Disabling Injuries in 2007**

| Cause | Cost ($ Billions) | Percentage |
|---|---|---|
| Overexertion | 12.7 | 24.0 |
| Falls on same level | 7.7 | 14.6 |
| Falls to lower level | 6.2 | 11.7 |
| Bodily reaction (after slipping or tripping) | 5.4 | 10.2 |
| Struck by object | 4.7 | 9.0 |
| Highway incident | 2.5 | 4.7 |
| Caught in/compressed by | 2.1 | 3.9 |
| Repetitive motion | 2.0 | 3.8 |
| Struck against object | 2.0 | 3.8 |
| Assaults/violent acts | 0.6 | 1.1 |

*Source*: Liberty Mutual (2009).

earlier which breaks out direct U.S. workers compensation costs for the most disabling workplace injuries using two-digit OIICS event or exposure codes (Table 5).

Classifications of this type guide injury prevention efforts by identifying particular groups of events (such as overexertion and falls) that can be focused upon with specific control strategies, as expanded upon in Section 6 in this chapter. However, simply identifying the event or exposure that resulted in the injury or illness tells us little about why it occurred. The latter issue has traditionally been approached by studying the causes of accidents, as expanded upon in the following section.

### 4.2 Accident Causation

Over the years, much has been learned about the causes of occupational injuries and illnesses by studying accidents. This work has led to numerous models and theories of accident causation that explain why accidents occur and also suggest generic strategies for accident prevention (Lehto and Salvendy, 1991; Goetsch, 2008).

#### 4.2.1 Early Developments

Some of the earliest research on accident causation was done by Heinrich in the 1920s. Based on his analysis of 75,000 industrial accidents, Heinrich concluded that 88% of the accidents were caused by unsafe acts, 10% by unsafe conditions, and 2% by unpreventable causes. His conclusion that 98% of accidents are potentially preventable by eliminating unsafe acts and conditions was a major departure from the prevailing opinion that industrial accidents were an unavoidable cost of progress and set the stage for a whole host of approaches for accident prevention.

Heinrich also developed what is now known as the Heinrich accident triangle, which showed that for every accident resulting in serious injury 29 resulted in minor injuries and 100 in no injury. Since serious accidents are rare events, many incidents resulting in minor or no injury are likely to occur before particular unsafe acts and conditions cause serious accidents to happen. Consequently, minor accidents and near misses can act as an early warning to an organization that serious problems

are present. The victims of serious accidents may also be unable or unwilling to tell analysts much about what caused the accident. For these and other reasons, it is now well accepted that it is important to study near misses to learn more about the causes of accidents, and many organizations, such as the Nuclear Regulatory Commission (NRC), have established policies of tracking all incidents, regardless of whether there is injury or contamination. Other examples are the Aviation Safety Reporting System (ASRS), administered by the National Aeronautics and Space Administratio (NASA), which collects voluntary safety-related reports submitted by pilots, air traffic controllers, flight attendants, mechanics, and dispatchers. Another example is the National Fire Fighter Near-Miss Reporting System, funded by the U.S. Department of Homeland Security, which collects voluntary reports submitted by fire service professionals.

### 4.2.2 Multifactor Theories

Heinrich went on to develop the domino theory, which organized the sequence of events leading to an injury in terms of five factors (or dominos): social environment and ancestry, fault of person, unsafe act or condition, accident, and finally injury. The domino theory proposed that an injury could be prevented by removing any single factor from the accident sequence. Over the years, a large number of other models and theories have been developed that build upon this basic contribution of the domino theory.

Extensions of Heinrich's early focus on the accident sequence include the development of multifactor models that show how multiple chains of events converge to cause most, if not all, accidents (Lehto and Salvendy, 1991). This approach often involves the use of event and fault trees to show how possible combinations of events and unsafe acts can come together to cause unsafe conditions, accidents, and ultimately injuries. The latter models can be used during probabilistic risk assessment to calculate the criticality of particular event sequences leading to accidents. Other approaches, including network models, such as Benner's multilinear sequencing model, flowcharts, or program evaluation review technique/critical-path method (PERT/CPM) networks, place the events and unsafe acts that lead to unsafe conditions and ultimately an accident on time lines to show both temporal and logical relationships between events and are especially useful during accident investigation to explain how and why accidents occurred.

The epidemiological model provides another useful framework for organizing the multiple factors that may play a role in accident causation. As first proposed by Gordon (1949), factors influencing accidents may be subdivided into those associated with the host (accident victim), agent (deliverer of the injury), and environment (the accident setting). The epidemiological model has guided an immense amount of accident research over the years. Extensions of this model include the industrial accident model (Johnson, 1973), which shows how characteristics of the victim, environment, and accident agent fit into the accident sequence as background factors, initiating factors, intermediate factors, immediate factors, and measurable results.

By organizing a large number of potential accident causes in a meaningful way, the model seems to have significant potential for guiding managerial efforts for reducing accidents. Another example is the Haddon matrix (Haddon, 1975), which organizes causes of traffic accidents and methods of improving traffic safety related to the driver, car, and road environment into preaccident, accident, and pos-accident stages. This framework has been used for years to guide safety programs of the National Highway Traffic Safety Administration.

Haddon (1975) also proposed 10 generic countermeasure strategies that focus on the role of energy as a cause of injury. Each countermeasure falls within different stages of the accident sequence and focuses on different elements of the epidemiological model. The strategies can be paraphrased as (1) prevent the initial buildup of energy, (2) reduce the potential energy, (3) prevent the release of the energy, (4) reduce the rate of release, (5) separate the host from the energy source, (6) place a barrier between the host and energy source, (7) absorb the energy, (8) strengthen the susceptible host, (9) move rapidly to detect and counter the release, and (10) take procedures to ameliorate the damage. The energy model serves a useful purpose in focusing attention on energy as a potential cause of accidents and can be applied in a wide variety of ways in industrial settings. The model can also be extended to generic categories of accidents involving undesired transfers or blockage of material flows. For example, the exposure to toxic materials can be viewed as a flow from some source to a susceptible host.

### 4.2.3 Human Error and Unsafe Behavior

Other models and theories have focused on the important role of human error and intentionally unsafe behavior as a cause of accidents (Lehto and Salvendy, 1991; Reason, 1990; Lehto, 2006). The role and significance of human error can be viewed from many different perspectives. At the most general level, accidents are often the predictable consequence of design errors and management oversights or omissions. Such failures include poorly designed facilities, inadequate risk assessments, safety policies, supervision, and less than adequate operating, inspection, and maintenance procedures. At the operational level, errors are often distinguished as either (1) errors of omission, such as skipping steps in critical procedures or failing to take precautions, or (2) errors of commission, such as operating a machine at the wrong speed. Errors or unsafe acts are also commonly classified in terms of their consequences, their revocability, and their detectability (Altman, 1964). Errors which have delayed consequences are often called latent errors (Reason, 1990). Latent errors and violations, such as failing to reactivate an alarm system after performing maintenance on it, are a particularly important cause of accidents, as they often are not detected or corrected until after an accident happens.

Much effort has also been directed toward determining why errors and violations occur. The overall conclusion is that a large number of performance-shaping factors can play a role, including task demands, social norms, incentives and rewards, operator objectives,

environmental factors, operator skill, and the presence or absence of feedback. One school of thought is that errors and violations are largely due to the lack of proper incentives to behave safely. This perspective is the cornerstone of behavior-based safety (Geller and Williams, 2001), a management system that focuses on developing a supportive culture in which workers receive positive reinforcement for behaving safely. A corollary to this view is that workers are sometimes rewarded for taking shortcuts or failing to follow precautions. This can result in the development of routine, or habitual, patterns of unsafe behavior such as not using personal protective equipment or not using designated pedestrian walkways. Once such patterns of behavior become engrained, it becomes difficult to eliminate them.

Flaws in human decision making (See Chapter 7 in this book) provide another explanation of why people may make poor choices leading to accidents. One issue is that measures of risk perception are sometimes poorly correlated with more objective measures such as fatality rates. People also show a general tendency to weight small, but certain, costs more heavily than large costs that are unlikely to occur. In practice, this effect corresponds to unwillingness to take a precaution that has a small cost (i.e., extra time or effort, inconvenience, or discomfort) that reduces the chance of incurring an unlikely, but serious, consequence (illness, injury, etc.). Risk compensation is another issue. For example, the benefit of modifying a forklift to make it more stable might be reduced because operators start driving faster. People also sometimes show an overreliance on safety devices and technology in general (e.g., a user might rely too heavily on an alerting system to detect a hazard). The latter tendencies are arguably rational reactions in that they are predicted by economic theory if we assume people adjust their choices to maintain an acceptable level of risk.

A large set of other modeling frameworks provide perspective into why errors occur. In particular, models of human information processing suggest that errors can be caused by lapses in attention, distractions, lack of awareness or knowledge, forgetting, time pressure, or information overload. It also can be shown that errors differ depending on the level of task performance. The latter approach has been extended by mapping preferred intervention strategies to modes of error at particular levels of task performance (Table 6). As shown in the table, several different strategies can be followed to address the root causes of errors. The following sections will address these and other strategies for preventing or controlling accidents in more detail.

### 4.3 Control Strategies

Strategies for preventing or controlling accidents and occupational exposure to hazards can be distinguished in several different ways. The most effective approaches often are focused on organizational objectives other than injury or illness prevention, such as increased reliability, quality, or productivity. Such solutions create a "win–win" situation, making it easier to gain support across the organization for their implementation. Simply put, few people would argue with the fact that

**Table 6  Error Modes and Preferred Intervention Strategies**

| Error Mode | Preferred Intervention Strategy |
| --- | --- |
| Routine behavior | *Product and Equipment Design* |
| Fail to perceive hazard condition or forgetting hazard condition — lack of situation awareness (skill based) | Hazard signals |
| | Interruptive features |
| | *Warnings* |
| | Signals: interactive, selective, nonvisual |
| | *Training* |
| | RE: product signals and warnings |
| Forget intended action (rule based) | *Job Design* |
| | Checklists |
| | *Product and Equipment Design* |
| | Hazard signals |
| | Warnings |
| | Reminder signals |
| | *Product and Equipment Design* |
| | Affordances and constraints |
| Psychomotor variability (skill based) | *Product and Equipment Design* |
| | Hazard signals |
| | Error tolerance |
| | *Training* |
| | Skill development |
| | Employee selection |
| Nonroutine behavior (knowledge based) | *Training* |
| | Impart knowledge |
| | *Job Design* |
| | Written procedures |
| | Warnings |
| | Describe precautions and safety procedures |
| Intentional violations (rule based or judgment based) | *Supervision* |
| | Enforcement |
| | Monitoring |
| | *Product and Equipment Design* |
| | Modify costs and benefits of compliance and noncompliance |

*Source*: Lehto (2006).

well-designed physical facilities, equipment, processes, and jobs combined with a well-trained, motivated workforce will result in higher productivity, better quality products and services, and safety.

The basic point is that efforts to improve productivity or other objectives of the organization can also lead to safety improvements. However, in many situations, additional hazard control measures are needed to attain a reasonable level of safety, such as changes in the design of facilities, equipment, processes, and jobs that in some cases conflict with productivity or other objectives of the organization. Many control measures are highly specific to particular hazards, as expanded upon in Section 6 of this chapter, which focuses on common hazards found in the workplace. The following discussion will introduce some general control strategies, beginning with job and process design. Attention will then shift to the so-called hierarchy of hazard control, which provides a way of organizing control measures.

### 4.3.1 Job and Process Design

Good job and process design can eliminate the root cause of accidents and occupational exposure to hazards in many different ways. To a large extent, this approach involves the application of techniques used in the fields of industrial engineering and ergonomics (Salvendy, 2001). Some of these approaches are as follows:

1. Principles of facility and plant layout focus on the separation of people from hazardous operations, efficient patterns of material flow which minimize the need for people to cross traffic areas, properly located and spaced aisles and walkways, and appropriate locations for exits and emergency egress.

2. Value stream mapping focuses on eliminating unnecessary movements of material from location to location, reducing forklift traffic and the need for potentially hazardous exertion while moving items.

3. Methods of production control schedule operations to help avoid ebbs and surges of activity that place excessive demands on operators to rush operations.

4. Methods of workplace layout and design help reduce unnecessary reaching, lifting, or unsafe postures.

5. Methods of task analysis can be used to systematically assess task requirements and develop appropriate solutions when they are excessive. Such solutions include providing power tools to reduce excessive exertion, training, checklists and instructions, modifying the task to make it less demanding, adding additional staff, scheduling rest breaks, and job rotation.

6. Standardization of operations, tools, and parts reduces critical confusions and errors.

7. Methods of inventory control help ensure that necessary parts, tools, and equipment are available at the time they are needed, allowing activities to be performed promptly and correctly.

8. Technical information systems help ensure that accurate updated maintenance and repair instructions, material data sheets, and other important information are made available in a timely manner when needed, which can help prevent critical errors.

9. Quality control and preventative maintenance programs help ensure that equipment and tools remain in a good state of operational readiness.

All of the above approaches address root causes of accidents. In recent years, these approaches have been repackaged in various ways. Examples include concepts such as total quality management (TQM), Six Sigma, 5S (see Box 2), the Visual Factory, Factory Physics, or lean manufacturing. Companies applying these approaches have often observed considerable improvements in both safety and productivity (Lehto and Buck, 2008). It also should be mentioned that these approaches,

---

**Box 2: 5S Programs**

As part of the so-called lean revolution in manufacturing, implementation of a 5S plus safety program (Hirano, 1996) can be viewed as a systematic approach for continuously improving safety. At the most basic level, 5S is a five-step process followed to continuously improve a selected work area, usually involving workers drawn from the particular work area selected for improvement and a 5S expert who facilitates each step of the process. The five steps in the 5S process are as follows:

$S1$ *Sort*. The first step is to take an inventory of all items currently present in the work area. This step can free up space and eliminate clutter.

$S2$ *Set in Order*. The second step is to arrange the remaining items in an orderly fashion and clearly designate a correct location for each item, which makes it obvious when something is out of place or missing. This can lead to some obvious productivity improvements by making it easier for people to quickly find things when they need them and reducing the time needed for people and materials to move around the facility.

$S3$ *Shine*. This third step, often also called "sanitize," is to carefully clean all parts of work areas and equipment and then paint them white so that dirt or grime will stand out. One of the advantages of this approach is that maintenance issues such as minor leaks in hoses and fittings become obvious long before serious problems occur. A second advantage is that the environment becomes much brighter.

$S4$. *Standardize*. The fourth step often overlaps greatly with steps 1 and 2. This follows because the number of tools, dies, fixtures, parts, and types of equipment needed for a particular process can often be reduced by standardizing processes, tools, and the products they provide to customers. This, in turn, helps reduce clutter and can greatly increase productivity.

$S5$. *Sustain*. The last step is to develop ways of sustaining the improvements that have been made. Ideally, the first four steps of 5S are made part of each worker's job so that the improvement process is continued on a permanent basis. Companies also might publicize the 5S program with a newsletter and conduct periodic 5S inspections to demonstrate commitment.

---

involving organizations such as the Purdue Regenstrief Center for Healthcare Engineering (http://www.purdue.edu/discoverypark/rche/), are currently driving major efforts in health care settings to apply methods of industrial engineering to improve patient care and safety through better delivery of critical services.

### 4.3.2 Hierarchy of Hazard Control

The so-called hierarchy of hazard control can be thought of as a simple model that prioritizes control methods

from most to least effective. One version of this model proposes the following sequence: design out the hazard, guard against the hazard, and warn (Laughery and Hammond, 1999). Another version is eliminate the hazard, contain or reduce the hazard, contain or control people, train or educate people, and warn people (Lehto and Clark, 1990). The basic idea is that designers should first consider design solutions that completely eliminate the hazard. If such solutions are technically or economically infeasible, solutions that reduce but do not eliminate the hazard should then be considered. Behavioral controls, such as training, education, warnings, employee selection, and supervision, fall in this latter category for obvious reasons. Simply put, these behaviorally oriented approaches will never completely eliminate human errors and violations.

On the other hand, few product design solutions completely eliminate human errors and violations. Furthermore, imperfect behavioral solutions can supplement and, in some cases, be preferable to product design solutions (Hoyos and Zimolong, 1988). For example, a behavioral solution that reduces the number of automobile collisions, such as enforcement of speed limits, obviously supplements design solutions such as better seat belts and is arguably acting at a more fundamental level by helping prevent the accident as well as reducing its consequences. The obvious conclusion is that designers and others need more guidance than the hierarchy of hazard control provides to select appropriate intervention strategies.

A more substantial issue is that an emphasis on nonbehavioral solutions seems to conflict with the traditional view of accident researchers that human error and intentionally unsafe behavior are the predominant cause of accidents (e.g., Cooper, 1961; Heinrich et al., 1980; Kowalsky et al., 1974; Lehto and Miller, 1986; Ramsden, 1976; Wagenaar, 1992). Part of the blame for human error can be given to poor product and equipment design (Norman, 1992). However, many other factors play a role. For example, one analysis of accidents at sea found that habits, incorrect diagnoses, lack of attention, lack of training, and unsuitable personality contributed to 93 of the 100 accidents studied (Wagenaar and Groeneweg, 1988). The authors of the latter study conclude that preventing human error is the most promising approach to reducing accidents. Some of their proposed solutions include better training, working conditions, behavioral controls, and incentives. This again supports the conclusion that behavioral solutions are an important tool in preventing accidents and deserve more consideration than the hierarchy of hazard control would suggest.

However, it should be emphasized that there are other ways of eliminating or reducing undesired behavior that can be quite effective. Some of these alternative solutions are:

1. Design for usability and understandability
2. Behavioral constraints—elements of product design that make the undesired behavior difficult or impossible

3. User selection—making the product available only to selected, qualified, and responsible users
4. Supervision, enforcement, and incentives

As pointed out by Norman (1988), human errors can often be eliminated by designing products to be more usable. Some of his suggested solutions include the use of affordances and constraints, visible and natural mappings, and the provision of feedback. Such features make correct and incorrect uses of the product more obvious to the user and reduce the need for instruction manuals, warning labels, and other types of product information.

Behavioral constraints, on the other hand, are features of the product that make it hard or impossible to perform certain behaviors (Norman, 1988). Examples include features such as interlocks, lock-ins, lock-outs, guards, or barriers. Other behavioral constraints require that the user have certain knowledge (such as a password) to operate the product and are often targeted to prevent use of a product by unqualified users. For example, consider the case of child-resistant caps. Some related strategies include screening out employees with alcohol or drug dependence or bad driving records or who have not taken training courses. Supervision- and enforcement-related strategies focus on detecting and stopping the behavior, as illustrated by the need for supervisors to detect and prevent willful violations of safety rules.

Behavioral incentives include methods of rewarding safe behavior, such as reduced insurance premiums to nonsmokers or to drivers who use seatbelts. Other uses of incentives include punishment, such as issuing tickets for failing to wear a seatbelt. At this point, it should almost be unnecessary to state that there are many fundamentally different approaches to modifying human behavior that will often be effective. Behavioral controls such as safety information supplement product design by making certain hazards more obvious. They can also supplement supervision, enforcement, and use of behavioral incentives by reminding or informing people such programs are in place. A warning sign that informs drivers that they have entered a radar speed control zone or that speeding penalties are doubled in highway construction work zones illustrates this role. A closely related role of a warning is that of providing feedback that informs the user when they make errors. The importance of the alerting and feedback roles of a warning implies that warning systems that detect and selectively respond to intermittent hazards are especially desirable.

Warnings and other forms of safety information, such as safety precautions, can also serve as performance aids that help people decide what to do (Lehto and Miller, 1986; Lehto, 1992). In the latter role, the information sources often are serving as concise forms of external memory that help people remember and apply what they already know. This can happen in at least three different ways. That is, safety information can identify or describe the hazard, describe actions that should be taken to reduce the hazard or its effects, or direct the person's attention to other sources of information. For

example, a warning label might inform the user that a product contains hydrochloric acid, briefly describe what to do if it enters a person's eye, and direct the user to the material safety data sheets (MSDSs) and other sources of more detailed information. Such information can be a useful supplement to training, instruction manuals, education, and experience. This role is especially likely to be important when people do not know or have forgotten how to perform certain safety-critical tasks.

## 4.4 Risk Management and Systems Safety

In any organization it is essential that risks be kept to an acceptable level. Rather than simply reacting to accidents, management should be proactively taking steps to prevent them from occurring. To do this properly, management must balance the severity and likelihood of the hazards faced against the effectiveness and cost of control measures. The first step in this process is to systematically identify which hazards are potentially present. The next step is to assess the criticality of each hazard on some type of risk index that takes into account both severity and likelihood. Management can then use this information to prioritize hazards and decide upon control measures that keep risks at an acceptable level.

Practitioners in the field of systems safety apply this proactive approach throughout the life cycle of a product, program, or activity (Roland and Moriarty, 1990). This process involves hazard analyses, design reviews, and specification of safety requirements at each stage of development as the design moves forward from an initial concept to production and deployment in the field. One advantage of this approach is that it is usually much easier and cheaper to eliminate hazards by taking steps early in the design process. Control measures added in response to accidents that occur after a design has been launched are also often less effective. The systems safety approach also requires hazard analyses and design reviews for proposed changes. This is important, as changes in systems, processes, or products can often create hazards. Change analysis (Johnson, 1973, 1980) is also a useful technique for identifying root causes of system failures and accidents after they occur.

Applications of the systems model often involve various forms of probabilistic risk assessment to evaluate the reliability of safety-critical subsystems and determine the effect of design changes and other control measures. In this approach, fault trees are used to determine which combinations of component failures and human errors of omission or commission can cause a subsystem to fail and how different subsystem failures can come together to cause unsafe conditions, accidents, and ultimately injuries. The probability of system and subsystem failures as well as accidents can then be calculated after assigning probabilities to component failures and human errors of omission or commission.

One major strength of following the systems safety approach is that it provides a "divide-and-conquer" strategy for systematically identifying management oversights and omissions that cause accidents (Johnson, 1975, 1980). Johnson used the systems approach to organize a large number of factors contributing to accidents in

MORT (the management oversight and risk tree). In its original form, MORT was a method for accident investigation (Johnson, 1975) but eventually evolved into a model of safety management that follows a systems approach to organize and address a large number of factors contributing to management oversights and omissions, including (1) failures to adequately assess risk and (2) less than adequate safety policies, (3) supervision, (4) engineering controls, and (5) standard operating, (6) inspection, and (7) maintenance procedures.

## 5 OCCUPATIONAL HEALTH AND SAFETY MANAGEMENT PROGRAMS

As stated in the Occupational Health and Safety Act, employers are required to provide a workplace free from recognized hazards. This is normally done by establishing a health and safety management program. Guidance on how to do this is available from a large set of sources. For example, the American National Standards Institute (ANSI) has developed a standard for occupational health and safety management, *ANSI/AIHA Z10-2005: American National Standard for Occupational Health and Safety Management Systems*. The standard contains seven sections: *Management Leadership* and *Employee Participation*, *Planning*, *Implementation and Operation*, *Evaluation and Corrective Action*, and *Management Review*. The standard outlines what has to be accomplished by the organization but is a "performance" standard and leaves the specific tasks to be determined by each organization for their unique circumstances.

Some of the features of successful safety management programs that have been found in studies of program effectiveness are summarized in Table 7.

Typical activities conducted in a health and safety management program include (1) ensuring compliance with safety standards and codes, (2) establishing a general housekeeping program, (3) accident and illness monitoring, (4) identifying and analyzing workplace hazards, and (5) implementing controls to reduce or eliminate hazards.

### 5.1 Compliance with Standards and Codes

An essential aspect of a health and safety management program is to determine which standards, codes, and regulations are relevant and then taking steps to ensure compliance. Many standards and codes are developed by consensual organizations and cover such topics as chemical labeling, personal protective equipment, and workplace warning signs. In addition, states develop regulations and standards which pertain to workplace safety and health.

The broadest applicable regulation in the United States is the OSHA General Requirements for all machines, [29 Code of Federal Regulations (CFR) 1910.212] In addition to the OSHA which place an obligation on employers to provide a workplace free from recognized hazards, OSHA also specifies a large set of detailed and mandatory health and safety design standards. OSHA also requires that many employers

**Table 7  Factors of Good Safety Management (Zimolong, 1997)**

| General Factor | Safety Practice Definitions |
|---|---|
| Management commitment | 1.  Safety officer holds high staff rank.<br>2.  Top officials are personally involved in safety activities; e.g., they make personal plant safety tours and give personal attention to accidental injury reports.<br>3.  High priority is given to safety in company meetings and in decisions on work operations.<br>4.  Management sets clear safety policy and goals. |
| Safety committee and safety rules | 1.  The safety committee holds regular, frequent meetings.<br>2.  Safety rules are regularly reviewed and updated in light of accident experience.<br>3.  There is evidence of management and staff compliance with rules. |
| Hazard control | 1.  There is a high level of housekeeping.<br>2.  There is orderly design / layout of work processes.<br>3.  There are good environmental qualities (ventilation, lighting, noise control).<br>4.  There is a greater number and variety of safety devices of on operating machinery. |
| Inspections and communications | 1.  There are daily worker-supervisor contacts on safety or other job matters.<br>2.  Formal inspections are made at regular, frequent intervals.<br>3.  There is a smaller span of supervisor control.<br>4.  There are numerous informal contacts between workers and top officials. |
| Accident investigations | 1.  Investigations and records are kept both on disabling (lost time) and record keeping injuries and nondisabling injuries.<br>2.  Investigations are made of property accidents and ''near misses.''<br>3.  There is regular use of reports for prompting hazard control measures. |
| Employee support | 1.  There are well-established procedures for job placement and advancement.<br>2.  There are personal counseling services.<br>3.  There are recreational facilities and programs for off-job hours. |
| Safety motivation | 1.  A humanistic approach is used in disciplining safety violators.<br>2.  Worker families are enlisted in safety promotions.<br>3.  Specially designed posters or displays are used for hazard recognition.<br>4.  Individual praise, recognition are given for safe job performance. |
| Safety training | 1.  Safety is included in new worker orientation.<br>2.  Workers are given initial and follow-up training in safe job procedures.<br>3.  Supervisors are given special safety training.<br>4.  A variety of safety training techniques (lectures, films, group discussions, simulations) are used. |
| Makeup of workforce | 1.  Workers are generally older.<br>2.  Workers generally have longer experience in their jobs.<br>3.  There are more married workers.<br>4.  There is less turnover and absenteeism in the workforce. |

implement comprehensive hazard communication programs for workers involving labeling, MSDSs, and employee training (29 CFR 1910.1200). OSHA enforces these standards by conducting inspections and imposing fines if the standards are not met. In addition, employers are required to maintain records of work-related injuries and illnesses and to prominently post both an annual summary of injury and illness and notices of noncompliance with standards.

In addition to federal OSHA regulations, 22 states have developed and operate state-run OSHA programs. Five additional states have state-run programs for public sector (state and local government) employment only. States must set job safety and health standards that are "at least as effective as" comparable federal standards. (Most states adopt standards identical to federal ones.) Additionally, states have the option to promulgate standards covering hazards not addressed by federal standards.

## 5.2  General Housekeeping and Preventative Maintenance

Good housekeeping and preventive maintenance conducted to prevent the development of unsafe work conditions is important in almost any imaginable work facility, especially important when toxic or hazardous materials are present or used in the production process. Some general requirements and elements of an adequate housekeeping and maintenance program are as follows:

1. Cleaning and maintenance should be scheduled on a frequent periodic basis to ensure dirt and clutter do not build up over time to unacceptable levels, minimize leaks from machinery, storage drums, and other sources, and ensure that air filters and ventilation systems are working properly.

2. Spilled liquids, dusts, and other objects should be immediately cleaned up using appropriate methods which do not add to the problem. In particular, toxic materials, acids, and otherwise reactive or hazardous materials should normally be neutralized or diluted before attempting to remove them.

3. Washrooms and showers should be provided to workers in dirty jobs.

4. Work and traffic areas should be clearly marked to separate them from temporary storage areas for work in progress (WIP).

5. Convenient, easily accessible locations should be designated for storing tools, parts, and other essential items used in the workplace.

6. Waste containers or other disposal devices should similarly be provided in convenient locations.

Despite the obvious benefits of following good housekeeping practices, any ergonomist with significant industrial experience will agree that many, if not most, organizations have difficulty maintaining a clean, uncluttered work environment for their employees (see Box 2). This tendency is especially true for small manufacturing faculties, but even larger organizations devoting significant efforts to housekeeping often have significant room for improvement. Part of the issue is that housekeeping is often viewed as a janitorial task, separate from the day-to-day responsibilities of most employees. Another issue is that clutter has a tendency to build up over long periods. In our experience, it is not unusual to find tools, equipment, and parts that have been sitting around unused for years, sometimes even taking up valuable space on the shop floor!

### 5.3 Accident and Illness Monitoring

The recording and reporting of work-related injuries and illnesses is another critical component of any safety management program. In most cases, this process involves some form of accident investigation. Another requirement is to post and maintain statistical records. These activities are mandated by the Occupational Health and Safety Act and help the employer evaluate the extent and severity of work-related incidents and identify patterns of occurrence that should be the focus of safety management efforts.

### 5.3.1 Accident Reporting

OSHA requires employers to submit a report of each individual incident of work-related injury or illness that meets certain criteria as well as maintain a log of all such incidents at each establishment or work site. Reportable incidents are those work-related injuries and illnesses that result in death, loss of consciousness, restricted work activity or job transfer, days away from work, or medical treatment beyond first aid. The specific information that must recorded is summarized below (OSHA Recordkeeping Handbook, http://www.osha.gov/recordkeeping/handbook/index.html).

OSHA form 301, "Injury and Illness Incident Report," is used to report a recordable workplace injury or illness to OSHA. Information must be provided on this form with regard to:

- Employee identification
- Medical treatment received
- Time of event and how long the employee had been at work
- What the employee was doing just before the incident occurred (s.a., "climbing a ladder while carrying roofing materials")
- What happened (s.a., "When ladder slipped on wet floor, worker fell 20 feet")
- What the specific injury or illness was (s.a., "strained back" or "broken wrist")
- What object/substance directly harmed the employee (s.a., "concrete floor") Whether the employee died

In addition, the same injury or illness must be recorded on OSHA form 300, "Log of Work-Related Injuries and Illnesses," which must be maintained at each work site. Each entry in the log identifies the employee and their job title, gives the date of the injury or onset of illness, describes where the event occurred, and gives a short description of the injury or illness, the parts of the body affected, and the object or substance that injured or made the person ill. The case is then classified into one of the following four categories based on outcome:

1. Death
2. Days away from work
3. Job transfer or restriction
4. Other recordable case

If applicable, the number of days away from work must be recorded as well as the number of days on job transfer or restriction. Finally, the case is classified as either an injury or one of five major types of illnesses.

Data from the form 300 log must be summarized at the end of each year on OSHA form 300A, "Summary of Work-Related Injuries and Illnesses." OSHA mandates that these summary statistics be posted at the work site, accessible for viewing by all employees, from February 1 to April 30 of the following year.

OSHA periodically visits plants to determine if they are in compliance with the law. When OSHA inspectors call, they have the right to see all of the reports/logs described above. Also, certain events will automatically trigger an OSHA visit, such as an accident resulting in a fatality.[*]

### 5.3.2 Calculation of Incidence Rates

Data submitted to OSHA can be used to calculate incidence rates for various types of injuries or illnesses.

---

[*] There is always an OSHA visit when a fatality occurs or more than six persons are injured in a single accident. At other times, OSHA targets certain industries because of an industry-wide problem and those cases can trigger visits.

An important component of an effective safety management program is to calculate and track these rates. *Incidence rates* should be tracked over time, and also compared to industry rates as a whole, in order to identify new problems in the workplace and/or progress made in preventing work-related injuries and illnesses.

An *incidence rate* is defined as the number of recordable injuries and illness occurring among a given standard number of full-time workers (usually 100 full-time workers) over a given period of time (usually one year). For example, the incidence rate of total nonfatal injury and illness cases in meatpacking plants in 2002 was 14.9 per 100 workers. This means that an average of 14.9 injuries and illnesses would occur for every 100 workers over the course of a year in a meatpacking plant.

The formula for calculating the incidence rate of injuries and illnesses per 100 workers is

$$\text{Incidence rate} = N \times (200,000/EH) \qquad (1)$$

where:

$N =$ number of reportable injuries and illnesses at an establishment during a one-year period

$EH =$ annual total number of hours worked by all employees at the establishment in a 50-week year

$200,000 =$ annual total number of hours worked by 100 workers (100 employees $\times$ 40 h/week $\times$ 50 weeks/yr)

The following example illustrates how to use this formula. Note that it is necessary in this example to annualize the number of accidents and number of hours worked by all employees at the establishment.

*Example:* If a plant experiences four accidents during a 26-week interval for their 80 employees who worked 40 h weekly, then the incidence rate is calculated as

$N = 2 \times 4$ accidents in a 26-week period

$= 8$ reportable accidents in an annual period

$EH = 80$ employees $\times$ 40 h/week $\times$ 50 weeks

$= 160,000$ total employee hours worked in 50 weeks

Incidence rate per 100 $= N \times (200,000/EH)$
employees per year

$= 8 \times (200,000/160,000)$

$= 10$

Another important statistic is the *lost-workday rate*. The lost-workday rate is defined as the number of lost workdays resulting from all recordable injuries and illness occurring among a given number of full-time workers over a given period of time. This statistic gives a measure of the severity of the injury and illness

experience of a given establishment. For example, two different establishments may have the same incidence rate of injury and illness. However, one establishment could have a higher lost-workday rate than the other for the same number of injuries/illnesses. In this way, the lost-workday rate tells us that the former establishment has a higher severity of injury/illness.

The formula for calculating the lost-workday rate per 100 workers is

$$\text{Lost-workday rate} = M \times (200,000/EH) \qquad (2)$$

where:

$M =$ number of lost workdays of employees due to reportable injuries and illnesses during a one-year period

$= N \times$ (avg. no. workdays lost per injury/illness)

*Example:* In the previous example, if the four cases average five days lost time each, then, again per 100 workers,

$\text{Lost-workday rate} = M \times (200,000/EH)$

$M = N \times (5 \text{ avg. workdays lost per injury/illness})$

$= 8 \times 5$

$= 40$ lost workdays due to reportable injuries and illnesses during a one-year period

It can be very beneficial to compare incidence and lost-workday rates for a particular company to industry averages. Note, however, that some caution must be observed when making these comparisons. While industry-wide rates are usually reported per 100 employees, in a few cases the rates are reported *per 10,000 employees*. Rates are typically reported per 10,000 employees only for incidents of low occurrence, such as certain types of illnesses. *When comparing company incidence rates to industry-wide statistics*, it is necessary to be sure whether the industry rates are reported per 100 employees or per 10,000 employees.

## 5.4 Hazard and Task Analysis

The process of identifying and classifying hazards has been recognized for many years as an important first step in injury and accident prevention. For example, over 50 years ago, Heinrich identified a generalized procedure for improved safety, which he called, "the hazard through-track." This through-track starts with a *hazard recognition phase* (Heinrich 1959). This phase includes developing knowledge about probable hazards and their relative importance. The second phase is to find and *name hazards explicitly*. Identifying the particular hazards involves analyses, surveys, inspections, and inquiries. Supervisory and investigative reports usually

form the basis of this identification. *Selection of a remedy* forms the third phase of this through-track. Some of the options here include engineering revisions of product or process, personnel adjustments or reassignments, and training, persuasion, and/or discipline. The last phase of this through-track consists of *implementing the remedy*. An important step here is to verify that the remedy is an improvement. Sometimes a postaudit is needed to verify that a remedy works.

### 5.4.1 Critical-Incident Analysis

One of the challenges of hazard and task analysis is the combinational explosion of conditions and events that occurs in a reasonably complex work environment. Whereas accident investigation can focus on "what happened" using deductive reasoning and physical evidence, hazard analysis must usually rely on inductive reasoning to determine "what can happen." The classic problem in accident analysis is that the analyst must consider how a possibly very large number of subsets of conditions and events might interact to produce unusual and undesirable results. Consequently, there often are not enough accident data to determine how likely certain types of accidents are.

The critical-incident technique (CIT) is a type of analysis that is helpful in bridging the gap between accident investigation and pure prospective hazard analysis (Flanagan, 1954). The CIT is used to gather useful information from "near misses," or accidents that almost happened. Since these events may have only lacked a single factor to result in an accident, their analysis can be as useful as investigating an actual accident.

### 5.4.2 Work Safety Analysis

A number of hazard identification techniques have emerged which focus on dividing work into sequences of subtasks which are then individually evaluated. Work safety analysis (WSA) is typical of such approaches and involves the development of a matrix in which hazards are identified for each subtask and then described in terms of causative factors and corrective actions (Suokas and Rouhiainen, 1984). To guide this process, checklists of generic accidents and their causes are provided to the analyst. The analyst also rates the probability and consequence of each hazard on five-point scales. The hazards are then ranked on a risk index obtained by multiplying the probability score by the consequence score, both before and after the corrective action.

Action error analysis (AEA) is a similar approach which also involves the division of work procedures into sequential stages (Suokas and Pyy, 1988). A matrix is then developed in which potential errors are identified for each stage and described in terms of primary consequences, secondary consequences, means of detection, and measures for prevention. As such, performing AEA is practically identical to developing the detectability/revocability/consequence (D/R/C) matrix (Altman 1964; 1967) as a way of prioritizing, in terms of importance, the human errors found during task analysis.

### 5.4.3 Human Reliability Analysis (HRA)

Several quantitative techniques have also been developed for analyzing human error. All such approaches involve task analysis. They diverge significantly, however, in the degree to which they attempt to develop quantitative measures of error. The technique for human error rate prediction (THERP) (Swain and Guttman, 1983) provides tabled estimates of human error probabilities for a limited set of tasks primarily relevant to those performed in the nuclear power industry. Application of THERP begins with human-error-related events in the system fault tree. The associated task is then divided into a sequence of elemental subtasks which are organized into an event tree. Human error probabilities are then assigned to the outcomes of each task from a tabled set of values. (The tabled values include provisions for the influence of performance-shaping factors and dependencies between subtask failures.) The probability of task failure is then calculated from the event tree. While advantageous in that it provides a quantitative measure of human error, the overall applicability of THERP to industrial tasks is severely limited by the limited set of tasks addressed.

The subjective likelihood index methodology (SLIM-MAUD) addresses this latter issue by focusing on expert judgment as a means of estimating the probability of an accident or human error (Embrey et al., 1984). The approach is based on decision-analytic methods for obtaining subjective probability estimates. It takes the format of asking experts to estimate (1) the importance of performance-shaping factors as causes of accidents or human error and (2) the degree to which each performance-shaping factor is present in the evaluated scenarios. This information is then used to develop a mathematical ordering of the likelihood of the analyzed errors.

### 5.4.4 Error Modes and Effects Analysis (EMEA)

Error modes and effects analysis (Lehto, 2006) is a systematic method for determining preferred intervention strategies for errors that might occur. The first step in EMEA is to identify and prioritize the hazards, or effects, associated with inappropriate behavior falling within particular stages of user or purchaser interaction with the product. The hazard identification stage in EMEA involves a process similar to that used in WSA of identifying inappropriate or missing responses, their frequency of occurrence, and the severity of associated hazards at the task or subtask level for different stages of interaction with a product (Lehto, 1996). Stages of interaction that might be evaluated include purchasing decisions, set-up or assembly tasks, ordinary use, troubleshooting, maintenance and repair, emergency procedures, and disposal of the product.

The second step in EMEA involves a systematic analysis of the behavioral basis of errors or violations that are occurring frequently or have severe consequences. This process involves determining potential skill, rule, knowledge, or judgment-based error modes for each of the analyzed errors or violations. For example, many purchasing decisions and ordinary

uses of a product are performed routinely at the skill- or rule-based levels. Other tasks such as assembly, troubleshooting, or maintenance are performed less often, perhaps at the knowledge-based level. The important point is that the behavior basis, or error mode, of the undesired behavior is likely to differ between users, tasks, and products. Once error modes have been determined for the analyzed errors or violations, the next step in EMEA is to consider appropriate intervention strategies, including product design features, training, job design, supervision, written procedures, checklists, or warnings.

## 6  COMMON HAZARDS AND CONTROL MEASURES

Hazards present in the work environment can have a significant effect on productivity, safety and health, worker satisfaction, and employee turnover. Dirty, cluttered, poorly organized work, traffic, and storage areas are one common problem. Other potential concerns include exposure to hazardous materials, temperature extremes, inadequate lighting, or noise levels. Addressing these issues requires knowledge of how environmental conditions impact people, assessment methods, and a toolbox of solutions. Engineering solutions which involve altering the environment are the most fundamental approach but are often expensive. Less expensive solutions include administrative controls, such as job rotation, rest breaks, and employee selection, as well as implementing better methods of housekeeping. Providing protective equipment and clothing is another potential solution in some situations (see Box 3). The following discussion will introduce some of these hazards along with methods of control.

### 6.1  Cleanliness, Clutter, and Disorder

Dirty, cluttered, or poorly organized work environments are a common unsafe condition that can lead to health problems and accidents, reduce employee morale and productivity, and reduce the quality of the products and services produced by a manufacturer or service

---

**Box 3: Personal Protective Equipment**

Personal protective equipment (PPE) is an especially important solution for common hazards present throughout a work environment, for example, PPE such as protective eyewear, protective footwear, gloves, hearing protection, uniforms or similar clothing, and hardhats. PPE can also be targeted to protect workers for specific tasks and hazards such as electrical lineman gloves, thermal protective clothing, and welding helmets. OSHA regulates PPE based on the industry, such as general industry, construction, longshoring, shipyard, and marine terminals. Types of personal equipment required by OSHA are given below (http://www.osha.gov/SLTC/personalprotectiveequipment/standards.html):

| PPE | Examples | Reference |
|---|---|---|
| Foot | Shoes/boots: steel toes/insoles; reinforced; rubber/plastic; thermal insulation; nonconductive; slip resistant; metal free; wooden soled; metatarsal instep guards; conductive; nonsparking; gaiter | 1910.136: Occupational foot protection |
| Leg | Leggings; knee pads; shin guards | |
| Body | Aprons; garments; full suits; cooling; puncture/cut-resistant clothing; high-visibility clothing; coats and smocks; coveralls; fire entry/proximity suits; rainwear; personal flotation device (PFD) | |
| Skin | Protective creams; cleaners | |
| Hand–finger–arm | Gloves/mittens; pads; finger guards and cots; wristlets; arm protectors; protective sleeves | 1910.138: Hand protection |
| Face | Shields; babbitting helmets; welding helmets (UV); acid-proof hoods; air-supplied hoods | 1910.133: Eye and face protection |
| Eyes | Safety glasses (frontal/side impact); protective goggles | 1910.133: Eye and face protection |
| Ears | Plugs; muffs | 1910.95: Occupational noise exposure |
| Head | Safety helmets or hard hats; bump caps; soft caps; hair nets and caps | 1910.135: Head protection |
| Fall | Safety belt; safety harness; lanyard; grabbing device; lifeline; fall arrestor; climbing safety systems | 1926 subpart M (construction) |
| Respiratory PPE | Self-contained (SCBA); supplied; purifying; filter | 1910.134: Respiratory protection |

provider. Some of the many ways this can happen are listed below:

1. People can slip on spilled liquids or powders or trip over small objects or clutter on the floor, resulting in serious injuries.

2. People's ability to move about their environment might be impeded by stacks of work in process (WIP) or other objects unnecessarily cluttering work or storage areas, aisles, and passageways.

3. Tools, parts, or other objects stacked on shelves or cluttering a work surface may use up much of the available space, and significantly interfere with people's ability to do their tasks as intended.

4. Dirt and grime accumulated on light fixtures, windows, walls, ceilings, and elsewhere in the facility can greatly reduce the brightness of the work environment, interfering with people's ability to perform essential visual tasks and in general creating unpleasant effect.

5. Toxic, irritating, allergenic, carcinogenic, tetragenic (potentially causing birth defects), or otherwise harmful substances may cause health problems when they contact the skin or are ingested when workers smoke or eat without first washing their hands.

6. Dusts, vapors, and gases may enter the air and be inhaled, resulting in serious health problems. They also might accumulate in significant quantities, creating fire and explosion hazards, or contaminate products that are being produced.

7. Poor sanitation might lead to the spread of disease within facilities and is a special concern in health care settings and food industries.

Other examples along these lines can be easily imagined. One traditional solution is better housekeeping and maintenance, as discussed earlier in this chapter.

## 6.2 Fall and Impact Hazards

Many accidents involve falling or being struck by a moving or falling object. Falls and impacts account for about 51% of the accidents reported by the Bureau of Labor Statistics (BLS) in 2008 for which there was a lost workday. These types of accidents are particularly common in the construction industry and are often easily preventable.

Falls often result from slipping or tripping. Wet floors are an invitation for slips and falls. Good housekeeping minimizes slip hazards, and marking wet floors alerts people to use caution. Keeping floor surfaces in good repair and clear of objects such as equipment and cords will also reduce trip hazards.

Falls from elevated work surfaces, such as stairs, ladders, and scaffolds, are also a frequent type of accident in this category. Safety harnesses and guardrails are important safety equipment for minimizing falls from elevated work surfaces. In addition, it is important that ladders, stairs, and guardrails meet OSHA and

---

**Box 4: Traffic Areas**

Standards such as military standard (MIL STD) 1452, OSHA (see 29 CFR 1910), or the Federal Aviation Administration (FAA) (HF-STD-001) contain a variety of general requirements for stairs, aisles, ramps, floors, and other traffic areas (also see Van Cott and Kincaide, 1972).

a. Aisles and work areas should not occupy the same floor space to help prevent interference and collisions

b. Providing necessary pull-outs or turning space in aisles for passage of wheelchairs or material-handling devices or equipment

c. Ensuring that adequate space is marked out and allocated for placing materials in storage or marshalling areas to eliminate interference with work or passage

d. Appropriate markings of aisles and passageways

e. Adequate clearance dimensions for aisles and passageways

f. Appropriate dimensions of stairs

g. Appropriate flooring materials free of protruding objects that might create tripping hazards

h. Eliminating obstacles that might be collided with

i. Avoiding blind corners

j. Making sure there is a connected, accessible path by which disabled users can reach most of the facility

k. Ensuring that doors do not open into corridors

l. Avoiding one-way traffic flow in aisles

m. Special requirements for emergency doors and corridors

Not all of these requirements are directly related to safety, but failure to comply with them often is an important contributing factor to accidents, such as falls and collisions involving pedestrians, forklifts, open doors, or other people.

---

architectural design standards (Box 4) (29 CFR 1910 Subpart D).

Elevated work surfaces also present dangers from falling objects on those working below. While work situations in which workers are working physically above others are unavoidable in many construction and maintenance operations, extreme care should be taken to prevent hand tools and other objects from falling on those working below. Sometimes fine mesh fencing can be placed horizontally to catch falling objects. Regardless, hard hats should be worn for protection at all times.

Less obvious, but important to preventing falls, is the screening and proper selection of employees for working in elevated workspaces. There are psychological and physiological differences among people with regard to their ability to tolerate heights. Some people are afraid of

being exposed to heights in which there are few barriers to falling or jumping. Such workers should be screened from jobs where work on elevated surfaces is required. New employees should always be accompanied by an experienced employee who can observe any adverse reactions to heights, or have a fear of falling, that might impair their safety. Also, people with colds or flu or who are recovering from those diseases often have an impaired sense of balance and therefore should not be working in situations where momentary imbalance can affect their safety. Some medications for colds are suspected of impairing the sense of balance.

### 6.3  Hazards of Mechanical Injury

Machinery or tools pose a potential for cutting, shearing, crushing, and pinching injuries when people contact sharp or moving parts.

### 6.3.1  Cutting/Shearing

Sharp cutting tools and machines are common hazards in many factories. Knives in the meatpacking industry are examples that come to mind easily, but there are numerous sharp edges in companies that make metal, glass, lumber, or ceramic products. Almost all manufacturing firms have sharp machine parts.

Powered cutting tools are a common cause of cutting and tearing accidents. Most of these tools have guards that prevent the human body from slipping into the rotating or reciprocating blades. While these guards are not infallible, removal of the guards should only be permitted in extreme situations where the job is impossible to perform with the guard in place. It is critical then that those guards be replaced immediately. The same is true for guards for shearing machines, which have been a source of many finger–hand amputations.

### 6.3.2  Crushing/Pinching

*Pinch points* are locations other than the point of operations where it is possible for the human body to get caught between moving parts of machinery. These points are particular hazards to crushing of body parts. Other causes of crushing include hitting one's finger with a hammer or getting one's hands between two heavy moving objects such as objects suspended from crane cables. Besides crushing, many of these same accidents can lead to broken bones.

The use of machine *guards* and *safety devices* can provide protection from mechanical injury accidents and is emphasized in OSHA standards for several industries (Table 8). Guards are intended to prevent any part of the human body from entering a hazardous area of the machine, while safety devices deactivate the machine when body parts are in hazardous areas. Sometimes safety devices are hooked to machine guards to prevent the machine from operating without the guards in place.

Preferred guards and safety devices work automatically, impose little or no restrictions on the operator, and are productive. Good guards should also be fail-safe and prevent the operator from bypassing or

**Table 8  OSHA Regulations for Specific Industries Addressing Mechanical Hazards**

| OSHA Regulation | Title |
| --- | --- |
| 1910.213 | Woodworking machinery requirements |
| 1910.217 | Mechanical power presses |
| 1910.219 | Mechanical power transmission apparatus |
| 1910.241-244 | Subpart P — Hand and Portable Powered Tools and Other Hand-Held Equipment |

deactivating them.[*] Some such guards totally enclose hazardous operations with limited, adjustable, or no access. Limited-access guards allow small product parts to be inserted by the operator but prevent entry of body parts. Adjustable-access guards permit changes to the openings around the hazard area in order to adapt to different sizes of product parts. Unfortunately, there is a tendency for operators to leave the guard at the maximum size opening at all times and hence deactivate much of the guard protection.

A number of other safety devices have been devised for worker protection. Optical screens, ultrasonic devices, and field sensing devices are common safety devices that are used around presses and press-brakes. Optical screens deactivate the machine whenever a light shield around the machine openings is broken. Ultrasonic devices and field sensing devices detect any intrusion of the hands into a hazard area and terminate machine operation. Screens and sensing devices may need to be supplemented with special stopping mechanisms in certain types of machines that complete a cycle even after the power is cut off.[†]

Other safety devices include two-hand control devices and pull-out devices. Two-hand control devices are really interlocks that require two control devices to be pressed[‡] before the machine activates. Pullout devices often have wrist cuffs that are connected to small cables that pull the operator's hands clear of the danger area when the machine activates. However, pullout devices often restrict movement and productivity.

Another type of protective approach is to use mechanical means to feed product units into dangerous

---

[*] There have been numerous instances of operators deactivating guards and other protective devices. Those deactivations often occur when the operators are working on incentive systems and they see a way to make more money by keeping the safety device from slowing them down. Incentive systems should not be employed when there are dangers of these kinds in the shop. Disciplinary action is also needed when a safety device is knowingly deactivated and not reported.

[†] Presses and press-brakes are examples of machine tools which frequently complete a cycle even after power is shut off. In such cases, special stopping mechanisms may be required.

[‡] The two devices are presumably depressed with separate hands. However, when the two-hand switch devices are too close, an operator can elbow one switch and depress the second with the hand on the same arm.

machines, thus creating protective distance between the person and the hazard. In addition to guards and safety devices, there is need for supervision to thoroughly train operators and remind them not to wear rings or loose clothing that can catch in the machine and cause accidents.[*] All of these efforts have resulted in fewer industrial accidents with mechanical injuries, but the rate is still unacceptably high.

Woodworking industries present some particular challenges in mechanical injury. One of the reasons is that wood is a highly variable material, and another is in the nature of power woodworking machines. In sawing operations, for example, a knot in the wood can be caught in a saw blade in a way that forces the piece of wood back out toward the operator. To prevent some of these situations, newer models of woodworking power tools have anti-kickback devices. Basically, those devices force the wooden workpiece to bind if the piece suddenly starts to move in the reverse direction. Other anti-kickback devices consist of camlike metal pieces next to the wooden workpiece that can move in one direction but will grip and dig into the workpiece when the direction of movement is reversed.

Mechanical injuries are a primary concern to ergonomic specialists involved in equipment design. A guiding principle is to design safety features that are simple and that minimally interfere with efficient operation of the equipment. Another principle is that danger points should be *labeled with an appropriate warning*.

### 6.4 Ergonomic Issues

Many injuries in the workplace are due to excessive lifting and exertion, exposure to vibration hazards, and repetitive motion. Each of these sources of injury is focused upon heavily by practitioners of ergonomics. The application of ergonomics to address these sources of injury will not be discussed here in detail because it is focused upon at length elsewhere in this book.

#### 6.4.1 Lifting and Overexertion

Workplace disorders and injuries due to lifting and overexertion accounted for 24% of the direct costs of worker compensation in 2007, making them the largest single contributor (Table 5). The overall cost estimate was $12.7 billion, demonstrating the major significance of such injuries. A number of generic control strategies have been developed for reducing the risk associated with lifting and force exertion. In particular, the NIOSH (http://www.osha.gov/dts/osta/otm/otm_vii/otm_vii_1 .html#app_vii:1_2) work practices guidelines specify a way of determining recommended weight limits for different lifting conditions. A number of computer modeling tools have also been developed by Don Chaffin and his colleagues at the University of Michigan Center for Ergonomics analyzing lifting, carrying, and

other forms of physical exertion. These programs are commercially available (i.e., information on the JACK program can be easily obtained from the Web).

#### 6.4.2 Vibration Hazards

Most power tools are vibratory in nature and those vibrations are imparted to people handling them. Chain saws, chipping hammers, jack hammers, and lawnmowers are a few that come to mind easily. These power tools tend to vibrate the operator's *hands and arms*. Also, some vehicles exhibit heavy vibration and people inside are subjected to *whole-body* vibration.

It is known that whole-body vibration can increase heart rate, oxygen uptake, and respiratory rate. In addition, whole-body vibration can produce fatigue, insomnia, headache, and "shakiness" during or shortly after exposure. However, it is unclear what the long-term effects are from exposure to whole-body vibration.

In the case of hand and arm vibrations, there tends to be a vasospastic syndrome known as *Raynaud's syndrome* or *dead fingers* (NIOSH, 1983). This circulatory disorder is usually permanent. Typically it takes several months of exposure to around 40–125 Hz vibration to occur, but there appears to be large individual differences among people relative to the onset. The two primary means of preventing or reducing the onset frequency of Raynaud's syndrome are (i) reducing the transfer of vibration from hand tools and (ii) protecting hands from extreme temperatures and direct air blast. Three ways to reduce vibration transfer from hand tools are (a) to make tool–hand contacts large and nonlocalized, (b) to dampen vibration intensities at the handles with rubber or other vibratory-dampening materials, and (c) to require operators to wear gloves, particularly those with vibration-arresting pads.

#### 6.4.3 Cumulative Trauma Disorders

Cumulative trauma disorders (CTDs) are frequently associated with certain occupational tasks and risk factors. Table 9 summarizes some typical occupational tasks and risk factors along with frequently associated CTDs. While most CTDs do not lead to life-threatening situations, they do lead to missed workdays and considerable inconvenience (NIOSH 1995).

OSHA provides ergonomic guidelines for certain industries to assist in the reduction of musculoskeletal disorders (MSDs). OSHA's resources related to ergonomic guidelines can be found at the following website: http://www.osha.gov/SLTC/ergonomics/resources .html. One common approach followed to address these issues is to perform a task analysis using checklists (see Box 5).

### 6.5 Noise Hazards

Excessive noise is a potential safety problem in vehicles and many work environments. Noise becomes a safety problem at certain levels and frequencies. The most commonly used scale is the adjusted decibel (dBA) scale. Even at low levels, noise can annoy or distract workers. At moderate levels, of 80–90 dBA, noise interferes with communication and causes hearing loss

---

[*] A number of companies appoint forklift operators or other roving operators to the additional duty of safety enforcement. As they go around the plant, it is their duty to remind anyone who is not following good safety practices to change behavior. Many people simply forget to wear safety glasses or hearing protectors, and simple reminders are good practice.

**Table 9  Typical CTDs, Associated Tasks, and Occupational Factors**

| Task Type | Occupational Safety Factor | Disorder |
|---|---|---|
| Assembling parts | Prolonged restricted posture<br>Forceful ulnar deviations<br>Thumb pressure<br>Repetitive wrist motion<br>Forearm rotation | Tension neck<br>Thoracic outlet syndrome<br>Wrist tendinitis<br>Epicondylitis |
| Manual materials handling | Heavy loads on shoulders | Thoracic outlet syndrome<br>Shoulder tendinitis |
| Packing boxes | Prolonged load on shoulders<br>Forceful ulnar deviation<br>Repetitive wrist motion | Tension neck<br>Carpal tunnel syndrome<br>DeQuervain's syndrome |
| Typing, cashiering | Static or restricted posture<br>Arms abducted or flexed<br>High-speed finger movement<br>Ulnar deviation | Tension neck<br>Thoracic outlet syndrome<br>Carpal tunnel syndrome |

*Source*: Adapted from Putz and Anderson (1988).

in susceptible people. Higher levels of noise (above 85 dBA) significantly increase the chance of hearing loss. The effects of long-term exposure to noise are cumulative and nonreversible.

According to the BLS, occupational hearing loss is the most commonly recorded occupational illness in manufacturing, accounting for one in nine recordable illnesses (NIOSH, 2010).

In the United States, the Occupational Safety and Health Act specifies allowable levels and duration of exposure to noise (Table 10). The act also requires that employers monitor noise levels, perform audiometric tests, furnish hearing protection, and maintain exposure

**Table 10  Permissible Noise Exposure**

| Duration Per Day (h) | Slow-Response Level, Adjusted Decibels (dBA) |
|---|---|
| 8 | 90 |
| 6 | 92 |
| 4 | 95 |
| 3 | 97 |
| 2 | 100 |
| 1 $\frac{1}{2}$ | 102 |
| 1 | 105 |
| $\frac{1}{2}$ | 110 |
| $\frac{1}{4}$ or less | 115 |

*Source*: OSHA 29 CFR 1910.95.

records whenever employee noise exposure equals or exceeds an 8-h time-weighted average level of 85 dBA (29 CFR 1910.95).

## 6.6  Pressure Hazards

Pressure hazards can be found in many industrial environments. There are a number of sources of highly pressurized gas around most industrial plants. Boilers are one common source. When the expansive force of an enclosed fluid exceeds the pressure vessel's strength, ruptures occur, often with explosive results and sometimes with great heat. In steam boilers, loss of water will create superheated steam with very high resulting pressures. Safety valves on most steam boilers are designed to trip at a safe upper pressure limit. While that practice reduces pressure hazards, that out-flowing steam is also a safety hazard for burns.

Unfired pressure vessels, such as portable gas cylinders, air tanks, or aerosol cans, can also explode when excessive pressure builds. A major cause of such accidents occurs when these vessels are exposed to sunlight or other heating sources. Special care must be exercised to keep these pressure vessels in cool environments. Note also that portable gas cylinders often contain liquefied gases, which themselves become dangerous if released. These cylinders should be prominently marked as to their contents. Also, these cylinders should always be secured when in use, during transport, and in storage to prevent falling and resulting rupture. When a fall causes a valve to break at the end of a cylinder, the cylinder can act like a flying missile.

Another pressure source that is common in industry is compressed air. Many industrial power hand tools, such as impact wrenches, use air pressure. Typically, a hose supplies compressed air to the hand tool. A typical source of accidents occurs when the hose accidentally uncouples from the tool and whips with great force. Long air hoses should be restrained in case of accidental uncoupling.

## 6.7  Electrical Hazards

*Shocks* and *burns* are by far the most common electrical injuries. Burns cause destruction of tissue, nerves, and muscles. Electric shocks can vary considerably in severity. Severe shocks can cause temporary nerve center paralysis along with chest muscle contractions, resulting in breathing impairment and, if prolonged, death by asphyxiation. Severe shocks can also cause *ventricular fibrillation,* in which fibers of the heart muscles begin contracting in a random uncoordinated pattern. This is one of the most serious electrical injuries because the only known treatment is defibrillation, which requires special equipment and skills to administer.

Other electrical hazards include:

1. Mechanical injuries from electrical motors
2. Fires or explosions resulting from electrical discharges in the presence of dusts and vapors
3. Falls that result from the electrical shock directly or as a result of human reaction afterward

---

**Box 5: Motion Appraisal Checklist**
Methods

| | | A | B |
|---|---|---|---|
| 1. | Are motions within the "normal" horizontal and vertical work areas? | — | — |
| 2. | Are motions restricted? | — | — |
| 3. | Are fixed locations in the proper sequence provided for tools and materials? | — | — |
| 4. | Is preorientation of tools and materials used to advantage? | — | — |
| 5. | Are materials conveyed mechanically or through gravity to the point of use? | — | — |
| 6. | Is the disposal point convenient so that the next movement can be made easily? | — | — |
| 7. | Have worker comforts been reasonably provided for: proper work height, back support, feet rests, etc.? | — | — |
| 8. | Are general working conditions suitable? Orderly? | — | — |
| 9. | Is use of the lowest practical class of motions made possible by the method? | — | — |
| 10. | Are both hands employed in useful work at most all times? | — | — |
| 11. | Are rhythmic motions used? Regular and free from sharp changes in direction? | — | — |
| 12. | Is prepositioning done in transit? | — | — |
| 13. | Is a change of control performed only when necessary? | — | — |
| 14. | Is the same method used by the operator at all times? | — | — |
| 15. | Do stops, guides, or pins aid positioning? | — | — |
| 16. | Does a holding device free hands for useful work? | — | — |
| 17. | Is effort minimized by combined tools, good mechanical leverage, or power tools? | — | — |
| 18. | Does a mechanical ejector remove finished parts? | — | — |
| 19. | Does the equipment effectively separate waste from finished parts? | — | — |
| 20. | Do safety devices eliminate all major hazards? | — | — |
| 21. | Is the foot or leg used to perform some part of the operation? | — | — |
| 22. | Are there postural restrictions that prevent direct forward viewing or cause unnatural postures? | — | — |
| 23. | Are all necessary work points visible by the operator for all realistic natural postures? | — | — |
| 24. | Can the operator both stand and sit? Can all foot pedals and hand-actuated controls be operated properly from both standing and sitting positions without reducing operator stability? | — | — |
| 25. | Are the maximum muscle-activated forces produced when the person's hands, arms, feet, or legs are nearly midway in the natural range of motion? | — | — |
| 26. | Can exertions be provided by either arm or leg when frequent repetitions are likely? | — | — |
| 27. | Are the largest appropriate muscle groups employed in the proper direction? | — | — |
| 28. | Is all the work performed below the person's heart? | — | — |

---

Electrical current flows inversely with resistance. The outer layers of human skin have a high resistance to electricity when the skin is not ruptured or wet. However, wet or broken skin loses 95% or more of its natural resistance. Typically, dry skin has about 400,000 $\Omega$ resistance, while wet skin resistance may only be 300–500 $\Omega$.

The following equation shows that current is equal to voltage divided by resistance:

$$I = E/R \qquad (3)$$

Using equation (3), one can determine that if a person contacts a 120-V circuit, current is only 0.3 mA with dry skin but 240 mA or more with wet skin.[*]

When a person comes into contact with an electrical source, current flows from source to ground. The path

of this source-to-ground flow is critical. If as little as 10 μA reaches the heart muscles, ventricular fibrillation can occur. One minute after the onset of ventricular fibrillation, the chance of survival is 99% if both a defibrillator and the people with the training to use it are available; otherwise the chances are about 1/10,000 for survival.

One cause of electrical accidents is contact with bare power conductors. Ladders, cranes, or vehicles can strike power lines and act as a conducting agent to whomever is touching it. While insulated electric wires offer some protection as long as the insulation is intact, the insulation can break down due to heat or weather or as a result of chemical or mechanical damage. As the insulation weakens, the hazard level approaches that of a bare conductor. In fact, there can be even more danger because the insulation makes the wire still appear to be safe. Insulation breakdown is often accelerated in high-voltage circuits. The corona around high-voltage wires often produces nitrous oxide, which becomes weak nitric

---

[*] Amperage $I$ = 120 V/400,000 $\Omega$ = 0.3 mA with dry skin but $I$ = 120/500 = 240 mA with wet skin.

acid in the presence of moisture; this compound further decomposes insulation.

Equipment failure is another cause of electrical accidents. Internal broken wires can sometimes ground out to the external casing of the equipment. To prevent this, manufacturers of electrical equipment have begun to provide double insulation or a direct-grounding circuit for internal wiring. In addition, ground-fault circuit interrupters are now widely used because these interrupters shut off the circuit almost instantly when the ground circuit exceeds the limit. Many building codes require ground-fault circuits in exterior electrical outlets and those in a garage.

Electrical safety can be improved by designing electrically powered equipment with several features. Most importantly, all electric-powered devices should be designed such that they can be placed in a zero-energy state (i.e., power can be shut off). An associated second feature is the ability to lock out the power via a tag, cover, or hasp and padlock. It is one thing to turn off power; it is another to make sure it stays off, particularly during service or maintenance. A third safety feature is important in power equipment that utilizes energy-storing devices such as capacitors or accumulators. Such equipment should be designed so that an operator can safely discharge the energy from these storage devices without contact.

Sometimes fires and burns can result from electrical devices overheating. Electrical heating systems are particularly prone to overheat. Circuit breakers and thermal lockouts are common devices to prevent those occurrences. These devices may have thermally activated fuses, regular circuit breakers that are thermally activated, or circuit breakers that bend with heat and open the circuit at set upper temperature limits.

Another source of electric hazard is static electricity. *Static electricity* exists when there is an excess or deficiency of electrons on a surface. That surface becomes the source or sink of a very high voltage flow,[*] particularly when the humidity is low. Static electricity occurs often in papermaking, printing, and textile manufacture. A primary danger is that dust and vapor explosions can be ignited with the static electrical discharge.[†] Static electricity can be reduced by using materials that are not prone to static buildup. Another protection is to ground out equipment so that there is no buildup of static electricity. Other prevention measures consist of neutralizing the static electricity and humidifying to reduce the buildup. Lightning is a natural discharge of static electricity. Large conductors and ground rods are used to safely ground out the discharges. Pulse suppressers are often used with computers and other equipment that are vulnerable to damage caused by power spikes, which lightning often produces.

## 6.8 Heat, Temperature, and Fire Hazards

### 6.8.1 Heat/Temperature Hazards

While most industrial hazards in this category are heat related, it should be remembered that very low temperatures associated with freezing[‡] or with cryogenic processes can also be a source of danger.

Burns are the principal result of accidents involving heat. The severity of a burn depends upon the temperature of the heat source and the duration and region of the contact. Milder burns, which usually only redden the skin, are *first-degree* burns. *Second-degree* burns are more serious and are usually associated with blisters on the burned areas of the skin. When the blisters break, there is a chance of infection. The most severe type of burns are *third-degree* burns, which penetrate all the layers of the skin and kill nerve endings. Third-degree burns can cause the affected tissue to appear white, red, or even a charred grey or black. Typically part of the tissue, capillaries, and muscle are destroyed, and gangrene often follows. Freezing burns are similar to heat burns in many ways, including the degrees of severity.

### 6.8.2 Ultraviolet Radiation Hazards

Ultraviolet (UV) radiation can also cause burns. Eyes are particularly vulnerable to UV burns, as eyes are much more sensitive to UV radiation than the skin. UV radiation is emitted during welding, so special eye protection must be worn to protect the welder's eyes. UV rays can also bounce off light-colored walls and pose a hazard to others in the vicinity. For that reason, welding booths are recommended. Other types of equipment that pose similar hazards include drying ovens, lasers, and radars.

### 6.8.3 Fire Hazards

A fire requires three principal elements. These are:

1. Fuel (or reducing agent), which gives up an electron to an oxidizer
2. Oxidizer, which is the substance that acquires the electrons from the fuel
3. Source of ignition, which is anything capable of commencing the oxidation reduction reaction to a sufficient degree for heat or flame

Many substances are always fuels or always oxidizers, but a few will switch roles.[§] Fuels include regular heating fuels, solvents, and any other flammable liquids, gases, or solids.[¶] Oxidizers include oxygen itself and compounds carrying oxygen.[‖]

---

[*] Typically, the static electical flow has very low amperage but extremely high voltage.

[†] In the late 1970s R. R. Donneley Company had a plant in Chicago explode due to static electricity. Paper dust in the overhead trusses was ignited by the static electricity caused by moving paper in the machines.

[‡] Freezing is associated with food processing.

[§] Sometimes a substance can be either a fuel or an oxidizer, depending upon what the other substance is.

[¶] Other examples of fuels are fuels for internal combustion engines or rocket engines, cleaning agents, lubricants, paints, lacquers, waxes, refrigerants, insecticides, plastics and polymers, hydraulic fluids, and products of wood, cloth, paper, or rubber and some metals, particularly as fine powders.

[‖] Other examples of oxidizers are halogen, nitrates, nitrites, peroxides, strong acids, potassium permanganate, and fluorine gas, which is even stronger than pure oxygen.

**Table 11 Flammability Properties for Selected Compounds**

| Compound | Lower Flammability Limit (LFL, %) | Upper Flammability Limit (UFL, %) | Flash Point ($^\circ$C/$^\circ$F) | Auto-Ignition ($^\circ$C/$^\circ$F) |
|---|---|---|---|---|
| Ethanol | 3.0–3.3 | 19 | 12.8 55 | 365/689 |
| Methanol | 6.0–6.7 | 36 | 11.0/52 | 385/725 |
| Methane | 4.5–5.0 | 15–17 | Flammable Gas | 580/1076 |
| Mineral spirits | 0.7 | 6.5 | 38–43/100–109 | 258/496 |
| Gasoline (100 octane) | 1.4 | 7.6 | <−40−40 | 24–280/475–536 |
| Kerosene Jet A-1 | 0.6–0.7 | 4.9–5 | >38/100 as jet fuel | 210/410 |

The key to prevention and control is *keeping these three parts separated in time and/or space*. Airborne gases, vapors, fumes, or dusts are particularly dangerous, since they can travel significant distances to reach ignition sources and often burn explosively.

A flammable material can be a gas, liquid, or solid. Flammable gases burn directly. In contrast to gases, liquids do not burn; they vaporize and their vapors burn. Solids often go through two phase changes before burning, first to liquid and then to gas, but that does not always hold. When the solid is in the form of a fine dust, it becomes much more flammable and even explosive.

Many gases are flammable in the presence of air when there is a sufficient concentration. The lower flammability limit (LFL) of a gas is the lowest concentration of the gas in air that will propagate a flame from an ignition source; below that it is too dilute. The upper flammability limit (UFL) is the highest concentration of a flammable gas in air that will propagate a flame from an ignition source; above that the concentration is too rich to burn. Both of these limits are measured as the percentage of gas by volume. Generally the wider the difference between the UFL and LFL, the greater the range of burning conditions and the more dangerous the substance (Table 11).

Flammable liquids burn when they vaporize and their vapors burn. Several variables affect the rate of vaporization, including the temperature of the liquid, latent heat of vaporization, size of exposed surface, and air velocity over the liquid. The f*lash point* is the lowest temperature at which a liquid has sufficient vapor pressure to form an ignitable mixture with air near the surface of the liquid. The lower the flash point, the easier it is to ignite the material. Note that this term is distinctive from the *fire point,*[*] which is the lowest temperature that will sustain a continuous flame. The fire point is usually a few degrees above the flash point. Clearly, the lower these critical temperatures are, the more sensitive that substance is to burning.

A liquid is classified as *flammable, combustible*, or *nonflammable* based on its flash point. A liquid is flammable if its flash point is below a specified temperature and combustible if its flash point is above this level but below a higher specified level (and so somewhat harder to ignite). Liquids with a flash point above the higher limit (and so most difficult to ignite)

are classified as nonflammable. Different organizations and agencies use different specified limits for these classifications.

Another term of significance to fires is the *autoignition temperature,*[†] which is the lowest temperature at which a material (solid, liquid, or gas) will spontaneously ignite (catch fire) without an external spark or flame. Most organic materials contain the mechanisms for autoignition, which is also referred to as *spontaneous combustion*. Lower grades of coal are volatile and may self-ignite if temperatures build up sufficiently high. Hay, wood shaving, and straw undergo organic decomposition that creates heat internally and may cause self-ignition. Oily rags are particularly dangerous because of the large exposed area to air. When these rags are in a pile, the outer rags insulate those in the center and hold the buildup of heat there. Also, oily insulation around steam lines and heat ducts is another source of self-ignition. *Hypergolic reactions* are special cases of self-ignition where the fuel and oxidizer combust at room temperatures upon mixing. Some of the substances known for hypergolic reactions are white phosphorus, some hydrides, and many iron sulfides that are common wastes in mineral processing or oil fractionation.

Fires inside buildings often have poor ventilation so there is incomplete combustion, which in turn produces carbon monoxide. Carbon monoxide is both toxic and flammable. It is the toxicity of the carbon monoxide more than the smothering effects of carbon dioxide and smoke that is responsible for the highest percentage of fire-connected fatalities. Also carbon monoxide tends to reignite as more ventilation improves the combustion mixture. It is this reignition that results in explosions that frequently occur as fire fighters break windows or holes in the side of a burning building. Toxic conditions of carbon monoxide occur at as low a concentration as 1.28% for a 3 min duration.

Fire-extinguishing methods are specific to the type of fire. Class A fires involve solids that produce glowing embers. Water is the most common fire extinguishing recommended for this class of fires. However, water is *not* recommended for other classes of fires except as a spray alone or with special additives. Class B fires involve gases and liquids such as automotive fuels, greases, or paints. Recommended extinguishants of this class of fires are bromotrifluoromethane, carbon dioxide,

---

[*] A term that is equivalent to fire point is *kindling temperature*.

[†] Spontaneous ignition temperature or combustion temperature has the same meaning as autoignition temperature.

dry chemical foam, and loaded steam. The administration of water to class B fires often spreads the fire. Class C fires are class A or B fires which involve electrical equipment. In this case, recommended extinguishants are bromotrifluoromethane, carbon dioxide, and dry chemical. Care should be taken in class C fires not to use metallic applicators that could conduct electrical current. Finally, class D fires consist of combusting metals, usually magnesium, titanium, zirconium, sodium, and potassium. Temperatures in this class of fires tend to be much greater than in the other classes. There is no general extinguishant for this class of fire, but specific extinguishants are recommended for each type of metal.

One of the best strategies to protect against fires is, of course, to prevent their occurrence by keeping separate the elements that make up fires. A second strategy is early detection. Of course, different detection systems are based on different features that accompany fires. Some detectors have bimetallic strips that expand with a sufficient rise in temperatures to close an alarm circuit. Similar kinds of detectors enclose a fluid that expands under heat until a pressure-sensitive switch is activated. Thermoconductive detectors, another variety, contain insulation between conductor wires. As the heat rises, the insulation value of this special material decreases to a sufficient extent that electrical power flows through that material, creating a new circuit that sets off the alarm. Accompanying the heat of a fire are infrared rays. Thus, infrared photoelectric cells are used in radiant energy detectors. Another form of detector is one based on light interference. These detectors have a glass or plastic tube, through which the air flows, and there is a light source on one side of the tube and a photoelectric cell on the other. When smoke comes from a fire into the tube, the light is refracted from the photoelectric cell and an alarm sounds. Finally, there are ionization detectors that measure changes in the ionization level. Byproducts of combustion cause the air to become ionized, which in turn activates these detectors.

## 6.9  Hazards of Toxic Materials

*Toxic materials* are poisons to most people, meaning that small quantities will cause injury in the average person. Those substances that only affect a small percentage of the population are *allergens.* Toxic substances are typically classified as:

1.  Irritants if they tend to inflame the skin or the respiratory track
2.  Systemic poisons when they damage internal organs
3.  Depressants if they act on the central nervous system
4.  Asphyxiates if they prevent oxygen from reaching the body cells
5.  Carcinogens if they cause cancer
6.  Teratogens if they affect the fetus
7.  Mutagens if the chromosomes of either males or females are affected

A number of substances are well-known *irritants*. Ammonia gas combines with moisture of mucous membranes in the respiratory tract to form ammonium hydroxide, which is a strong caustic agent that causes irritation. Acid baths in plating operations often use chromic acid that not only irritates but also eats holes in the nasal septum. Most of the halogens, including chlorine and phosgene, are strong irritants. Chlorinated hydrocarbon solvents and some refrigerants are converted to phosgene when exposed to open flames. Dusts of various origins[*] inflame the respiratory tract of most people and can cause pneumoconiosis. Long-term exposure to some dusts, even to low concentrations, can result in permanent damage to the health such as lung scarring.

*Systemic poisons* are far more dangerous than irritants. Lead poisoning from paint has been a well-publicized problem. Other metals, such as cadmium and mercury, and some chlorinated hydrocarbons and methyl alcohol, which are used as solvents and degreasing agents, are not as-well-known poisons.

There are a number of substances that are *depressants* or narcotics to the nervous system. While the effects of many are only temporary, depressants pose strong safety hazards because they interfere with human judgment. Best known as a depressant is ethanol, also known as ethyl alcohol, whose impairments to judgment and control are clear. Other depressants include acetylene (used in welding), as well as methanol and benzene, both of which are popular solvents. In addition to being depressants, methanol and benzene are also systemic poisons, while benzene is also a carcinogen (i.e., leukemia).

*Simple asphyxiates* are gases that displace inhaled oxygen, which in turn reduces the oxygen available in the human body. One simple asphyxiate is methane, which is a by-product of fermentation. Other simple asphyxiates include argon, helium, and nitrogen, which are used in welding, as well as carbon dioxide, which has numerous industrial uses and is a byproduct of complete combustion. Problems arise with simple asphyxiates when people must work in closed spaces with large concentrations of these gases. More insidious are the *chemical asphyxiates* that interfere directly with oxygenation of the blood in the lungs and thereby create oxygen deficiencies. One of the best known chemical asphyxiates is carbon monoxide, which is the by-product of incomplete combustion and is typically associated with exhausts of internal combustion engines. This gas has an affinity for hemoglobin that is over 200 times stronger than oxygen and is known as the cause of many deaths. The industrial insecticide hydrogen cyanide is another very dangerous chemical asphyxiate.

*Carcinogenic* substances are in the news frequently, but the degree of danger of these substances varies greatly. As mentioned above, lead is a carcinogen. Another carcinogen is vinyl chloride. Vinyl chloride is extremely dangerous in other ways as it is flammable, giving off phosgene when burning, and it is explosive.

---

[*] Asbestos fibers, silica, and iron oxide are examples of irritants commonly found in industry.

Extreme care must to be exercised in handling these carcinogenic compounds as well as those that are *teratogens* or *mutagens*.

Guidance regarding dangerous substances is available from many sources. OSHA and the U.S. Environmental Protection Agency (EPA) both publish lists of products that should display warnings of danger (29 CFR 1910 Subpart Z). These lists also carry the Chemical Abstracts Service (CAS) identifier numbers to help clarify chemicals with multiple names. However, these lists are very long and some firms are using expert systems and other computer systems to identify these chemical hazards and the various limits of permissible exposure.

### 6.9.1 Measures of Exposure and Exposure Limits

One measure of substance toxicity is the *threshold limit value* (TLV). A TLV is the maximum allowable concentration of a substance to which a person should be exposed; exposures to higher concentrations are hazardous. TLVs are used primarily with reference to airborne contaminants that reach the human body through the respiratory system.

Published TLVs indicate the average airborne concentration that can be tolerated by an average person during exposure for a 40-h week continually over a normal working lifetime. These values are published periodically. As new information becomes available, the American Conference of Governmental Industrial Hygienists (ACGIH, 2010) reviews this information and sometimes changes the TLVs. Therefore, TLVs are usually stated as of certain dates of publication. It should also be stated that the TLVs issued by ACGIH are *indicators* of relative toxicity. The actual danger depends upon other things as well, such as volatility, because more volatile substances generate more gas exposure. OSHA reviews the ACGIH publications of TLVs and may or may not accept those limits. Then OSHA publishes its *permissible exposure limits* (PELs), a term used to distinguish OSHA's exposure limits from consensus standards by ACGIH or any other noted group. PELs and TLVs are highly correlated but not always identical. Many firms adopt the most restrictive limits. Note that some states publish more restrictive limits than those published by OSHA.

TLVs (and PELs) are generally expressed in milligrams per cubic meter. If a substance exists as a gas or vapor at normal room temperature and pressure, its TLV (or PEL) can also be expressed in parts per million (ppm). The relationship between concentrations expressed in units of milligrams per cubic meter and those expressed in units of parts per million (at 25°C and 1 atm of pressure) is

$$\text{TLV (ppm)} = \frac{24.45 \times \text{TLV(mg/m}^3)}{\text{gram molecular weight of substance}} \quad (4)$$

For a few substances, such as asbestos, the TLV (or PEL) is stated in terms of units, or fibers, per cubic centimeter.

Exposure can be measured at a single point in time or as a weighted average over a period of time. TLV and PEL values are generally stated as 8-h *time-weighted averages* (TWAs). An 8-h TWA exposure is computed as

$$\text{TWA} = [C_1 T_1 + C_2 T_2 + \cdots + C_n T_n]/8 \quad (5)$$

where:

  TWA = equivalent 8-h TWA concentration
   $C_i$ = observed concentration in time period $i$
   $T_i$ = duration of time period $i$ in hours
   $n$ = number of time periods studied

When a person is exposed to a single kind of substance hazard, one merely inputs the concentration and time data in the above equation and compares the answer with the PEL corresponding to that substance. If the computed TWA is less than the PEL, then conditions appear moderately safe and the legal requirements are met.

On the other hand, if there is exposure to multiple hazards, then a mixture exposure should be computed as

$$E_m = [C_1/L_1] + [C_2/L_2] + \cdots + [Cm/Lm]$$

where:

  $E_m$ = equivalent ratio for entire mixture
  $C_j$ = concentration of substance $j$
  $L_j$ = PEL of substance $j$
  $m$ = number of different contaminants present in the atmosphere

Safe mixtures occur when $E_m$ is less than *unity*. Even when the exposure to individual substances in a mixture is below each relevant PEL value, the mixture ratio can exceed unity and hence be dangerous; this is more likely as more substances are involved.

There are a number of different ways of finding the concentrations of various substances. Most typically, samples are taken at the site and these bottled samples or special filters are taken to a laboratory for analysis. For certain substances, there are direct-reading instruments that can be used. In addition, for some substances dosimeters are available, which can be worn by employees to alert them when they are in the presence of dangerous exposure levels. Miners frequently use dosimeters for the most common contaminants in mines and people around radioactive materials constantly wear radiation dosimeters on the job.

Additionally, TLVs can also be expressed as *short-term exposure limit* (STEL) or *Emergency exposure limits* (EELs). The STEL denotes the maximum acceptable concentration for a short specified duration of exposure, usually 15 min. These STEL measures are intended for people who are only occasionally exposed to toxic substances. EELs have been introduced more recently and indicate the approximate duration of time a person can

**Table 12  Selected PELs and TLVs from OSHA Chemical Sampling Information File (2010)**

| Substance | CAS No. | PEL mg/m3 | PEL ppm | TLV ppm (TWA) | TLV ppm (STEL)[a] |
|---|---|---|---|---|---|
| Acetic acid | 64-19-7 | 25 | 10 | 10 | 15 |
| Ammonia | 7664-41-7 | 35 | 50 | 25 | 35 |
| Asbestos | 1332-21-4 | 0.1 fiber/cm$^3$ | — | — | 0.1 fiber/cm$^3$ |
| Bromine | 7726-95-6 | 0.7 | 0.1 | 0.1 | 0.2 |
| Carbon dioxide | 124-38-9 | 9000 | 5000 | 5000 | 30,000 |
| Carbon disulfide | 75-15-0 | — | 20 | — | 10 |
| Carbon monoxide | 630-08-0 | 55 | 50 | 25 | — |
| Ethanol | 64-17-5 | 1900 | 1000 | — | — |
| Fluorine | 7782-41-4 | 0.2 | 0.1 | 1 | 2 |
| Furfuryl alcohol | 98-00-0 | 200 | 50 | 10 | 15 |
| Hydrogen cyanide | 74-90-8 | 11 | 10 | — | 4.7 |
| Methyl alcohol | 67-56-1 | 260 | 200 | 200 | 250 |
| Nitric acid | 7697-37-2 | 5 | 2 | 2 | 4 |
| Pentane | 109-66-0 | 2950 | 1000 | 600 | — |
| Phenol | 108-95-2 | 19 | 5 | 5 | — |
| Propane | 74-98-6 | 1800 | 1000 | 2500 | — |
| Stoddard's solvent | 8052-41-3 | 2900 | 500 | 100 | — |
| Sulfur dioxide | 7446-09-5 | 13 | 5 | — | 0.25 |
| Uranium soluble | 7440-61-1 | 0.05 | — | — | — |

[a]STELs are generally set by the ACGIH, not OSHA, and should be compared to TLVs.

be exposed to specified concentrations without ill effect. EELs are important for personnel who work with toxic substance problems.

Table 12 presents some of these limits for a few selected common substances. This table is presented for illustrative purposes only; PELs, TLVs, and STELs are updated regularly and can be found on the OSHA website.

### 6.9.2  Protection from Airborne Contaminants

*Ventilation* is the most common control strategy used for airborne contaminants. The idea is to remove the substance from the air before it reaches people. Ventilating systems are typically designed by mechanical engineers who specialize in this application. An effective ventilating system consists of a contaminant collector, properly sized ductwork laid out to allow smooth airflow from the source to the collector, and exhaust fans which maintain a near-constant negative pressure throughout the system. A number of different types of collectors are used depending upon the substance. Low-pressure cyclone collectors are often used for large particles and sawdust. Electrostatic precipitators, wet-type dust collectors, scrubbers, and fabric collectors are other forms of collectors used for different types of substances.

*Personal protection devices* are also used to protect people from airborne contaminants. Respirators with filters are commonly used in places where there is a lot of dust or vapor that can be neutralized by a chemical filter. Some types of respirators receive air from a compressor much like a diver in the sea. In addition to the respiratory tract, the skin must also be protected. Suits made of special fabrics are frequently used to protect the skin along with gloves for the hands.

*Emergency response* and *cleanup* are other important issues that arises in the use of chemicals and potentially hazardous substances (Box 6).

### 6.10  Transportation Hazards

Transportation-related accidents are the leading cause of occupational deaths (40%). Transportation safety requirements generally fall outside the jurisdiction of OSHA and are controlled by other governmental agencies such as the Department of Transportation (DOT), FAA, and Federal Railroad Administration (FRA).

The most effective means to control transportation hazards is the use of mandated safety features, such as those required by standards/regulations (DOT, FMVSS, SAE). Training employees to properly use modern vehicle safety features (bumpers, side-impact systems, active and passive passenger restraints, etc.) coupled with roadway safety features (signage, traffic lights, lane markings, traffic separation devices, etc.) can be effective in minimizing transportation hazards.

Transportation hazardous can also be related to the cargo being transported. Common hazards from cargo faced when transporting it include materials that are explosive, flammable/combustible, radioactive, oxidizing/corrosive, poison, and etiologic. Controls of hazards related to cargo include excluding materials, limiting quantities, defining storage rules, packaging design and selection, labeling, restricting transportation routes, and training. In addition, the U.S. DOT regulates the transportation of many of these items that have hazards discussed above.

---

**Box 6: Hazardous Waste Operations and Emergency Response (HazWOpER), 29 CFR 1910.120 (1987)**

HazWOpER applies to employers and employees engaged in hazardous substance response and cleanup operations; operations involving hazardous waste storage, disposal, and treatment; operations at sites designated for cleanup; and emergency operations for release of hazardous substances. A HazWOpER plan contains the following elements:

*Site Characterization*. Identify conditions that might be immediately dangerous to life or health and determine appropriate safety and health control measures needed to protect employees.

*Site Control*. Prevent contamination of employees by designating work zones, operating procedures, communication, and medical assistance.

*Training*. Provide hazard and procedure information to employees with 40 h off-site and three days under supervision on-site leading to certification. Define names of responsible personnel, hazards present, signs and symptoms of overexposure, use of engineering controls and equipment, work practices, and PPE.

*Medical Surveillance*. Monitor employees exposed to hazards, including potential exposure and respirator use for at least 30 days/year. This required a medical examination and consultation with a licensed examining physician, informing employee with written opinion of physician, and recordkeeping of above.

*Controls*. Design engineering controls or work practices and ensure effective use of PPE. PPE program must address hazards, task duration, in-use monitoring, and decontamination.

*Monitoring*. Monitor hazards for identification and control purposes.

*Informational*. Inform employees of plan addressing hazards. Requires formal, written, site-available plan, including information on names of key or responsible personnel, task safety analysis, training assignments, PPE, medical surveillance/monitoring plans, control measures/contingency plans, decontamination procedures, and confined-space entry procedures.

*Material Handling*. Set standards for the handling of containers of hazardous materials and wastes.

*Decontamination*. Requires decontamination procedures (employees, clothing, equipment, site), exposure control, communication, assessment, and improvement.

*Emergency Response*. Requires preemergency planning, delineation of responsibilities, recognition and prevention, safe distances and refuge, site security and control, evacuation/decontamination, emergency medical treatment/first aid, PPE and emergency equipment, and evaluation.

*Illumination*. Adequate lighting during associated tasks requires a minimum of 5 footcandles (fc) general, 10 fc in general shops, 30 fc in offices, first aid/medical areas.

*Sanitation*. Requires adequate supply of potable water on-site, identification of nonpotable water supplies, toilet facilities, and, if provided, food, sleeping, and washing facilities in compliance with law.

---

## 7 HAZARD COMMUNICATION*

As outlined in the previous sections, there is much that can be learned about the different types of hazards that might be present in the workplace and on how to avoid them. From an ethical and legal perspective employees have a right to know about these hazards so they can make informed decisions about how to respond to them. Providing safety information to workers can also be thought of from the control perspective as either a way of preventing unsafe acts by alerting, reminding, or instructing people what to do or as a strategy for ensuring safe behavior by building safety awareness and motivating people to behave safely.

### 7.1 Legal Requirements

In most industrialized countries, governmental regulations require that certain forms of safety information be provided to workers. For example, in the United States, the EPA has developed several labeling requirements for toxic chemicals. The DOT makes specific provisions

regarding the labeling of transported hazardous materials. OSHA has promulgated a hazard communication standard that applies to workplaces where toxic or hazardous materials are in use. Training, container labeling, and material data safety sheets are all required elements of the OSHA hazard communication standard.

In the United States, the failure to warn also can be grounds for litigation holding manufacturers and others liable for injuries incurred by workers. In establishing liability, the *theory of negligence* considers whether the failure to adequately warn is unreasonable conduct based on (1) the foreseeability of the danger to the manufacturer, (2) the reasonableness of the assumption that a user would realize the danger, and (3) the degree of care that the manufacturer took to inform the user of the danger. The *theory of strict liability* only requires that the failure to warn caused the injury or loss.

### 7.2 Sources of Safety Information

Manufacturers and employers throughout the world provide a vast amount of safety information to workers. The many sources of safety information now made available to workers include materials provided in training courses, material safety data sheets, written procedures, safety signs, product labels, and instruction manuals.

---

* The material in this section is adapted for the most part from Lehto and Miller (1997).

**Table 13  Objectives and Example Sources of Safety Information Mapped to Accident Sequence**

| | Task Stage in Accident Sequence | | | |
| | Prior to Task | Routine Task Performance | Abnormal Task Conditions | Accident Conditions |
| --- | --- | --- | --- | --- |
| Objectives (behavioral) | Educate and persuade worker of the nature and level of risk, precautions, remedial measures, and emergency procedures | Instruct or remind worker to follow safe procedures or take precautions | Alert worker of abnormal conditions; specify needed actions | Indicate locations of safety and first aid equipment, exits, and emergency procedures; specify remedial and emergency procedures |
| Example sources | Training Manuals, videos, or programs; hazard communication programs; MSDSs; safety propaganda safety feedback | Instruction manuals; job performance aids; checklists; written procedures; warning signs and labels | Warning signals: visual, auditory, or olfactory; temporary tags, signs, barriers, or lock-outs | Safety information signs, labels, and markings; MSDSs |

*Source*: Reprinted with permission from Lehto and Miller (1997).

Information provided by each of these sources varies in its behavioral objectives, intended audience, content, level of detail, format, and mode of presentation. Each source also provides its information at different stages in a way that corresponds to behavioral objectives and when it is provided (Table 13).

Sources of information, such as safety training materials, hazard communication programs, and various forms of safety propaganda, including safety posters and campaigns, are used to educate workers about risks and persuade them to behave safely. Such information is often provided away from the job in the classroom or safety meetings. Occasional safety and health meetings are especially appropriate when things are changing in manufacturing areas that may have safety and health implications. Meetings called to inform operators of some expected changes and to elicit employee suggestions about details of those changes are also very valuable, especially if they affect safety and health.[*] Inexperienced workers are often the target audience, and the information provided is often quite detailed and focused on building safety awareness and motivating people to behave safely.

Other sources of information, such as written procedures, checklists, instructions, warning signs, and product labels, often provide critical safety information to the operator during routine task performance. This information usually consists of brief statements that either instruct less skilled workers or remind skilled workers to take necessary precautions. Following this approach can help prevent workers from omitting precautions or other critical steps in a task. Statements providing such information are often embedded at the appropriate stage within step-by-step instructions describing how to perform a task. Warning signs at appropriate locations can

play a similar role. For example, a warning sign located at the entrance to a workplace might state that hard hats must be worn before entering.

At the next stage in the accident sequence, highly conspicuous and easily perceived sources of safety information alert workers of abnormal or unusually hazardous conditions. Examples include warning signals, safety markings, tags, signs, barriers, or lock-outs. Warning signals can be visual (flashing lights, movements, etc.), auditory (buzzers, horns, tones, etc.), olfactory (odors), tactile (vibrations), or kinesthetic. Certain warning signals are inherent to products when they are in hazardous states (i.e., the odor released upon opening a container of acetone). Others are designed into machinery or work environments. Safety markings refer to methods of nonverbally identifying or highlighting potentially hazardous elements of the environment (i.e., by painting step edges yellow or emergency stops red). Safety tags, barriers, signs, or lock-outs are placed at the point of hazard and are often used to prevent workers from entering areas or activating equipment during maintenance, repair, or other abnormal conditions.

At the final stage in the accident sequence, the focus is on expediting worker performance of emergency procedures at the time an accident is occurring or performance of remedial measures shortly after the accident. Safety information signs and markings conspicuously indicate facts critical to adequate performance of emergency procedures (e.g., the locations of exits, fire extinguishers, first-aid stations, emergency showers, eye wash stations, emergency releases). Product safety labels and material safety data sheets may specify remedial and emergency procedures to be followed.

Before safety information can be effective, at any stage in the accident sequence, it must first be noticed and understood and, if the information is not immediately applicable, also be remembered. Then, the worker must both decide to comply with the provided message and be physically able to do so. Successfully attaining

---

[*] Employee participation greatly helps employee motivation and the building of a team effort.

each of these steps for effectiveness can be difficult. Guidelines describing how to design safety information are of some assistance, as expanded upon below.

### 7.3 Design Guidelines

Standard-setting organizations, regulatory agencies, and the courts through their decisions have traditionally both provided guidelines and imposed requirements regarding when and how safety information is to be provided. More recently, there has been a trend toward developing guidelines based on scientific research concerning the factors which influence the effectiveness of safety information.

### 7.3.1 Voluntary Standards

A large set of existing standards provide voluntary recommendations regarding the use and design of safety information. These standards have been developed by both (1) international groups, such as the United Nations, the European Economic Community (EURONORM), the International Organization for Standardization (ISO), and the International Electrotechnical Commission (IEC), and (2) national groups, such as the American National Standards Institute (ANSI), the British Standards Institute, the Canadian Standards Association, the German Institute for Normalization (DIN), and the Japanese Industrial Standards Committee.

Among consensus standards, those developed by ANSI in the United States are of special significance. Seven ANSI standards focusing on safety signs, labels, and manuals have been developed and revised on a recurring basis: (1) ANSI Z535.1, *Safety Color Code*; (2) ANSI Z535.2, *Environmental and Facility Safety Signs*; (3) ANSI Z535.3, *Criteria for Safety Symbols*; (4) ANSI Z535.4, *Product Safety Signs and Labels*; (5) ANSI Z535.5, *Safety Tags and Barricade Tapes (for Temporary Hazards)*; (6) ANSI Z535.6, *Product Safety Information in Product Manuals, Instructions, and Other Collateral Materials*. Other standards, such as (7) ANSI Z129.1, *Hazardous Industrial Chemicals—Precautionary Labeling*, should be reviewed and applied when required.

### 7.3.2 Design Specifications

Design specifications can be found in the consensus and governmental safety standards discussed above specifying how to design (1) material safety data sheets (MSDSs), (2) instructional labels and manuals, (3) safety symbols, and (4) warning signs, labels, and tags.

***Material Safety Data Sheets*** The OSHA hazard communication standard (29 CFR 1910.1200) specifies that employers must have a MSDS in the workplace for each chemical that is considered a health hazard. Most chemicals used today will have MSDSs provided by the supplier. In addition, many online resources are available for MSDSs, such as Cornell University (http://www.ehs.cornell.edu/msds/msds.cfm).

For in-process or internally used chemicals only, the standard requires that each sheet be written in English, list its date of preparation, and provide the chemical and common name of hazardous chemicals contained. It also requires the MSDS to describe (1) physical and chemical characteristics of the hazardous chemical, (2) physical hazards, including potential for fire, explosion, and reactivity, (3) health hazards, including signs and symptoms of exposure, and health conditions potentially aggravated by the chemical, (4) the primary route of entry, (5) the OSHA permissible exposure limit, the ACGIH threshold limit value, or other recommended limits, (6) carcinogenic properties, (7) generally applicable precautions, (8) generally applicable control measures, (9) emergency and first-aid procedures, and (10) the name, address, and telephone number of a party able to provide, if necessary, additional information on the hazardous chemical and emergency procedures.

Several publications provide assistance in creating MSDSs, such as OSHA form 174 (http://www.osha.gov/dsg/hazcom/msds-osha174/msdsform.html) and ANSI Z400.1: Hazardous Industrial Chemicals—Material Safety Data Sheets—Preparation.

***Instructional Labels and Manuals*** The recent ANSI Z535.6 product safety information in product manuals, instructions, and other collateral materials was published in 2006. This standard focuses on the elements and format that should be considered when developing safety information other than warning labels and signs. This usually takes the form of user manuals and instructions.

***Safety Symbols*** Numerous standards throughout the world contain provisions regarding safety symbols. Among such standards, the ANSI Z535.3 standard, Criteria for Safety Symbols, is particularly relevant for industrial practitioners. The standard presents a significant set of selected symbols shown in previous studies to be well understood by workers in the United States. Perhaps more importantly, the standard also specifies methods for designing and evaluating safety symbols. Important provisions include (1) new symbols must be correctly identified during testing by at least 85% of 50 or more representative subjects, (2) symbols which do not meet the understandability criteria should only be used when equivalent word messages are also provided, and (3) employers and product manufacturers should train users regarding the intended meaning of the symbols. The standard also makes new symbols developed under these guidelines eligible to be considered for inclusion in future revisions of the standard.

***Warning Signs, Labels, and Tags*** ANSI and other standards provide very specific recommendations for how to design warning signs, labels, and tags. These include, among other factors, particular signal words and text, color-coding schemes, typography, symbols, arrangement, and hazard identification (Table 14).

The most commonly used signal words are DANGER, to indicate the highest level of hazard; WARNING, to represent an intermediate hazard; and CAUTION, to indicate the lowest level of hazard. Color-coding methods, also referred to as a "color system," consistently associate colors with particular levels of hazard. For example, red is used in all of the

**Table 14 Summary of Recommendations in Selected Warning Systems**

| System | Signal Words | Color Coding | Typography | Symbols | Arrangement |
|---|---|---|---|---|---|
| ANSI Z129.1 Precautionary Labeling of Hazardous Chemicals | Danger<br>Warning<br>Caution<br>Poison<br>optional words for "delayed" hazards | Not specified | Not specified | Skull-and-crossbones as supplement to words; acceptable symbols for three other hazard types | Label arrangement not specified; examples given |
| ANSI Z535.2 Environmental and Facility Safety Signs | Danger<br>Warning<br>Caution<br>Notice<br>[general safety]<br>[arrows] | Red<br>Orange<br>Yellow<br>Blue<br>Green<br>as above; B&W otherwise<br>per ANSI Z535.1 | Sans serif, uppercase, acceptable typefaces, letter heights | Symbols and pictographs per ANSI Z535.3 | Defines signal word, word message, symbol panels in 1–3 panel designs; four shapes for special use; can use ANSI Z535.4 for uniformity |
| ANSI Z535.4 Product Safety Signs and Labels | Danger<br>Warning<br>Caution | Red<br>Orange<br>Yellow<br>per ANSI Z535.1 | Sans serif, uppercase, suggested typefaces, letter heights | Symbols and pictographs per ANSI Z535.3; also SAE J284 Safety Alert Symbol | Defines signal word, message, pictorial panels in order of general to specific; can use ANSI Z535.2 for uniformity; use ANSI Z129.1 for chemical hazards |
| ANSI Z535.6 Product safety information in product manuals, instructions, and other collateral materials | Danger<br>Warning<br>Caution | Red<br>Orange<br>Yellow<br>per ANSI Z535.1 | Sans serif, uppercase, suggested typefaces, letter heights | Symbols and pictographs per ANSI Z535.3; also SAE J284 Safety Alert Symbol | Defines signal word, message, pictorial panels in order of general to specific; can use ANSI Z535.2 for uniformity |
| OSHA 1910.145 Specification for Accident Prevention Signs and Tags | Danger<br>Warning (tags only)<br>Caution<br>Biological Hazard, BIOHAZARD, or symbol<br>[safety instruction]<br>[slow-moving vehicle] | Red<br>Yellow<br>Yellow<br>Fluorescent Orange/Orange-Red<br>Green<br>Fluorescent Yellow-Orange & Dark Red<br>per ANSI Z535.1 | Readable at 5 ft or as required by task | Biological Hazard Symbol. Major message can be supplied by pictograph (tags only); Slow-Moving Vehicle (SAE J943) | Signal word and major message (tags only) |
| OSHA 1910.1200 [Chemical] Hazard Communication | Per applicable requirements of EPA, FDA, BATF, and CPSC | | In English | | Only as MSDS |

*Source*: Adapted from Lehto and Miller (1986) and Lehto and Clark (1990).

standards in Table 9 to represent the highest level of danger. Explicit recommendations regarding typography are given in nearly all the systems. The most general commonalty between the systems is the recommended use of san serif typefaces. Varied recommendations are given regarding the use of symbols and pictographs.

Certain standards also specify the content and wording of warning signs or labels in some detail. For example, ANSI Z129.1 specifies that chemical warning labels include (1) identification of the chemical product or its hazardous component(s), (2) signal word, (3) statement of hazard(s), (4) precautionary measures, (5) instructions in case of contact or exposure, (6) antidotes, (7) notes to physicians, (8) instructions in case of fire and spill or leak, and (9) instructions for container handling and storage. This standard also specifies a general format for chemical labels that incorporate these items. The standard also provides extensive and specific recommended wordings for particular messages.

### 7.3.3 Cognitive Guidelines

Design specifications, such as those discussed above, can be useful to developers of safety information. However, many products and situations are not directly addressed by standards or regulations. Certain design specifications are also scientifically unproven. In extreme cases, conforming with standards and regulations can reduce the effectiveness of safety information. To ensure effectiveness, developers of safety information consequently may need to go beyond safety standards. To address this issue, the International Ergonomics Association (IEA) and International Foundation for Industrial Ergonomics and Safety Research (IFIESR) supported an effort to develop guidelines for warning signs and labels (Lehto, 1992) which reflect published and unpublished studies on effectiveness and have implications regarding the design of nearly all forms of safety information.

Six of these guidelines, presented in slightly modified form, are as follows: (1) Match sources of safety information to the level of performance at which critical errors occur for the relevant population. (2) Integrate safety information into the task and hazard-related context. (3) Be selective. (4) Make sure that the cost of complying with safety information is within a reasonable level. (5) Make symbols and text as concrete as possible. (6) Simplify the syntax of text and combinations of symbols. Satisfying these guidelines requires consideration of a substantial number of detailed issues as addressed in the earlier parts of this chapter.

## 8  FINAL REMARKS

Much more information is available on each of the topics noted here, and many issues are not addressed. However, the presented material should be enough to familiarize the reader with the importance of safety and health, types of occupational hazards, causes of accidents, control strategies, and important functions of occupational safety and health management. There are numerous books on the subject of occupational safety and health management and engineering as well as specialty books on subjects such as toxicology, human error, system engineering, reliability, and human factors engineering. The Web is another important source of information. Most safety organizations, such as the NSC, Board of Certified Safety Professionals, OSHA, NIOSH, and BLS, have websites offering large amounts of helpful information. Contact information for many of these sources of additional information is provided below.

### Acknowledgment

## REFERENCES

Altman, J. W. (1964), "A Central Store of Human Performance Data," in *Proceedings of the Symposium on Quantification of Human Performance*, Albuquerque, NM, August 17–19, pp. 97–108.

Altman, J. W. (1967, May), "Classification of Human Error," in *Symposium on Reliability of Human Performance in Work*, W.B. Askgren, Ed., Technical Report No. 67-88, Aerospace Medical Labs, Wright-Patterson AFB, OH.

American Conference of Governmental Industrial Hygienists (ACGIH), *Threshold Limit Value* (*TLV*®) *and Biological Exposure Indices* (*BEIs*®), ACGIH, 2010.

American Society for Testing and Materials (ASTM) (1975), *The Voluntary Standards System of the United States of America: An Appraisal by the American Society for Testing and Materials*, ASTM, Philadelphia, PA.

Asfahl, C. R., and Rieske D. W. (2010), *Industrial Safety and Health Management*, 6th ed., Prentice-Hall, Englewood Cliffs, NJ.

Behm, M. (2005), "Linking Construction Fatalities to the Design for Construction Safety Concept," *Safety Science*, No. 43, pp. 589–611.

Blake, R. P. (1963), *Industrial Safety*, Prentice-Hall, Englewood Cliffs, NJ.

Brauer, R. L. (2005), *Safety and Health for Engineers*, 2nd ed., Wiley, New York.

Bureau of Labor Statistics (BLS) (2008), "TABLE R4: Number of Nonfatal Occupational Injuries and Illnesses Involving Days Away from Work by Industry and Selected Events or Exposures Leading to Injury or Illness," BLS, Washington, DC.

Cohen, A., and Margolis, B. (1973), "Initial Psychological Research Related to the Occupational Safety and Health Act of 1970," *American Psychologist*, July, pp. 600–606.

Cooper, J. (1961), "Human Initiated Failures and Malfunction Reporting," *IRE Transactions*, Human Factors Electronics, Vol. 2, pp. 104–109.

Department of Labor (2005), OSHA Recordkeeping Handbook: The Regulation and Related Interpretations for Recording and Reporting Occupational Injuries and Illnesses, Publication 3245-09R, U.S. Department of Labor, Washington, DC.

DeReamar, R. (1958), *Modern Safety Practices*, Wiley, New York.

Embrey, D. E., Humphreys, P., Rosa, E. A., Kirwin, B., and Rea, K. (1984, March), *SLIM-MAUD: An Approach to*

*Assessing Human Error Probabilities Using Structural Expert Judgement*, Vol. 1, *Overview of SLIM-MAUD*, NUREG/CR-3518, U.S. Nuclear Regulatory Commission, Washington, DC.

Epp, S., and Konz, S. (1975), "Appliance Noise: Annoyance and Speech Interference," *Home Economics Research Journal*, Vol. 3, No. 3, pp. 205–209.

Flanagan, J. C. (1954), "The Critical Incident Technique," *Psychological Bulletin*, Vol. 51.4, pp. 327–359.

Geller, E. S., and Williams, J. H. (2001), *Keys to Behavior-Based Safety: From Safety Performance Solutions*, ABS Consulting, Rockville, MD.

Goetsch, D. L. (2008), *Occupational Safety and Health for Technologists, Engineers, and Managers*, 7th ed., Prentice-Hall, Englewood Cliffs, NJ.

Haddon, W. (1975), "Reducing the Damage of Motor Vehicle Use," *Technology Review*, Vol. 77, No. 8, pp. 53–59.

Haddon, W., Suchman, E. A., and Klein, D. (1964), *Accident Research*, Harper and Row, New York.

Hammer, W., and Price, D. (2000), *Occupational Safety Management and Engineering*, 5th ed., Prentice-Hall, Englewood Cliffs, NJ.

Hanley, W., Ed. (1969), *Industrial Safety Handbook*, McGraw-Hill, New York.

Heinrich, H. W. (1959), *Industrial Accident Prevention— A Scientific Approach*, McGraw-Hill, New York, New York.

Heinrich, H. W., Peterson, D., and Roos, N. (1980), *Industrial Accident Prevention*, 5th ed., McGraw-Hill.

Hirano, H. (1996), 5S for Operators: 5 pillars of the Visual Workplace for Your Organization! Productivity Press, Portland, OR.

Hoyos, C., and Zimolong, B. (1988), *Occupational Safety and Accident Prevention: Behavioral Strategies and Methods*, Elsevier, Amsterdam.

Johnson, W. G. (1973), "Sequences in Accident Causation," guest editorial, *Journal of Safety Research*, Vol. 5, No. 2, pp. 54–57.

Johnson, W. G. (1975), "MORT: The Management Oversight and Risk Tree," *Journal of Safety Research*, Vol. 7, pp. 4–15.

Johnson, W. (1980), *MORT Safety Assurance Systems*, Marcel Dekker, New York.

Kowalsky, N., Masters, R., Stone, R., Babcock G., and Rypka, E. (1974), "An Analysis of Pilot Error-Related Aircraft Accidents," Report No. CR-2444, NASA, Washington, DC.

Krlyter, K. (1970), *The Effects of Noise on Man*, Academic, New York.

Laughery, K. R., and Hammond, A. (1999), "Overview," in *Warnings and Risk Communication*, M. Wogalter, D. DeJoy, and K. Laughery. Eds., Taylor & Francis, London, pp. 3–14 .

Lehto, M. R. (1992), "Designing Warning Signs and Warning Labels: Part I—Guidelines for the Practitioner," *International Journal of Industrial Ergonomics*, Vol. 10, Nos. 1--2, pp. 105–113.

Lehto, M. R. (2006), "Human Factors Models," in *Handbook of Warnings*, M. S. Wogalter, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, Chapter 6, pp. 63–87.

Lehto, M. R., and Buck, J. R. (2008), *An Introduction to Human Factors and Ergonomics for Engineers*, Lawrence Erlbaum Associates, Taylor and Francis Group, New York.

Lehto, M. R., and Clark, D. R. (1990), "Warning Signs and Labels in the Workplace," in *Workspace, Equipment and Tool Design*, W. Karwowski and A. Mital, Eds., Elsevier, Amsterdam, pp. 303–344.

Lehto, M. R., and Miller, J. M. (1986), *Warnings*, Vol. I, *Fundamentals, Design, and Evaluation Methodologies*, Fuller Technical Publications, Ann Arbor, MI.

Lehto, M. R., and Miller, J. M. (1997), "Principles of Prevention: Safety Information," in J. M. Stellman, Ed., *ILO Encyclopedia of Occupational Health and Safety*, 4th ed., International Labour Office, Geneva, Switzerland, pp. 56.33–56.38.

Lehto, M. R., and Salvendy, G. (1991), "Models of Accident Causation and Their Application: Review and Reappraisal," *Journal of Engineering and Technology Management*, Vol. 8, pp. 173–205

Liberty Mutual (2009), *2009 Liberty Mutual Workplace Safety Index*, Liberty Mutual, Hopkinton, MA.

Mattila, M., and Karwowski, W., Eds. (1992), "Computer Applications in Ergonomics, Occupational Safety and Health," in *Proceedings of the International Conference on Computer-Aided Ergonomics and Safety '92-CAES' 92*, Tampere, Finland, North Holland, Amsterdam, Netherlands.

Miller, J. M., Lehto, M. R., and Frantz, J. P. (1990), *Instructions & Warnings*, Fuller Technical Publications, Ann Arbor, MI.

National Institute for Occupational Safety and Health (NIOSH) (1983, March 29), *Vibration Syndrome*, Publication No. 83-110, NIOSH, Washington, DC.

National Institute for Occupational Safety and Health (NIOSH) (1995, September), *Cumulative Trauma Disorders in the Workplace: Bibliography*, Publication No. 95–119, NIOSH, Washington, DC.

National Institute for Occupational Safety and Health (NIOSH) (2010), *Occupationally-Induced Hearing Loss*, Publication No. 2010-136, NIOSH, Washington, DC.

National Safety Council (NSC) (2009, January), *Accident Prevention Manual*, 13th ed., NSC, Itasca, IL.

National Safety Council (NSC) (2010), *Injury Facts*, NSC, Chicago, IL.

Norman, D. A. (1988), *The Psychology of Everyday Things*, Basic Books, New York.

Norman, D. A. (1992), "Turn Signals Are the Facial Expressions of Automobiles," Addison-Westley, Reading, MA.

Occupational Injury and Illness Classification System (OIICS), developed by the Bureau of Labor Statistics (BLS) in the United States, available: http://www.bls.gov/iif/oshoiics.htm.

Olishifski, J. B., and McElroy, F. E. (1971), *Fundamentals of Industrial Hygiene*, National Safety Council, Chicago.

Perkins, H. C. (1975), *Air Pollution*, McGraw-Hill Book, New York.

Ramsden, J. (1976), *The Safe Airline*, McDonnalds and James, London.

Reason, J. (1990), *Human Error*, Cambridge University Press, Cambridge.

Rogers, W. P. (1971), *Introduction to System Safety Engineering*, Wiley, New York.

Roland, H. E., and Moriarty, B. (1990), *Systems Safety Engineering and Management*, Wiley, New York.

Roughton, J. E., and Crutchfield, N. (2008), *Job Hazard Analysis*, Elsevier, Amsterdam.

Salvendy, G. (2001), *Handbook of Industrial Engineering, Technology, and Operations Management*, 3rd ed., Salvendy, G., Ed, Wiley, pp. 2172–2223.

Simons, R. H., and Grimaldi, J. V. (1963), *Safety Management*, R. D. Irwin, Homewood, IL.

Suokas, J., and Pyy, P. (1988), "Evaluation of the Validity of Four Hazard Identification Methods with Event Descriptions," Research Report 516, Technical Research Centre of Finland, Espoo, Finland.

Suokas, J., and Rouhiainen, V. (1984), "Work Safety Analysis: Method Description and User's Guide," Research Report No. 314, Technical Research Centre of Finland, Tampere, Finland.

Swain, A. D., and Guttmann, H. E. (1983), *Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications*, NUREG/CR-1278, U.S. Nuclear Regulatory commission, Washington, DC.

Van Cott, H. P., and Kincade, R. G. (1972), *Human Engineering Guide to Equipment Design*, U.S. Government Printing Office, Washington, DC.

Wagenaar, W. A. (1992), "Risk Taking and Accident Causation," in *Risk-Taking Behavior*, J. F. Yates, Ed., Wiley, New York, pp. 257–281.

Wagenaar, W. A., and Groeneweg, J. (1988), "Accidents at Sea: Multiple Causes and Impossible Consequences," in E. Hollnagel, G. Mancini, and D. D. Woods, Eds., *Cognitive Engineering in Complex Dynamic Worlds*, Academic, London, pp. 133–144.

Zimolong, B. (1997), "Occupational Risk Management," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York, Chapter 31 pp. 989–1020.

## RESOURCES

American National Standards Institute (ANSI), 1819 L Street, NW, Suite 600, Washington, DC 20036, www.ansi.org.

American Society of Safety Engineers (ASSE), 1800 E Oakton St, Des Plaines, IL 60018, www.asse.org.

ASTM International (originally known as the American Society for Testing and Materials), 100 Barr Harbor Drive, West Conshohocken, PA, 19428-2959, www.astm.org.

Board of Certification in Professional Ergonomics, PO Box 2811, Bellingham WA 98227-2811, www.bcpe.org.

Board of Certified Safety Professionals, 208 Burwash Avenue, Savoy, IL 61874, www.bcsp.org.

Department of Labor, 29 Code of Federal Regulation (CFR) 1910.95: Occupational Noise Exposure, www.osha.gov.

Department of Labor, 29 Code of Federal Regulation (CFR) 1910 Subpart D: Walking-Working Surfaces, www.osha.gov.

Department of Labor, 29 Code of Federal Regulation (CFR) 1910 Subpart O: Machine Guarding, Part 212: General Requirements for all machines, www.osha.gov.

Department of Labor, 29 Code of Federal Regulation (CFR) 1910 Subpart Z: Toxic and Hazardous Substances, Part 1200: Hazard Communication, www.osha.gov.

Department of Labor, 29 Code of Federal Regulation (CFR) 1910 Subpart Z: Toxic and Hazardous Substances, www.osha.gov.

Human Factors and Ergonomics Society (HFES), P.O. Box 1369, Santa Monica, CA 90406-1369, www.hfes.org.

National Safety Council (NSC), 1121 Spring Lake Drive, Itasca, IL 60143-3201, www.nsc.org.

National Society of Professional Engineers, 1420 King Street, Alexandria, VA 22314, www.nspe.org.

# CHAPTER 26

# HUMAN ERROR AND HUMAN RELIABILITY ANALYSIS

**Joseph Sharit**
**University of Miami**
**Coral Gables, Florida**

## 1 INTRODUCTION

### 1.1 Some Perspectives on Human Error

A fundamental objective of human factors and ergonomics is to design and facilitate the control of various artifacts, such as devices, systems, interfaces, rules, and procedures, to enable safe and effective performance outcomes. In principle, the realization of this goal entails a detailed understanding of how these artifacts might bear upon the limitations and capabilities of their users. It is when the consequences related to the artifact's use are judged to be sufficiently harmful or inappropriate that people may be conferred with the attribution of human error. These people might include those responsible for conceptualizing and designing the artifact; those responsible for installing, maintaining, or providing instruction on its use; those who determine and oversee the rules governing its use; or those who actually use it.

Concerns for human error were a major influence in establishing the area of human factors (Helander, 1997)

and have since become increasingly emphasized in product and system design and in the operations of various organizations. Human error is also often on the minds of the general public as they acknowledge, if not fully understand, the failures in their everyday interactions with products or in their situational assessments or are swept up into the media's coverage of high-profile accidents that are often attributed to faulty human actions or decisions. Yet, despite the apparent ubiquity of human error, its attribution has been far from straightforward.

For example, for a good part of the twentieth century the dominant perspective on human error by many U.S. industries was to attribute adverse outcomes to the persons whose actions were most closely associated to these events—that is, to the people who were working at what is now often referred to as the "sharp end." Likewise, most aircraft crashes were historically blamed on pilot error and, as in the industrial sector, there was little inclination to scrutinize the design of the tools or system

or the situations with which the human was expected to coexist.

In contrast, in the more current perspective the human is deemed to be a reasonable entity at the mercy of an array of design, organizational, and situational factors which can lead to behaviors external observers come to regard, although to some often unfairly, as human errors. The appeal of this view should be readily apparent in each of the following two cases. The first case involves a worker who is subjected to performing a task in a restricted space. While attempting to reach for a tool, the worker's forearm inadvertently brushes against a switch whose activation results in the emission of heat from a device. Visual feedback concerning the activation is not possible, due to the awkward posture the worker must assume; tactile cues are not detectable due to requirements for wearing protective clothing; and auditory feedback from the switch's activation, which is normally barely audible, is not perceived due to ambient noise levels. Residual vapors originating from a rarely performed procedure during the previous shift ignite, resulting in an explosion.

In the second case, a worker adapts the relatively rigid and unrealistic procedural requirements dictated in a written work procedure to demands that continually materialize in the form of shifting objectives, constraints on resources, and changes in production schedules. Management tacitly condones these procedural adaptations, in effect relying on the resourcefulness of the worker for ensuring that its goals are met. However, when an unanticipated scenario causes the worker's adaptations to result in an accident, management is swift to renounce any support of the worker's actions that were in violation of work procedures.

In the first case the worker's action that led to the accident was unintentional; in the second case the worker's actions were intentional. In both cases, however, whether the person committed an *error* is debatable. One view that is consistent with this position would shift the blame for the adverse consequences from the *actor* to management or the designers. *Latent management* or *latent designer errors* (Reason, 1997)—that is, actions and decisions that occurred at the "blunt end"—would thus absolve the actor from human error in each of these cases. The worker, after all, was in the heat of the battle, performing "normal work," responding to the work context in reasonable, even skillful ways. The human does not want his or her actions or decisions to result in a negative consequence; in fact, such a desire would constitute sabotage, which is not within the realm of the topic of human error. In a sense, the human was "set up to fail" due to the context or situation in which they were operating (Spurgin, 2010).

Of course, the process of shifting blame does not have to end with designers. In the current landscape of global competition designers may face pressures that limit their ability to adequately investigate the conditions under which their products will be used or to become sufficiently informed about the knowledge and resources users would have available to them when using these products. Management, or the organizations they represent and lead, would then seem to be the true architects of human failure, but even this attribution of blame may be misleading. Regulatory agencies or even the federal government may have laid the groundwork, by virtue of poor assessments of needed control mechanisms and through misdirected priorities, for the faulty policies and decisions on the part of organizations and ultimately for shaping or at least impacting the entrepreneurial and managerial cultures of these organizations (Section 6). Governments themselves, however, could hardly be expected to provide certainty in solutions and policies as they struggle to make sense of the information streaming from the complex and fluid milieu of social, political, and economic forces.

Nonetheless, societal and organizational prescriptions in the form of policies and other types of remedies are powerful forces. Workers who disagree with these policies, such as the worker in the case above who adapted a procedure in the face of existing evidence, are thus risking being blamed for negative outcomes, especially when such forces that run through an organization are "hidden and undisclosed" (Dervin, 1998). At any rate, what should be fairly clear is that attempts at pinpointing the latent sources of human error or resolving how latent sources can collectively contribute to human error can be far from straightforward.

Another basis for dismissing attributions of human error derives from the doubt capable of being cast on the error attribution process itself (Dekker, 2005). By virtue of having knowledge of events, especially bad events such as accidents, outside observers are able (and perhaps even motivated) to invoke a backward series of rationalizations and logical connections that has neatly filtered out the subtle and complex situational details that are likely to be the basis for the perpetrating actions. Whether this process of establishing causality is due to convenience, *hindsight bias* (Fischhoff, 1975; Christoffersen and Woods, 1999; Dekker, 2001), or the inability to determine or comprehend the perceptions and assessments made by the actor that interlace the more prominently observable events, the end result is a considerable underestimation of the influence of the context within which the person acts. Ultimately, this obstruction at establishing cause and effect jeopardizes the ability to learn from accidents and consequently the ability to predict or prevent future failures.

Even the workers themselves, if given the opportunity in each of these cases to examine or reflect upon their performance, may acknowledge their actions as errors, easily spotting all the poor decisions and improperly executed actions, when in reality, within the frames of references at the time the behaviors occurred, their actions were in fact reasonable and constituted "mostly normal work." The challenge, according to Dekker (2005), is "to understand how assessments and actions that from the outside look like errors become neutralized or normalized so that from the inside they appear unremarkable, routine, normal" (p. 75).

This view is also very consistent with that of Hollnagel (2004), who considers both normal human performance and performance failures (outcomes of actions that differ from what was intended or required) as *emergent properties* of mutual dependencies that are

induced by the complexity and demands arising from the entire system. Therefore, it is not so much the variability of human actions that is responsible for failures but the variability in the context and conditions to which the human is trying to adjust. This variability can result in irregular and unpredictable inputs (e.g., other people in the system acting in unexpected ways); incompatibility between demands (e.g., conflicting or unreasonable production requests) and available resources (e.g., lack of time, lack of training or experience for handling the situation, or limits in cognitive capacity); and working conditions falling outside of normal limits (e.g., noise, poor communication channels, or inappropriate work schedules). Furthermore, system outputs that fail to comply with expectations can result in protracted irregularity in inputs, leading to cycles of variability to which the human must adjust.

While it is fair to assume that normally humans do not want to commit errors, there are situations where human error is not only acceptable but also desirable. Mistakes during training exercises are often essential for developing the deductive, analogical, and inferential skills needed to acquire expertise for handling routine problems as well as the adaptability and creativity required for coping with less foreseen situations and, more generally, for learning.

In fact, it is natural for humans, when faced with uncertainty, to resort to exploratory trial-and-error behavior in order to replace false beliefs and assumptions with valid frames of reference for assessing and solving problems and situations. In these learning situations, the benefits of making errors are expected to outweigh the costs. During the early stages of the U.S. space rocket program there is anecdotal evidence that scientists actually desired failures during testing phases, in much the same way that designers of complex software sometimes do, as these failures provide insights into improvements, and ultimately more effective and robust designs, that would otherwise not have been apparent.

## 1.2 Defining Human Error

The position taken here is that human error is a real phenomenon, if only for the simple fact that humans are fallible. When this fallibility, in the form of committed or omitted human actions, appears in retrospect to be linked to undesirable consequences, an attribution of human error is often made. It can be argued that the choice of the term *human error* is unfortunate as in many circumstances (some may even claim in all circumstances barring malicious behavior) the stigma that is bestowed by virtue of using this term is inappropriate and misleading.

Human error, especially in the form of unintended or mistaken actions, is very much a two-sided coin, as it has at its roots many of the same processes of attention and architectural features of memory that also enable humans to adapt, abstract, infer, and create. It is certainly not incorrect, though perhaps a bit too convenient, to explain unintended *action slips* (Section 3.1), such as the activation of an incorrect control or the selection of the wrong medication, as rational responses in contexts characterized by pressures, conflicts, ambiguities, and

fatigue. In reality, it is human fallibility, in all its guises, that infiltrates these contexts. It is the task of human factors researchers and practitioners to examine and understand this interplay between fallibility and context as humans carry out their various activities. This knowledge could then be used to predict the increased possibility for certain types of errors, ultimately enabling safer and more productive designs.

It is not easy arriving at a satisfying definition of human error. Hollnagel (1993) preferred the term *erroneous action* to human error, which he defined as "an action which fails to produce the expected result and which therefore leads to an unwanted consequence" (p. 67). This definition as well as Sheridan's (2008) definition of human error as an action that fails to meet some arbitrary implicit or explicit criterion both allude to the subjective element that definitions of human error must incorporate.

Another term often used by Hollnagel, and which is frequently used throughout this chapter, is *performance failure*. While this term also implies some form of negative outcome related to human actions, it does so with the recognition that this outcome derives mostly from the intersection of "normal" human performance variability with "normal" system variability. The implication is that a different point of intersection may have very well brought about a favorable result.

Dekker's (2005) view of errors as "*ex post facto* constructs rather than as objective, observed facts" (p. 67) is based on the accumulated evidence for the predisposition of hindsight bias (Section 1.1). Specifically, observers (including the people who may have been recent participants of the unwanted events being investigated) impose their knowledge in the form of assumptions, facts, past experiences, and future intentions to transform what was in fact inaccessible information at the time into neatly unfolding sequences of events and deterministic schemes that are capable of explaining any adverse consequence. These observer and hindsight biases presumably do not bring us any closer to understanding the experiences of the actor in the actual situation for whom there is no error—"the error only exists by virtue of the observer and his or her position on the outside of the stream of experience" (p. 66).

What seems to be indisputable, at least in current thinking, is that human error involves some form of *attribution* that is based on the circumstances surrounding the offending behavior and the expectations held by some entity concerning the corresponding actor. The entity—a supervisor, designer, work team, regulatory agency, organization, the public, or even the person whose performance was directly linked to the adverse event—decides, based on the circumstances, whether an attribution of human error is called for.

The process of attribution of error obviously will be subject to a variety of influences. These would include cultural norms that dictate, for example, the standards to which designers, managers, and operators are held to by their organizations and to which regulatory agencies and the public hold organizations. Thus, a highly experienced pilot or nuclear power plant maintenance worker would probably not be expected to omit an important

step in a check-off procedure, even if distraction at an inopportune time and poor design of the procedure were obvious culprits. But there would be less expectation that a second-year medical resident in a trauma center, thrust into a leadership role in the absence of more senior personnel, would not make an error related to the management of multiple patients with traumatic injuries. There may, however, be an attribution of error by a state regulatory agency directed at the health care organization's management stemming from the absence or poor oversight of protocols intended for preventing these highly vulnerable allocations of responsibility.

The attribution of human error thus also encompasses actions or decisions whose unwanted outcomes may occur at much later points in time or following the interjection of many other actions by other people. Even in cases where such "blunter" actions have not resulted in adverse outcomes, an entity may consider such decisions to be in error based on its belief that unwanted consequences had been fortuitously averted.

Although intentional violations of procedures are a great concern in many industries, these acts are typically excluded from attributions of human error when the actions have gone as planned. For example, violations in rigid "ultrasafe and ultraregulated systems" are often required for effectively managing work constraints (Amalberti, 2001). However, when violations result in unforeseen and potentially hazardous conditions, managers responsible for the design of and compliance with the violated procedures may attribute human error to these actions (Section 1.1).

The attribution of human error becomes more blurred when humans knowingly implement strategies in performance that will result in some degree of unwanted consequences. For example, a worker may believe an action in a particular circumstance would avert the possibility of more harmful consequences. Even if these strategies come off as intended, depending on the boundaries of acceptable outcomes established or perceived by external observers such as managers or the public, the human's actions may in fact be considered to be in error.

Accordingly, a person's ability to provide a reasonable argument for behaviors that resulted in unwanted consequences does not necessarily exonerate the person from the *attribution* of error. What of actions the person intends to commit that are normally associated with acceptable outcomes but that, due to an unusual collection of circumstances, result in adverse outcomes? These would generally not be attributed to human error except perhaps by unforgiving stakeholders who are compelled to exact blame.

## 2 UNDERSTANDING HUMAN ERROR

### 2.1 Basic Framework: Human Fallibility, Context, and Barriers

Figure 1 presents a very basic framework for understanding human error that consists of three components. The *human fallibility* component addresses fundamental sensory, cognitive, and motor limitations of humans as well as a host of other *behavioral tendencies* that predispose humans to error. The *context* component refers to situational variables that can shape, influence, force, or otherwise affect the ways in which human fallibility, in the form of normal human performance variability, can play a role in bringing about adverse consequences. This variability encompasses not only the variability that derives from fundamental sensory, cognitive, and motor considerations but also the more "deliberate and purposeful" variability that, within the context of complex system operations, gives rise to the adaptive adjustments people make (Hollnagel, 2004). Finally, the *barriers* component concerns the various ways in which human errors or performance failures can be contained.

A number of points concerning this framework should be noted. First, human error is viewed as arising primarily from some form of *interplay* between human fallibility and context. This is probably the most intuitive way for practitioners to understand how human errors come about. Interventions that minimize human dispositions to fallibility, for example, by placing fewer memory demands on the human, are helpful, but only to the extent that they do not create new contexts that,
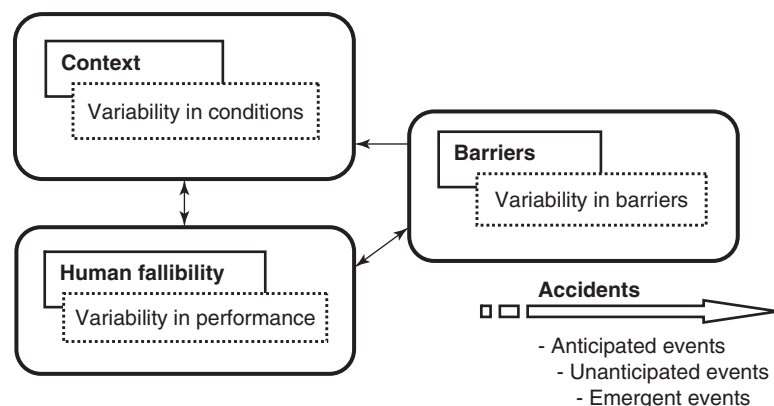


**Figure 1** Framework for understanding human error and its potential for adverse consequences.

in turn, can create new ways in which human performance variability can translate into negative outcomes. Similarly, interventions intended to reduce the error-producing potential of work contexts, for instance, by introducing new protocols for communication, could unsuspectingly produce new ways in which human fallibility can be brought to bear.

Second, many of the elements that comprise human fallibility can potentially overlap, as can many of the elements that encompass context, reflecting the interactive complexity that can be manifest among these factors. Third, because of the variability that exists in both the fallibility elements and the contextual elements, the product of their interplay will also necessarily be dynamic in nature. One consequence of this interplay is the need for anticipation, which produces human performance that is *proactive*, in addition to being *reactive*, making possible the human's ongoing adaptive responses. These responses, in turn, can alter the context that, at the same time, is experiencing its own exogenously driven variability.

From this superimposition of human performance variability on situational variability, accidents can *emerge* (Figure 1). This does not exclude the possibility for predictions of accidents based on underlying *linear* (and to some extent interactive) mechanisms, but it does dramatically alter the conceptualization of the accident process and the implications for its management.

Fourth, barriers intended to prevent the propagation of errors to adverse outcomes such as accidents could also affect the context, as well as human perceptions of the work context, and thus ultimately human performance. These interactions are often ignored or misunderstood in evaluating a system's risk potential.

In some accident models, the possibility for progressing from human error to an adverse outcome depends on how the "gaps" (the windows of opportunity for penetration) in existing barriers are aligned (Reason, 1990). Generally, the likelihood that errors will traverse these juxtaposed barriers is low, which is the reason for the much larger number of near misses that are observed compared to events with serious consequences. The avoidance or containment of or rapid recovery from accidents, including those resulting from *emerging phenomena*, may very well characterize the *resilience* of an organization (Section 6).

Finally, this framework (Figure 1) is intended to encompass various perspectives on human error that have been proposed, in particular, the human factors, cognitive engineering, and sociotechnical perspectives [Center for Chemical Process Safety (CCPS), 1994]. In the human factors perspective, error is the result of a mismatch between task demands and human mental and physical capabilities. Presumably, this perspective allows only general predictions of human error to be made. For example, cluttered displays or interfaces that impose heavy demands on working memory are likely to overload perceptual and memory processes (Section 2.2.1), possibly leading to the omission of actions or the confusion of one control with another. Guidelines that have been proposed for designing displays (Wickens et al., 2004) are offered as a means

for diminishing mismatches between demands and capabilities and thus the potential for error.

The cognitive engineering perspective, in contrast, emphasizes detailed analysis of work contexts (Section 3) coupled with analysis of the human's intentions and goals. Although both the human factors and cognitive engineering perspectives on human error are very concerned with human information processing, cognitive engineering approaches attempt to derive more detailed information about how humans acquire and represent information and how they use it to guide actions. This emphasis provides a stronger basis for linking underlying cognitive processes with the external form of the error and thus should lead to more effective classifications of human performance and human errors. As a simple illustration of the cognitive engineering perspective, Table 1 demonstrates how the same external expression of an error could derive from various underlying causes.

Sociotechnical perspectives on human error focus on the potential impact of management policies and organizational culture on shaping the contexts within which people act. These "higher order" contextual factors are capable of exacting considerable influence on the designs of workplaces, operating procedures, training programs, job aids, and communication protocols and can produce excessive workload demands by imposing multiple conflicting and shifting performance objectives and by exerting pressure to meet production goals, often at the expense of safety considerations (Section 6).

## 2.2  Human Fallibility

### 2.2.1  Human Information Processing

A fundamental basis for many human errors derives from underlying limitations and tendencies that characterize human sensory, cognitive, and motor processes (Chapters 3–5). These limitations are best understood by considering a generic model of human information processing that conceptualizes the existence of various *processing resources* for handling the flow and transformation of information (Figure 2).

According to this model, sensory information received by the body's various receptor cells gets stored in a system of sensory registers that has an enormous storage capacity. Through the process of *selective attention*, subsets of this vast collection of briefly available information become designated for further processing in an early stage of information processing known as *perception*. Here, information can become meaningful through comparison with information in *long-term memory* (LTM). This could promptly trigger some form of response or require the need for further processing in a short-term memory store referred to as *working memory* (WM).

A good deal of our conscious effort is dedicated to WM activities such as visualizing, planning, evaluating, conceptualizing, and making decisions, and much of this WM activity depends on information that can be accessed from LTM. The rehearsal of information in WM enables it to be encoded into LTM; otherwise, it decays rapidly. In addition to this time constraint, WM

**Table 1  Examples of Different Underlying Causes of Same External Error Mode**

*Situation:* A worker in a chemical processing plant closes valve B instead of nearby valve A, which is the required action as set out in the procedures. Although there are many possible causes of this error, consider the following five possible explanations.

1. The valves were close together and badly labeled. The worker was not familiar with the valves and therefore chose the wrong one.

   *Possible cause:* wrong identification compounded by lack of familiarity leading to wrong intention (once the wrong identification occurred, the worker intended to close the wrong valve).

2. The worker may have misheard instructions issued by the supervisor and thought that valve B was the required valve.

   *Possible cause:* communications failure giving rise to a mistaken intention.

3. Because of the close proximity of the valves, even though he intended to close valve A, he inadvertently operated valve B when he reached for the valves.

   *Possible cause:* correct intention but wrong execution of action.

4. The worker closed valve B very frequently as part of his everyday job. The operation of A was embedded within a long sequence of other operations that were similar to those normally associated with valve B. The worker knew that he had to close A in this case, but he was distracted by a colleague and reverted back to the strong habit of operating B.

   *Possible cause:* intrusion of a strong habit due to external distraction (correct intention but wrong execution).

5. The worker believed that valve A had to be closed. However, it was believed by the workforce that despite the operating instructions, closing B had an effect similar to closing A and in fact produced less disruption to downstream production.

   *Possible cause:* violation as a result of mistaken information and an informal company culture to concentrate on production rather than safety goals (wrong intention).

*Source*: Adapted from CCPS (1994). Copyright 1994 by the American Institute of Chemical Engineers. Reproduced by permission of AIChE.

also has relatively severe capacity constraints governing the amount of information that can be kept active. The current contention is that within WM there are separate limited-capacity storage systems for accommodating visual information presented in an analog spatial form and verbal information presented in an acoustical form as well as an attentional control system for coordinating these two storage systems. Ultimately, the results of WM–LTM analysis can lead to a response (e.g., a motor action or decision) or to the revision of one's thoughts.

This overall sequence of information processing, though depicted in Figure 2 as flowing from left to right,

in fact can assume other pathways. For example, it could be manifest in the form of an attention-WM-LTM loop if one was contemplating how to modify a work operation.

With the exception of the system of sensory registers and LTM, the processing resources in this model may require *attention*. Often thought of as *mental effort*, attention is conceptualized here as a finite and flexible endogenous energy source under conscious control whose intensity can be modulated over time. Although the human has the capability for distributing attention among the various information-processing resources, fundamental limitations in attention constrain the capacities of these resources, implying that there is only so much information that can, for example, undergo perceptual coding or WM analysis. Focusing attention on one of these resources will usually handicap, to some degree, the information-processing capabilities of the other resources.

In many situations, attention may be *focused* almost exclusively on WM, for example, during intense problem solving or when conceiving or evaluating plans. Other situations may require the need for dividing attention, which is the basis for *time sharing*. This ability is often observed in people who have learned to rapidly shift attention between tasks. Time-sharing skill may depend on having an understanding of the temporal and knowledge demands of the tasks and the possibility that one (or more) of the tasks has become automated in the sense that very little attention is needed for its performance. Various dichotomies within the information-processing system have been proposed, for example, between the visual and auditory modalities and between early (perceptual) versus later (central and response) processing (Figure 2), to account for how people are able, in time-sharing situations, to more effectively utilize their processing capacities (Wickens, 1984).

Many design considerations arise from the errors that human sensory and motor limitations can cause or contribute to. Indeed, human factors studies are often preoccupied with deriving design guidelines for minimizing such errors. Knowledge concerning human limitations in contrast sensitivity, hearing, bandwidth in motor movement, and sensing tactile feedback can be used to design visual displays, auditory alarms, manual control systems, and protective clothing (such as gloves that are worn in surgery) that are less likely to produce errors in detection and response.

Much of the focus on human error, however, is on the role that cognitive processing plays. Even seemingly simple situations involving errors in visual processing may in fact be rooted in much more complex information processing. For example, consider the following prescription medication error, which actually occurred. A physician opted to change the order for 50 mg of a leukemia drug to 25 mg by putting a line through the zero in the "50" and inserting a "2" in front of the "5." The resulting dose was perceived by the pharmacist as 250 mg and led to the death of a 14-year-old boy.

On the surface, this error can be viewed as resulting from normal human variability associated with visual processing—that is, at any given moment, the attention
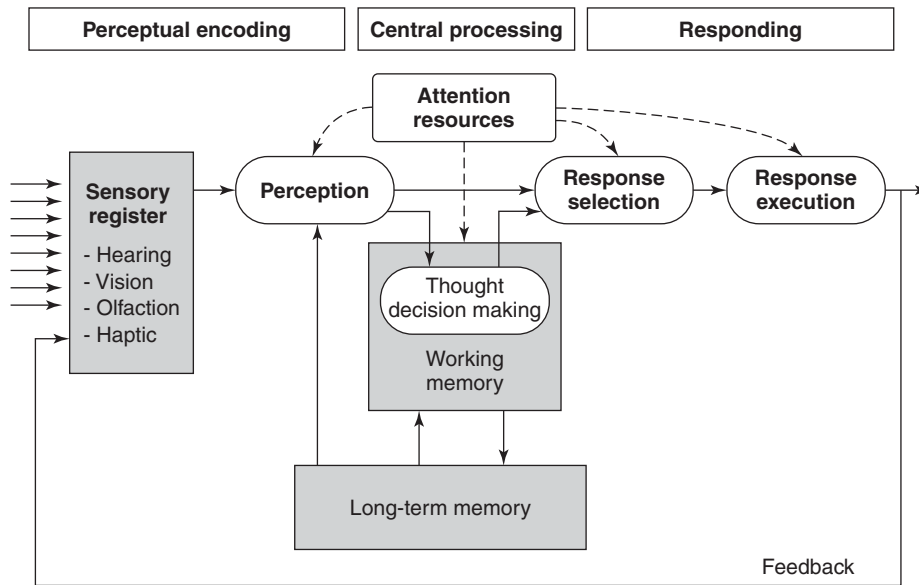
**Figure 2**   Generic model of human information processing. (Adapted from Wickens et al., 2004. Reproduced by permission of Pearson Education, Inc.)

being directed to a given stimulus is varying and at that critical moment the line through the zero was missed. However, a closer examination of the context may suggest ways in which this normal variability can be influenced, beginning with the fact that the line that was meant to indicate a cross-out was not centered but (due to normal psychomotor variability) was much closer to the right side of the circle. The cross-out at that given moment could then have easily been construed as just a badly written zero. Also, when one considers that perception relies on both *bottom-up* processing (where the stimulus pattern is decomposed into features) and *top-down* processing (where context and the expectations that are drawn from the context are used for recognition of the stimulus pattern), the possibility that a digit was crossed out may have countered expectations (i.e., it does not usually occur).

If one were to further presume that the pharmacist had a high workload (and thus diminished cognitive resources for processing the prescription) and a relative lack of experience or knowledge concerning dosage ranges for this drug, it is easy to understand how this error can come about. The progression from faulty visual processing or misinterpretation of the stimulus to adverse consequences can be put into a more complete perspective when potential barriers are considered, such as an automatic checking system that could have screened the order for a potentially harmful dosage or interactions with other drugs or a procedure that would have required the physician to rewrite any order that had been altered. However, even if these safeguards were in place, which was not the case, it is still possible that they could have been bypassed (Section 2.4).

### 2.2.2   Long-Term Memory's Role in Human Error

Long-term memory has been described as a parallel distributed architecture that is continuously being reconfigured within the brain through selective activation and inhibition of massively interconnected neuronal units (Rumelhart and McClelland, 1986). In the process of adapting to new stimuli or thoughts, the complex interactions that are produced between these neuronal units give rise to the generalizations and rules and ultimately to the knowledge that is so critical to human performance. When we consider the forms in which this knowledge is stored in LTM, we usually distinguish between the general knowledge we have about the world, referred to as *semantic memory*, and knowledge about events, referred to as *episodic memory*.

Items of information, such as visual images, sounds, and thoughts that are processed in WM at the same time and to a sufficient degree, usually become associated with each other in LTM. The ability to retrieve this information from LTM, however, will depend on the strengths of the individual items as well as the strengths of their associations with other items. Increased *frequency* and *recency* of activation are assumed to promote stronger (i.e., more stable) memory traces, which are otherwise subject to negative exponential decays.

Much of our basic knowledge about things can be thought of as being stored in the form of *semantic networks*, which are implemented within LTM through parallel distributed architectures. Other knowledge representation schemes commonly invoked in the human factors literature are *schemas* and *mental models*. Schemas typically represent knowledge organized about

a concept or topic. When they reflect processes or systems for which there are relationships between inputs and outputs that the human can mentally visualize and "experiment with" (i.e., "run," like a simulation program), the schemas are often referred to as *mental models* (Wickens et al., 2004). The organization of knowledge in LTM as schemas or mental models is also likely based on semantic networks.

The constraints associated with LTM architecture can provide many insights into human fallibility and how this fallibility can interact with situational contexts to produce errors. For example, many of the contexts within which humans operate produce what Reason (1990) has termed *cognitive underspecification*, which implies that at some point in the processing of information the specification of information may be incomplete. It may be incomplete due to perceptual processing constraints, WM constraints, LTM (i.e., knowledge) limitations, or external constraints, as when there is little information available on the medical history of a patient undergoing emergency treatment or when piping and instrumentation diagrams have not been updated.

Because the parallel associative networks in our brain have the ability to recall both items of information and patterns (i.e., associations) of information based on *partial matching* of this incomplete input information with the contents of memory, the limitations associated with cognitively underspecified information can be overcome, but at a risk. Specifically, LTM can retrieve items of information that provide a match to the inputs, and these retrieved items of information may enable, by virtue of LTM's associative structure, an entire

rule or idea to become activated. Even if this rule is not appropriate for the particular situation, if the pattern characterizing this rule in LTM is sufficiently *similar* to the input pattern of information, it may still get triggered, possibly resulting in a *mistaken action* (Section 2.2.5).

## 2.2.3 Information Processing and Decision-Making Errors

Human decision making, particularly the kind that takes place in complex dynamic environments without the luxury of extended time and other resources needed for accommodating *normative prescriptive models* (Chapter 7), is an activity fraught with fallibility. As illustrated in Figure 3, this fallibility can arise from a number of information-processing considerations (Figure 2). For example, if the information the human opts to select for examination in WM is fuzzy or incomplete, whether it be facts, rules, or schemas residing in LTM, or information available from external sources such as equipment monitors, computer databases, or other people, intensive interpretation or integration of this information in WM may be needed. Unfortunately, WM is relatively fragile as it is subject to both time and capacity constraints (Section 2.2.2).

Decision-making situations that involve the consideration of different hypotheses as a basis for performing some action also can place heavy demands on WM. Initially, these demands derive from the process of *generating* hypotheses, which is highly dependent on information that can be retrieved from LTM. The *evaluation* of hypotheses in WM may then entail searching
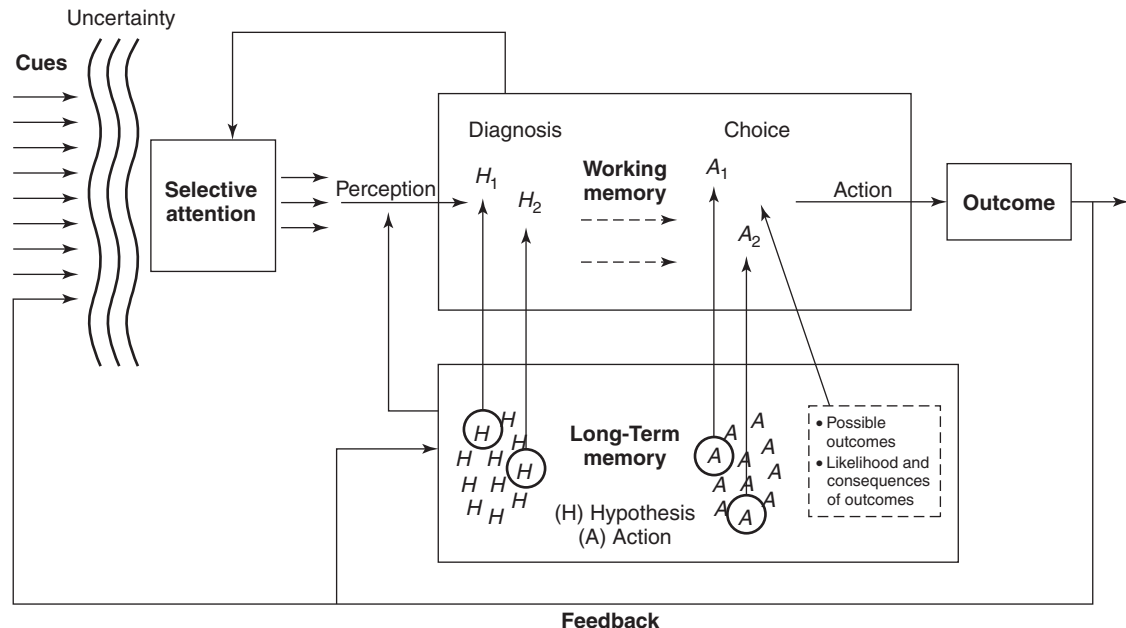


**Figure 3** Information-processing model of decision making. (Adapted from Wickens et al., 2004. Reproduced by permission of Pearson Education, Inc.)

for additional information, which would further increase the load on WM. Although any hypothesis for which adequate support is found can become the basis for an action, there may be a number of possible actions associated with this hypothesis, and they also would need to be retrieved from LTM in order to be evaluated in WM. Finally, the possible outcomes associated with each action, the estimates of the likelihoods of these outcomes, and the negative and positive implications of these outcomes would also require retrieval from LTM for evaluation in WM (Figure 3).

From an information-processing perspective, there are numerous factors that could constrain this decision-making process, particularly those that could influence the amount or quality of information brought into WM and the retrieval of information from LTM. These constraints often lead to shortcuts in decision making, such as *satisficing* (Simon, 1966), whereby people adopt strategies for sampling information that they perceive to be most relevant and opt for choices that appear to them to be good enough for their purposes.

In general, the human's natural tendency to minimize cognitive effort (Section 2.2.5) opens the door to a wide variety of shortcuts or heuristics (Tversky and Kahneman, 1974). These tendencies are usually effective in negotiating environmental complexity but under the right coincidence of circumstances can bias the human toward ineffective choices or actions that can become designated as errors. For example, with respect to the cues of information that we perceive, there is a tendency to overweight cues occurring earlier rather than later in time or that change over time. Often, the information that is acquired early on can influence the shaping of an initial hypothesis; this could, in turn, influence the interpretation of the information that is subsequently acquired. In trying to make sense of this information, WM will only allow for a limited number of possible hypotheses, actions, or outcomes of actions to be evaluated at any time. Moreover, LTM architecture will accommodate these limitations by making information that has been considered more frequently or recently more readily available (the *availability heuristic*) and by enabling its partial-matching capabilities to classify cues as more representative of a hypothesis than may be warranted.

There are many other heuristics (Wickens et al., 2004) that are capable of becoming invoked by virtue of the human's fundamental tendency to conserve cognitive effort. These include *confirmation bias* (the tendency to consider confirming and not disconfirming evidence when evaluating hypotheses); *cognitive fixation* (remaining fixated on initial hypotheses and underutilizing subsequent information); and the tendency to judge an "event" as likely if its features are *representative* of that event (e.g., judging a person as having a particular occupation based on the person's appearance or political ideology, even though the likelihood of having that occupation is extremely low).

Similarly, the human is often found to be biased in matters related to making statistical or probabilistic assessments. One important type of statistical assessment is the ability to recognize the existence of *covariation* between events. This ability can prove essential in ensuring desired outcomes (and avoiding adverse ones), as it provides humans with the capability to control the present and predict the future by virtue of explaining past events (Alloy and Tabachnik, 1984). While debates continue regarding human capabilities at such assessments, there is ample evidence that, when estimating the degree to which two events are correlated, people overemphasize instances in which the events co-occurred and disregard cases in which one event occurred but not the other, leading to overestimation of the relationship between the two events (Peterson and Beach, 1967). Top-down expectancies or preconceptions by people can alter the detection of covariation by making it unlikely that it will be detected if the variables are not expected to be related. Conversely, when relationships between variables are expected, their covariation can be given undue weight at the expense of overlooking or discounting disconfirming evidence, especially when people believe there to be a cause–effect relationship between the variables. In fact, this tendency by people can be viewed as one of the many manifestations of the confirmation bias (Nickerson, 1998).

People also typically overestimate the probability of the joint occurrence of independent events (relative to the objective or estimated probabilities of the individual events) and underestimate the probability that at least one of them will occur (Peterson and Beach, 1967; Tversky and Kahneman, 1974). These tendencies have a number of practical implications, especially when estimation of the probability of success depends on the conjunction of two or more events. For example, in the execution of sequential stepwise procedures, it can lead to overestimation of the probability that the *entire* operation will be performed successfully or completed by a specified time and to underestimation that *some* problem will be encountered in executing the procedure (Nickerson, 2004).

While the human's lack of knowledge of certain concepts and principles that are fundamental to probability theory may explain a few of the findings in this area, limitations in information-processing capacities coupled with overreliance on heuristics that work well in many but not all contexts is probably at the root of many of these human tendencies. Generally, however, one should be cautious when providing explanations of human judgments and behaviors on the basis of cognitive biases. To exclude the possibility that a human's situational assessments are in fact rational, a sound understanding of the specific context is required (Fraser et al., 1992).

## 2.2.4 Levels of Human Performance and Dispositions for Errors

Rasmussen (1986) has described fundamentally different approaches that humans take to processing information based on distinctions between skill-based, rule-based, and knowledge-based (SRK) levels of performance. The distinctions that underlie this *SRK framework* have been found to be particularly appealing for analyzing and predicting different types of human errors.

Activities performed at the *skill-based level* are highly practiced routines that require little conscious

attention. Following an intention for action, which could originate in WM or from environmental cues, the responses associated with the intended activity are so well integrated with the activity's sensory features that they are elicited in the form of highly automatic routines that are "hardwired" to the human's motor response system, bypassing WM (Figure 2).

At the *rule-based level* of performance, use is made of rules that have been established in LTM based on past experiences. WM is now a factor, as rules (of the if–then type) or schemas may be brought into play following the assessment of a situation or problem. More attention by the human is thus required at this level of performance, and the partial matching characteristics of LTM can prove critical.

When stored rules are not effective, as is often the case when new or challenging problems arise, the human is usually forced to devise plans that involve exploring and testing hypotheses and must continuously refine the results of these efforts into a mental model or representation that can provide a satisfactory solution. At this *knowledge-based level* of performance heavy demands on information-processing resources are exacted, especially on WM, and performance is vulnerable to LTM's architectural constraints to the extent that WM is dependent on LTM for problem solving.

In reality, many of the meaningful tasks that people perform represent mixtures of SRK levels of performance. Although performance at the skill-based level results in a significant economy in cognitive effort, the reduction in resources of attention comes at a risk. For example, consider an alternative task that contains features similar to those of an intended task. If the alternative activity is frequently performed and therefore associated with skill-based automatic response patterns, all that is needed is a context that can distract the human from the intention and allow the human to be "captured" by the alternative (incorrect) task. This situation represents example 4 in Table 1 in the case of an inadvertent closure of a valve.

In other situations, the *capture* by a skill-based routine may result in the exclusion of an activity. For example, suppose that task A is performed infrequently and task B is performed routinely at the skill-based level. If the initial steps are identical for both tasks but task A requires an additional step, this step is likely to be omitted during execution of this task. Untimely interruptions are often the basis for such omissions at the skill-based level of performance. In some circumstances, interruptions or moments of inactivity during skill-based routines may instigate thinking about where one is in the sequence of steps. By directing attention to routines that are not designed to be examined, steps could be performed out of sequence (reversal errors) or be repeated (Reason, 1990).

Many of the errors that occur at the rule-based level involve inappropriate matching of either external cues or internally generated information with the conditional components of rules stored in LTM. Conditional components of rules that have been satisfied on a frequent basis or that appear to closely match prevailing conditions are more likely to be activated. Generally, the

prediction of errors at this level of performance would require knowing what rules the human might consider. This, in turn, would require having detailed knowledge not only about the task but also about the process (e.g., training or experience) by which the person acquired rule-based knowledge.

When applying rules, a mistake that can easily occur is the *misapplication of a rule* with proven success (Reason, 1990). This type of mistake often occurs when *first exceptions* are encountered. Consider the case of an endoscopist who relies on indirect visual information when performing a colonoscopy. Based on past experiences and available knowledge, the sighting of an anatomical landmark during the performance of this procedure may be interpreted to mean that the instrument is situated at a particular location within the colon, when in fact the presence of an anatomical deformity in this patient may render the physician's interpretation as incorrect (Cao and Milgram, 2000). These first exception errors often result in the decomposition of general rules into more specific rule forms and reflect the acquisition of expertise. General rules, however, given their increased likelihood of encounter, usually have higher activation levels in LTM, and under contextual conditions involving high workload and time constraints will be the ones more likely to be invoked.

At the knowledge-based level of performance, needed associations or schemas are not readily available in LTM. Formulating solutions to problems or situations therefore will require intensive WM activity, implying a much greater repertory of behavioral responses and corresponding expressions of error. Contextual factors that include task characteristics and personal factors that include emotional state, risk attitude, and confidence in intuitive abilities can play a significant role in shaping the error modes, making these types of errors much harder to predict. It is at this level of performance that we observe undue weights given to perceptually salient cues or early data, confirmation bias, use of the availability and representative heuristics (especially for assessing relationships between causes and effects), underestimation and overestimation of the likelihood of events in response to observed data, *vagabonding* (darting from issue to issue, often not even realizing that issues are being revisited), and *encysting* (overattention to a few details at the expense of other, perhaps more relevant information).

## 2.2.5 Tendency to Minimize Cognitive Effort

The tendency for the human to minimize cognitive effort is a way of partly explaining shortcuts people unintentionally take in their mental processing, including their use of heuristics. It also explains why many people, especially in the course of their work activities, do not adopt various aiding devices intended to support their activities (Sharit, 2003).

A classic manifestation of this tendency is the reluctance to invest mental resources to peruse service manuals, technical publications, or other forms of documentation, whether printed or computer based, unless left with no option. More palatable options

generally consist of trial-and-error assembly or use of a device or asking a co-worker for help. For example, residents performing morning rounds in intensive care units (ICUs) will often find it easier, especially when under time pressure to process a relatively large number of patients, to obtain needed information concerning patient status from ICU nurses rather than comb through various sources of information for the purpose of constructing mental models of patient problems. Similarly, a mechanic who encounters difficulty when trying to execute an assembly strategy may be inclined to ask a fellow mechanic for assistance, especially if there are a number of impending tasks to be performed.

In contrast to the automatic processing mode that largely characterizes efficient skill-based performance, performance that requires a significant outlay of attention is effortful and potentially exhaustive of information-processing resources. From an evolutionary standpoint, this type of processing leaves us vulnerable: Being consumed with activities requiring *focused* or *divided attention* leaves little capacity for negotiating other environmental inputs that can prove threatening. In practical work situations, especially in contexts with changing conditions and objectives, this type of processing can disable or weaken performance that is based on either *feedforward control*, whereby the human devises strategies or plans for controlling a work process, or *feedback control*, whereby the human monitors and assesses conditions and adjusts or adapts performance according to system outputs.

Most work and, for that matter, everyday situations are, however, characterized by sufficient regularity and predictability to warrant the use of shortcuts in mental processing. In fact, the argument can be made that at any given time the human's normal work performance reflects a subconscious attempt to optimally balance use of these efficient shortcuts with more capacity-demanding mental processing—what Hollnagel (2004) has referred to as the "efficiency-thoroughness trade-off" (ETTO). Because any protective function can fail, it should not be surprising that conditions and events can become aligned in ways that allow shortcuts, heuristics, or expectation-driven behaviors to lead to negative outcomes. Although such outcomes may be due to the momentary existence of conditions that were not favorable to the particular type of ETTO that was manifest, and thus reflect *normal* performance variability, they still derive in part from human fallibility related to the tendency to minimize cognitive effort.

Some typical ETTO rules noted by Hollnagel (2004, p. 154) that characterize how people (or groups of people) cope with particular work situations are as follows:

- *Looks ok*. The worker resorts to a quick judgment rather than a more thorough check of the status and conditions but takes responsibility for the assessment.
- *Not really important*. Even though there are cues to warrant a closer examination of the work issue, the consequences of not dealing with the issue are rationalized as not being that serious.

- *Normally ok, no need to check it now*. The tendency to defer closer examination of an issue is often traded off with the riskier decision resulting from internal or external pressure to meet production goals.
- *It will be checked by someone else later/it has been checked by someone else earlier*. Time pressure and impending deadlines often lead to a lowered criterion for the assumption that someone else will take care or has taken care of the issue.
- *Insufficient time or resources; will do it later*. The perception that there is insufficient time or resources to perform certain activities can create the tendency to minimize the importance or urgency to complete those activities and increase the importance of the activities in which one is currently engaged.
- *It worked the last time around; don't worry, it's perfectly safe and nothing will happen*. Referencing anecdotal evidence, resorting to wishful thinking, and referring to authority or experience rather than facts are all ways of averting more time- and resource-consuming activities that involve checks and closer examination of work processes.

### 2.2.6 Other Aspects of Human Fallibility

There are many facets to human fallibility, and all have the potential to contribute to human error. Peters and Peters (2006) refer to these attributes as "behavioral vectors" and suggest that "the overestimation of human capability (to adapt) and lack of meaningful consideration of individual differences is a prime cause of undesired human error" (p. 47).

One class of individual differences that has not been given sufficient attention with regard to its ability to influence the possibility for human error is personality traits. For example, in many scenarios that involve hand-offs of work operations across shifts, it is essential that the incoming worker receive all pertinent information regarding the work activities that will be inherited. An incoming worker with a passive or submissive personality, however, may be reluctant to interrupt, interrogate, or question the outgoing worker concerning the information that is being communicated or to actively pursue information from that person, especially if that worker is perceived to have an aggressive personality or assumes a higher job status. These situations are more pervasive than one might expect, and whether they involve maintenance personnel in process control industries or medical providers in hospitals, the end result can be the same: The incoming worker may develop an incomplete or incorrect mental model of the problem. This, in turn, could lead to *false assumptions*, for example, about how an assembly procedure may need to be completed or how a new patient arrival into the ICU should be managed.

Personality traits that reflect dispositions toward confidence, conscientiousness, and perseverance could

also influence the possibility for errors. Overconfidence in particular can lead to risk-taking behaviors and has been implicated as a contributory factor in a number of accidents. Similarly, people often can be characterized in terms of having a propensity for taking risks (risk-prone behavior), avoiding risks (risk-averse behavior), or being risk neutral (Clemens, 1996). As implied in Section 2.2.5, these behavioral propensities can impact the criterion by which ETTO rules become invoked.

Another important type of fallibility concerns the human's vulnerability to sleep deprivation and fatigue. These physiological states can often be induced by work conditions and have aroused media attention as possible contributory factors in several high-profile accidents. In fact, in the maritime and commercial aviation industries, conditions of sleep deprivation and fatigue are often attributed to company or regulatory agency rules governing hours of operation and rest time. The effects of fatigue on human performance may be to regress skilled performers to the level of unskilled performers (CCPS, 1994) through widespread degradation of abilities that include decision making and judgment, memory, reaction time, and vigilance. The National Aeronautics and Space Administration (NASA) has determined that about 20% of incidents reported to its Aviation Safety Reporting System (Section 5.4.1), which asks pilots to report problems anonymously, are fatigue related (Kaye, 1999). On numerous occasions pilots have been found to fall asleep at the controls, although they usually wake up in time to make the landing.

An aspect of human fallibility with important implications for human error is *situation awareness* (Chapter 19), which refers to a person's understanding or mental model of the immediate environment (Endsley, 1995). Presumably, any factor that could disrupt a human's ability to acquire or perceive relevant data concerning the elements in the environment, or compromise one's ability to understand the importance of that data and relate them to events that may be unfolding in the near future, can degrade situation awareness. Comprehending the importance of the various types of information in the environment also implies the need for temporal awareness—the need to be aware of how much time tasks require and how much time is available for their performance (Grosjean and Terrier, 1999).

Many factors related to human fallibility and context can potentially influence situation awareness. Increased knowledge (perhaps through training) or expertise (through experience) should allow for better overall assessments of situations, especially under contextual conditions of high workload and time constraints, by enabling elements of the problem and their relationships to be identified and considered in ways that would be difficult for those who are less familiar with the problem. In contrast, poor display designs that make integration of data difficult can easily impair the process of assessing situations. In operations involving teamwork, situation awareness can become disrupted by virtue of the confusion created by the presence of too many persons being involved in activities.

Human limitations in sensory processes and motor movement (Chapters 3, 4) can also contribute to unintended or inadequate outcomes that are often attributed to human error. Because sensory, motor, and cognitive abilities tend to decline with age (Chapter 52), there is the inclination to associate aging with an increased likelihood of human error. However, the literature on aging and work performance is somewhat shaky on this subject (Czaja and Sharit, 2009), and we know that many factors can counteract or compensate for the effects of these declines. Examples of such compensatory factors include the availability of environmental support in the form of memory and other aiding devices; the provision of favorable ergonomic work conditions such as increased illumination levels; continued practice on job activities that are frequently encountered; and the use of knowledge gained from experience to devise more efficient work strategies. The fact that older people usually are more conservative in their estimations of risk, either because of awareness of their physiological declines or as a result of their knowledge accumulated from experience, also tends to mitigate the propensity for their actions to produce adverse outcomes. Declines with age in the speed of cognitive processing, however, suggest that despite such compensatory abilities, older individuals are generally not suitable for work activities that rely heavily on fundamental information-processing abilities.

Finally, the human's vulnerability to a number of affective factors can corrupt human information-processing capabilities and thus predispose the human to error. Personal crises could lead to distractions, and emotionally loaded information can lead to the substitution of relevant job-related information with "information trash." Similarly, a human's susceptibility to panic reactions and fear can impair information-processing activities critical to human performance. Conversely, the tendency to inhibit emotional responses during emergencies can contribute to effective team communication and an increased likelihood of preventing serious accidents.

## 2.3 Context

Human actions are embedded in contexts and can only be described meaningfully in reference to the details of the context that accompanied and influenced them (Dekker, 2005). The attribution and expression of human error will thus depend on the context in which task activities occur.

The notion of a context is not easy to define. Commonly encountered alternative expressions include scenario, situation, situational context, contextual features, contextual factors, and contextual dynamics. Building on a definition of context proposed by Dey (2001) in the domain of context-aware computer applications, *context* is defined as any information that can be used to characterize the situation of a person, place, or object as well as the dynamic interactions among these entities. This definition of context also encompasses information concerning how situations are changing and the human's responses to these situations.

Table 2 lists some representative contextual factors capable of influencing human performance and thus contributing to human errors and violations. Because many of these contextual factors can be described

**Table 2  Contextual Factors Capable of Influencing Human Performance**

| *Attributes of Production Processes* | *Equipment/Interface Design* |
|---|---|
| Degree to which processes are understood | Workspace layout and design |
| Degree to which failed components can be isolated | Personnel protective equipment |
| Degree to which personnel are specialized | Communications equipment |
| Degree to which materials and tools can be substituted | Tool design |
| | Location/access to tools |
| Number of control parameters and interactions among them | Labeling of equipment and supplies |
| | Use of display design principles |
| Degree to which system interdependencies are well defined | Use of control design principles |
| | Design of menus |
| Degree to which system feedback is clear and identifiable | Availability and design of help systems |
| | Availability and design of job aids |
| Degree of slack possible in supplies and equipment | Design of alarms and warnings |
| Degree to which production processes are invariant | Design of voice recognition systems |
| | Demands on memory |

*Work Environment/Work Schedule*

Noise and lighting
Thermal conditions
Vibration and atmospheric conditions
Time constraints
Perceived danger or risks
Interrup
tions and distractions
Suddenness of onset of events
Novel and unanticipated events
Good housekeeping
Work hours and rest breaks
Shift rotation and circadian disruptions

*Job Aids and Procedures*

Designed using task analysis
Instructions are clear and unambiguous
Level of description is adequate
Specification of entry/exit conditions
Instruction is available on their use
Operator feedback on their design
Updated when needed without adding excessive complexity
Capability for referencing procedures during work operations

*Training*

Training in identifying hazardous conditions
Training individuals and teams in using new technologies
Practice with unfamiliar situations
Training on using emergency procedures
Simulator training
Training in interacting with automation
Use of just-in-time training
Ensuring workers have adequate supportive information

*Organization and Social Factors*

Teamwork and communication
Clarity of responsibilities
Clarity in safety–productivity priorities
Authority and leadership
Feedback channels from workers on procedures and policies
Safety culture
Absence of culture of blame and retribution
Management commitment to safety and organizational learning

*Higher Order Factors*

Social, political, and economic factors
Regulatory agency factors

at much greater levels of detail, for any particular domain of application practitioners and analysts would need to determine the appropriate level of contextual analysis.

The presumption is that higher order factors such as sociopolitical or government regulatory factors can influence or shape organizational factors. Organizations, in turn, are assumed to be capable of influencing contextual factors that are more directly linked to human performance. Contexts ultimately derive from the characterization of these factors and the interactions among them. Analysis of the interplay of human fallibility and context as a basis for understanding human error (Section 2.1) will be beneficial to the extent that relevant contextual factors can be identified and analyzed in detail.

A number of quantitative approaches to human reliability analysis (Section 4) employ concepts that are related to context. For example, several of these approaches use *performance-shaping factors* (PSFs) or *influencing factors* (IFs) either to modify the probability estimate assigned to an activity the human performed in error or as the basis for the estimation of that error. These approaches to adjusting or estimating human error probabilities generally assume *additive effects* of PSFs on human performance rather than interactive effects.

Implicit to the concept of a context, however, is the *interactive complexity* among contextual factors with regard to their potential for influencing the reliability of human performance. In this regard, a sociotechnical approach to assessing human reliability

known as STAHR (Phillips et al., 1990) is somewhat more consistent with the concept of a context. STAHR utilizes a hierarchical network of *influence diagrams* to represent the effects of direct influences on human error, such as time pressure and quality of training, as well as the effects of less direct influences, such as organizational and policy issues, which project their influences through the more direct factors.

While STAHR imposes a hierarchical constraint on influences, the dynamic and interactive complexity that underlies the concept of context, in theory, imposes no such constraint. Influences thus could be represented as an *unconstrained network*. Although it may be exceedingly difficult to generate quantitative estimates of human error from such a conceptualization of context, it can still serve as a qualitative tool for the analysis of work contexts and for the prediction of the *possibility* of human errors. To ensure that a manageable number of meaningful contexts are exposed, a defined set of influence mechanisms would need to be imposed on this network structure that could assess the extent to which a contextual factor is present (i.e., the level of activation of a *network node* within the network); the extent to which the factor can influence and be influenced by other factors (i.e., the level of activation of a *network arc*, its direction, and whether the arc's effect is excitatory or inhibitory); and the temporal characteristics underlying these influences.

As the human becomes engaged in setting goals, planning, assessing the situation, and carrying out activities, the possibility for human performance failures or errors would then depend on the interplay—the mutually dependent coupling—between human fallibility, as broadly defined in Section 2.2, and context, conceptualized as a dynamic unconstrained network of contextual factors. In terms of this interplay, the context would influence not only the final external manifestation of the human failure (e.g., activation of the wrong control, performing a soldering operation incorrectly) but also all the precursor conditions (e.g., communication of incorrect or ambiguous information to a co-worker, making an incorrect diagnosis) that could lead to these external error modes.

Knowledge of the existence of these precursor or intermediate states may in fact be of greater interest to organizations as it can be diagnostic of numerous design and process deficiencies and thus provide the basis for an organization's ability to learn (Section 5.4). Moreover, intermediate conditions, such as communication failures, may not lead to observable manifestations of error and adverse consequences only because of barriers, either designed in or a result of fortuitous circumstances, which may have been in place (Section 2.4), but which cannot always be relied on. Many highly instructive but much less formalized accounts of the interplay between human fallibility and context can be found in a number of the reconstructed accidents presented by Stephen Casey (1993, 2006).

The unconstrained network conceptualization of context is similar to the *nonhierarchical* representation of context underlying Hollnagel's (1998) method for human reliability analysis known as CREAM

(Section 4.10). This approach distinguishes between *phenotypes*, which are the external manifestation of erroneous human actions or "error modes," and *genotypes*, which are the factors that can "cause" these failures. Only a limited set of phenotypes are considered, such as an action that is performed too late, is in the wrong direction, is applied to the wrong object, or is omitted. In contrast, a much larger number of genotypes, the possible "causes" of these external "error modes," is proposed, which are categorized according to whether they are person related, technology related, or organization related.

In this scheme, an influence from a factor associated with one category to one or more factors associated with other categories gives rise to "antecedent-consequent" links that reflect cause–effect relationships. The consequents in these links can then serve as the antecedents of yet other categories of factors. The critical analysis lies in the examination of the paths formed by these antecedent-consequent links. The phenotypes are either the starting point in this analytical process, for example, in a *retrospective analysis* of an adverse event, or the endpoint in this analysis, for example, in the *prospective prediction* of the possibility for erroneous actions. Further details related to these concepts are presented in Section 3.2.

Just as contextual factors can be resolved downward to more refined levels of detail, the possibility also exists for describing larger scale work domain contexts that are likewise capable of bringing about adverse outcomes. In this regard, the views of Perrow (1999), which constitute a *system theory of accidents*, have received considerable attention. According to Perrow, the structural analysis of any system, whether technological, social, or political, reveals two loosely related concepts or dimensions: *interactive complexity* and *coupling*. These dimensions have their own characteristic sets of attributes that govern the potential for system accidents and human recovery of these events. Some of the contextual factors corresponding to these attributes are listed in Table 2 under the category "Attributes of Production Processes."

The dimension of interactive complexity can be categorized as either *complex* or *linear* and applies to all possible system components, including people, materials, procedures, equipment, design, and the environment. The relatively stronger presence of features such as increased interconnectivity of subsystems, the potential for unintended or unfamiliar feedback loops, the existence of multiple and interacting controls (which can be administrative as well as technological), the presence of information that tends to be more indirect and incomplete, and the inability to easily substitute people in task activities all serve to predispose systems toward being complex as opposed to linear. Complex interactions are more likely to be produced by complex systems than linear systems. Because these interactions tend to be less perceptible and comprehensible, the human's responses to problems that occur in complex systems can further increase the system's interactive complexity.

Systems also can be characterized by their degree of coupling. Tightly coupled systems are much less tolerant of delays in system processes than are loosely coupled

systems and are much more invariant to materials and operational sequences. Although each type of system has both advantages and disadvantages, loosely coupled systems have greater *slack*, which enables them to more easily absorb the variability of system demands. This attribute provides more opportunities for recovery from events with potentially adverse consequences, often through creative, flexible, and adaptive responses by people. To compensate for the fewer opportunities for recovery that are provided by tightly coupled systems, these systems generally require more built-in safety devices and redundancy than do loosely coupled systems.

Because Perrow's account of technological accidents focuses on the properties of systems themselves rather than human error associated with design, operation, or management of these systems, there has been criticism that his model marginalizes factors at the root of technological accidents (Evan and Manion, 2002). These criticisms, however, do not preclude the possibility of augmenting Perrow's model with additional perspectives on system processes that could endow the model with the capability for providing a reasonably compelling basis for how normal human variability in performance can predispose a system to adverse consequences.

Finally, a contextual factor that can have an especially powerful effect on predisposing the human to error during task performance is *stress*, due to the variety of ways that this phenomenon can influence human fallibility. For example, under stress people tend to become more reluctant to make an immediate decision; seek confirming evidence and disregard disconfirming evidence; become less able to recognize all the alternatives that are available for consideration; offer explanations based on a single global cause rather than a combination of causes; and take greater risks when operating in a group (Kontogiannis and Lucas, 1990).

### 2.4  Barriers

Barriers are entities that are capable of preventing errors or potentially hazardous events from taking place or, if these events manage to occur, can lessen the impact of their consequences. As such, they represent a key construct in the analysis of accidents and in the design of accident prevention systems.

The consideration of barriers was part of the Management Oversight and Risk Tree (MORT) program that was developed for the analysis of accidents and safety programs (Johnson, 1980; Trost and Nertney, 1985; Gertman and Blackman, 1994). MORT relies on a number of tree diagrams to examine factors such as lines of responsibility, barriers toward unwanted energy, and management factors. Its strategies for the elimination of system hazards, in order of importance, largely reflect the use of the following types of barriers: the elimination through design; installation of appropriate safety devices; installation of warning devices; and the use of special procedures. Distinctions between the different purposes of barriers (prevention, control, and minimization of consequences) and types of barriers (physical, equipment design, warning devices,

procedures, knowledge and skills, and supervision) are also proposed within the MORT program.

Human error and barriers are linked in a number of ways. One way in which they are connected relates to whether human actions are capable of becoming classified as human errors. Human actions that fail to result in adverse consequences due to the barriers that were in place may not be conferred with the attribution of human error, even if these actions were capable of generating hazardous conditions. They might instead, at best, be designated as near misses (Section 5.4). If analysts failed to select such actions when conducting human reliability analysis (Section 4.1), the contribution of human–system interactions to system risks could be greatly underestimated as these barriers could fail in ways that were not anticipated.

A second important connection between barriers and human error is that many barriers depend on some type of human intervention, whether it be in their detection or interpretation. Consequently, the presence of that barrier may contribute to defining a context that predisposes the human to commit actions that can produce hazardous conditions or accidents (Section 2.1). Similarly, barriers that allow for their modification, such as turning off alarms, can result in work contexts with hidden dangers that, when suddenly exposed, can define new work contexts with increased human error potential. In some cases, the introduction of a barrier may so thoroughly disturb the nature of work that many new and unanticipated forms of human error can arise (as exemplified in Section 5.1.1).

A third and often overlooked connection between barriers and human error concerns how the perception of barriers, such as intelligent sensing systems and corrective devices, may alter human performance. This connection is based in part on characterizations of human fallibility in terms of *risk attitude*, where individuals who are risk prone or even risk neutral, may be more willing to take risks when they perceive barriers to be in place. Adjusting risk-taking behavior to maintain a constant level of risk is in line with *risk-homeostasis theory* (Wilde, 1982). These adjustments presume that humans are reasonably good at estimating the magnitude of risk, which generally does not appear to be the case. A disturbing implication of this theory is the possibility that some interventions by organizations directed at improving the safety climate (Section 6) could instead result in work cultures that promote attitudes that are nonconducive to safe operations. The real danger of these behaviors is that they can establish new contexts that the barriers were not designed to prevent.

#### 2.4.1  Classification of Barrier Systems

Hollnagel (2004) has proposed a classification of barriers that, for our purposes, can serve to highlight the link between human error and barriers that can arise by virtue of human interaction with the barrier system. In his approach, barrier systems are grouped into four categories: physical or material barrier systems, functional barrier systems, symbolic barrier systems, and incorporeal barrier systems. The possibility also exists

**Table 3  Barrier functions**

| Barrier Function | Example |
|---|---|
| **Barrier Functions for Physical Barrier Systems** | |
| Containing or protecting | Walls, doors, buildings, restricted physical access, railings, fences, filters, containers, tanks, valves, rectifiers |
| Prevent transporting something from the present location (e.g., release) or into the present location (penetration) | |
| Restraining or preventing movement or transportation of mass or energy | Safety belts, harnesses, fences, cages, restricted physical movements, spatial distance (gulfs, gaps) |
| Keeping together; cohesion, resilience, indestructibility | Components that do not break or fracture easily (e.g., safety glasses) |
| Separating, protecting, blocking | Crumble zones, scrubbers, filters |
| **Barrier Functions for Functional Barrier Systems** | |
| Preventing movement or action (mechanical, hard) | Locks, equipment alignment, physical interlocking, equipment match |
| Preventing movement or action (logical, soft) | Passwords, entry codes, action sequences, preconditions, physiological matching (e.g., iris, fingerprint, alcohol level) |
| Hindering or impeding actions (spatial-temporal) | Distance (too far for a single person to reach), persistence (deadman button), delays, synchronization |
| Dampening, attenuation | Active noise reduction, active suspension |
| Dissipating energy, quenching, extinguishing | Air bags, sprinklers |
| **Barrier Functions for Symbolic Barrier Systems** | |
| Countering, preventing, or thwarting actions | Coding of functions (e.g., by color, shape, spatial layout), demarcations, labels, and (static) warnings (facilitating correct actions may be as effective as countering incorrect ones) |
| Regulating actions | Instructions, procedures, precautions/conditions, dialogues |
| Indicating system status | Signs (e.g., traffic signs), signals (visual, auditory), warnings, alarms |
| Permission or authorization (or the lack thereof) | Work permit, work order |
| Communication, interpersonal dependency | Clearance, approval (on-line or off-line) in the sense that the lack of clearance, etc., is a barrier |
| **Barrier Functions for Incorporeal Barrier Systems** | |
| Complying, conforming to | Self-restraint, ethical norms, morals, social or group pressure |
| Prescribing: rules, laws, guidelines, prohibitions | Rules, restrictions, laws (all either conditional or unconditional) |

*Source*: Hollnagel (2004).

for barriers to consist of some composite of these types of systems. A summary of the functions associated with each of these categories of barrier systems is given in Table 3.

### 2.4.2  Paradoxical Effects of Barriers

The possibility for barriers having paradoxical effects was exemplified in a study by Koppel et al. (2005), who found that the introduction of a hospital-computerized physician order entry (CPOE) system, a type of barrier system intended to significantly reduce medication-prescribing errors, actually facilitated errors by users. In this study, errors were grouped into two categories: (1) information errors arising from the fragmentation of data and the failure to integrate information

across the various hospital information systems and (2) human–machine interface flaws that fail to adequately consider the practitioner's behaviors in response to the constraints of the hospital's organizational work structure. An example of an error related to the first category is when the physician orders new medications or modifies existing medications. If current doses are not first discontinued, the medications may actually become increased or decreased or be added on as duplicative or conflicting medication. Detection of these errors was hindered by flaws in the interface that could require 20 screens for viewing a single patient's medications.

Complex organizational systems such as hospitals can make it extremely difficult for designers to anticipate

the many contexts and associated problems that can arise from interactions with the systems that they design (Section 5.1). It may seem to make more sense to have systems such as CPOEs monitored by practitioners and other workers for their error-inducing potential rather than have designers attempt to anticipate all the contexts associated with the use of these systems. However, this imposes the added burden of ensuring that mechanisms are in place for collecting the appropriate data, communicating this information to designers, and validating that the appropriate interventions have been incorporated.

With a number of electronic information devices, the benefits of reducing or even eliminating the possibility for certain types of errors may come at the risk of exposing new windows of opportunity for errors through the alteration of existing contexts. In hospital systems, for example, the reliance on information in electronic form can disturb critical communication flows and is less likely than face-to-face communication to provide the cues and other information necessary for constructing appropriate models of patient problems.

### 2.4.3  Forcing Functions and Work Procedures

A common method often employed by designers for creating barriers to human error is through the use of *forcing functions*, which are design constraints that alert system users to their errors by blocking their actions. For example, computer-interactive systems can force the user to correct an invalid entry prior to proceeding, provide warnings about actions that are potentially error inducing, and employ self-correction algorithms that attempt to infer the user's intentions. Unfortunately, each of these methods can also be breached, depending on the context in which it is used. For example, forcing functions can initiate a process of backtracking by the user that can lead to total confusion and thus more opportunity for error (Reason, 1990), and warnings can be ignored under high workloads.

One of the most frequently used symbolic barrier systems (Table 3) in industry—the *written work procedure*—is also one that is highly vulnerable to misinterpretation, often due to a variety of latent factors. For example, the designers of these procedures may not have adequately considered the human's abilities or users' concerns for their own safety or the work contexts in which the procedure would need to be carried out (Sharit, 1998). Even if procedures are well designed, inadequate training on their execution can provoke actions that can lead to adverse consequences.

Many of the procedures designed for high-hazard operations include warnings, contingencies (information on when and how to "back out" when dangerous conditions arise during operations), and other supporting features. To avoid the recurrence of past incidents, these procedures are often frequently updated. Consequently, they grow in size and complexity to the point where they can contribute to information overload, increasing the possibility even more that their users will miss or confuse important information (Reason, 1997). In addition, procedures that disrupt the momentum of

human work operations will be especially vulnerable to violation.

### 2.4.4  Use of Redundancy for Error Detection

Redundancy in the form of cues presented in multiple modalities is a simple and very effective way of increasing a person's likelihood of detecting and correcting errors. This strategy is illustrated in the case of the ampoule-swap error in hospital operating rooms (Levy et al., 2002). Many drug solutions are contained in ampoules that do not vary much in size and shape, often contain clear liquid solutions, and have few distinguishing features. If an anesthesiologist uses the wrong ampoule to fill a syringe and inadvertently "swaps in" a risky drug such as potassium chloride, serious consequences could ensue. Contextual factors such as fatigue and distractions make it unreasonable to expect medical providers to invest the resources of attention necessary for averting these types of errors. Moreover, the use of warning signs on bins that store ampoules containing "risky solutions" are poor solutions to this problem, as they require that the human maintain *knowledge in the head*—specifically, in WM—thus making this information vulnerable to memory loss resulting from delays or distractions between retrieving the ampoule and preparing the solution. The more reliable solution that was suggested by the investigators of this study was to provide tactile cues on both the storage bins and the ampoules. For example, wrapping a rubber band around the ampoule following its removal from the bin provides an alerting cue in the form of tactile feedback prior to loading the ampoule into the syringe.

Another approach to error detection through redundancy is to have other people available for detecting errors. As with hardware components, human redundancy will usually lead to more reliable systems. However, successful human redundancy often requires that the "other people" be external to the operational situation. Consequently, they would be less likely to be subject to tendencies by people to explain away inconsistencies or evidence that contradict one's assessment of the situation and thus less likely to exhibit *cognitive fixation errors*. In a study of 99 simulated emergency scenarios involving nuclear power plant crews, Woods (1984) found that while none of the errors involving diagnosis of the system state were detected by the operators who made them, other people were able to detect a number of them. In contrast, half the errors categorized as slips (i.e., errors in execution of correct intentions) were detected by the operators who made them.

That barriers to human error based on human redundancy need not always be in place by design is often demonstrated in large-scale hospital systems. In these systems, one typically encounters an assortment of patient problem scenarios, a variety of health care services, complex flows of patient information across various media on a continual 24-h basis, and a large variability in the skill levels of health care providers who must often perform under conditions of overload and fatigue while being subjected to various administrative constraints. Fortunately, there usually exist multiple layers of redundancy in the form of alternative materials

(e.g., equipment), treatment schedules, and health care workers to thwart the serious propagation of many potential errors. Thus, despite a number of constraints that are present in hospital systems, these systems are sufficiently loosely coupled (Section 2.3) to overcome many of the risks that arise in patient care, including those that are generated by virtue of discontinuities or gaps in treatment (Cook et al., 2000).

### 2.4.5 Cognitive Strategies in Error Detection

As implied in the study by Woods (1984), humans are quite adept at detecting and correcting many of the skill-based errors they make, which is why people are often relied upon to serve as barriers. Self-correction, however, implies two conditions: that the human depart from automated processing, even if only momentarily, and that the human periodically invest attentional resources to check whether the intentions are being met and that cues are available to alert one to deviation from intention (Reason, 1990). This would apply to both slips and omissions of actions.

These error detection processes, as well as other error detection processes such as forcing functions or human redundancy, are for the most part relatively spontaneous in nature and do not require significant outlays of effort. At the knowledge-based level of performance, however, the human's error detection abilities are greatly reduced. Error detection in these more complex situations will depend on intensive cognitive processing activities such as the ability to think about possible errors that might occur, predicting the time course of multiple processes, or discovering that the wrong goal has been selected.

Human error detection and recovery at the knowledge-based level of performance may in fact represent a highly evolved form of expertise. Interestingly, whereas knowledge-based errors decrease with increased expertise, skill-based errors increase. Also, experienced workers, as compared to beginners, tend to disregard a larger number of errors that have no work-related consequences, suggesting that with expertise comes the ability to apply higher order criteria for regulating the work system, thus enabling the allocation of attention to errors to occur on a more selective basis (Amalberti, 2001).

Kontogiannis and Malakis (2009) have developed and discuss in detail a taxonomy of cognitive strategies in error detection and identification that is based on the following four stages in error detection:

- *Awareness-Based Detection*. At this stage, introspection is used to critique one's mental models in terms of their completeness, coherence, and reliability, in order to enable revisions of situational assessments to consider hidden and untested assumptions through the collection of additional data.

- *Planning-Based Detection*. These strategies include the consideration of a time scale for revising plans in the face of new evidence; balancing conflicting goals through mental simulation of the risks associated with carrying out alternative plans; regulation of plan complexity to fit

the circumstances; and relying on loosely coupled rather than integrated plans to enable greater flexibility in error detection.

- *Action-Based Detection*. These proactive strategies include running preaction and postaction checks on highly routine tasks in order to avert slips and lapses; creating barriers in the form of reminders to combat the susceptibility to interruptions and "task triggers" to combat capture errors (Section 2.2.4) when tasks need to be performed in a different way; and rehearsing tasks that may need to be carried out later on under time pressure.

- *Outcome-Based Detection*. This stage includes strategies such as examining changes in relational and temporal data patterns over time; cross-checking data to manage mismatches between expected outcomes and observed outcomes; and the use of a mental model to consider the effects of interventions by other agents.

These cognitive strategies for error detection are clearly effortful. For example, they may call for the human to engage in simultaneous belief and doubt or to forego the use of well-used rules in order to cast familiar data in new ways. However, they have important implications for error management training (Kontogiannis and Malakis, 2009) and thus constitute a potentially critical consideration in the development of highly reliable and resilient organizations (Section 6).

## 3 ERROR TAXONOMIES AND PREDICTING HUMAN ERROR

### 3.1 Classifying Human Error

Many areas of scientific investigation use classification systems or taxonomies as a means for organizing knowledge about a subject matter. The subject of human error is no exception. Taxonomies of human error can be used retrospectively to gather data on trends that point to weaknesses in design, training, and operations. They can also be used prospectively, in conjunction with detailed analyses of tasks and situational contexts, to predict possible errors.

Earlier (Section 2.3), a distinction was made between phenotypes, which are the error modes that describe the external (i.e., observable) manifestation of an erroneous action, and genotypes, which are the factors that can influence or "cause" these failures. Eight basic phenotypes, or error modes, have been defined by Hollnagel (1998):

- *Timing*: actions performed too early or too late or omitted

- *Duration*: actions that continued for too long or were terminated too early

- *Force*: actions performed with insufficient or too much force

- *Distance/magnitude*: movements taken too far or not far enough

- *Speed*: actions performed too quickly or too slowly
- *Direction*: movements in the wrong direction or of the wrong kind
- *Wrong object*: a neighboring, similar, or unrelated object used by mistake
- *Sequence*: within a series of actions, actions that were omitted, repeated, reversed in their order, or inappropriately added

When directed at highly specific tasks or operations, this kind of taxonomy can be used to characterize the various ways that a particular task can be performed incorrectly. For example, in the health care industry, the diversity of medical procedures and the variety of circumstances under which these procedures are performed may, in fact, call for highly specific error taxonomies that are derived from more general error classification systems.

For more cognitively complex tasks, it may be useful to classify "cognitive failures." One approach is to categorize these errors in cognition according to stages of information processing (Figure 2), thereby differentiating, for example, errors related to perception from errors related to failures in working memory. The characterization of performance as skill, rule, or knowledge based (Section 2.2.4) also has proven particularly useful in thinking about the ways in which cognitively based failures can arise, in light of the different kinds of information-processing activities that are presumed to be occurring at each of these levels.

Figure 4 and Tables 4 and 5 illustrate several other types of error classification systems. The flowchart in Figure 4 classifies different types of human errors that can occur under SRK levels of performance. This flowchart seeks to answer questions concerning *how* an error occurred. Similar flowcharts are provided by Rasmussen (1982) to address *why* an error occurred as well as *what* type of error occurred.

Reason's (1990) taxonomy (Table 4) also exploits the distinctions among skill, rule, and knowledge-based levels of performance, but draws attention to how error modes related to skill-based *slips* and *lapses* differ from error modes related to rule and knowledge-based *mistakes*. The taxonomy presented in Table 5 illustrates the classification of external error modes into different aspects of information processing.

## 3.2 Predicting Human Error

The use of taxonomies for the purpose of revealing patterns or tendencies related to human performance failures can provide valuable data about weaknesses in design, training, and operations. These classification schemes can also support the gathering of data for performing quantitative human error assessments, which are often required in probabilistic system risk assessments (Section 4.1) and are integral to the analysis of accidents for root causes (Chapter 38).

In addition to these benefits, taxonomies of human error, especially those that emphasize cognitive or causal factors, have predictive value as well. However, as implied in Figure 5, predicting the types of errors humans might commit under actual work conditions is a difficult undertaking. The multidimensional complexity surrounding actual work situations and the uncertainty associated with the human's goals, intentions, and attentional and affective states that unfold over time introduce many layers of guesswork into the process of establishing reliable mappings between human fallibility and situational contexts.

In 1991, Senders and Moray stated: "To understand and predict errors . . . usually requires a detailed task analysis" (p. 60). Very little has changed since then to diminish the validity of this assertion. In fact, our greater understanding of mechanisms underlying human interaction with complex systems (e.g., Woods and Hollnagel, 2005) have probably made the process of predicting human error more laborious than ever, as it should be. Expectations of shortcuts are unreasonable; error prediction by its very nature should be a tedious process and will often be influenced by the scheme selected for error classification.

As implied by Senders and Moray (1991), task analysis (TA) is an essential tool for predicting human error or performance failures. TA describes the human's involvement with a system in terms of the goals to be accomplished and all the human's activities, both physical and cognitive, that are necessary to meet these goals.

Within a TA, the analysis of human–system interactions could be performed using a variety of perspectives and methods. For example, the analyst may resort to simple models of human information processing to determine if the human is receiving sufficiently salient, clear, complete, and interpretable input; has adequate time to respond to the input with respect to being able to mentally code, classify, and resolve the information or in terms of the time the system allows for executing an action; and whether feedback is available to enable the human to determine whether the action was executed correctly and was appropriate for dealing with the goal in question. More complex information-processing schemes can also be used.

These human–system interaction descriptions could also include activity time lines; dependencies that might exist among activities; alternative plans for performing an operation; contingencies that may arise during the course of activities and options for handling these contingencies; characterizations of information flow between different subsystems; and descriptions of displays, controls, training, and interactions with other people. Depending on whether the analysis is to be applied to a process that is still in the conceptual stages, to a newly implemented process, or to an existing process, broad applications of TA techniques that may include mock-ups, walkthroughs, simulations, interviews, and direct observations may be needed to identify the relevant contextual elements.
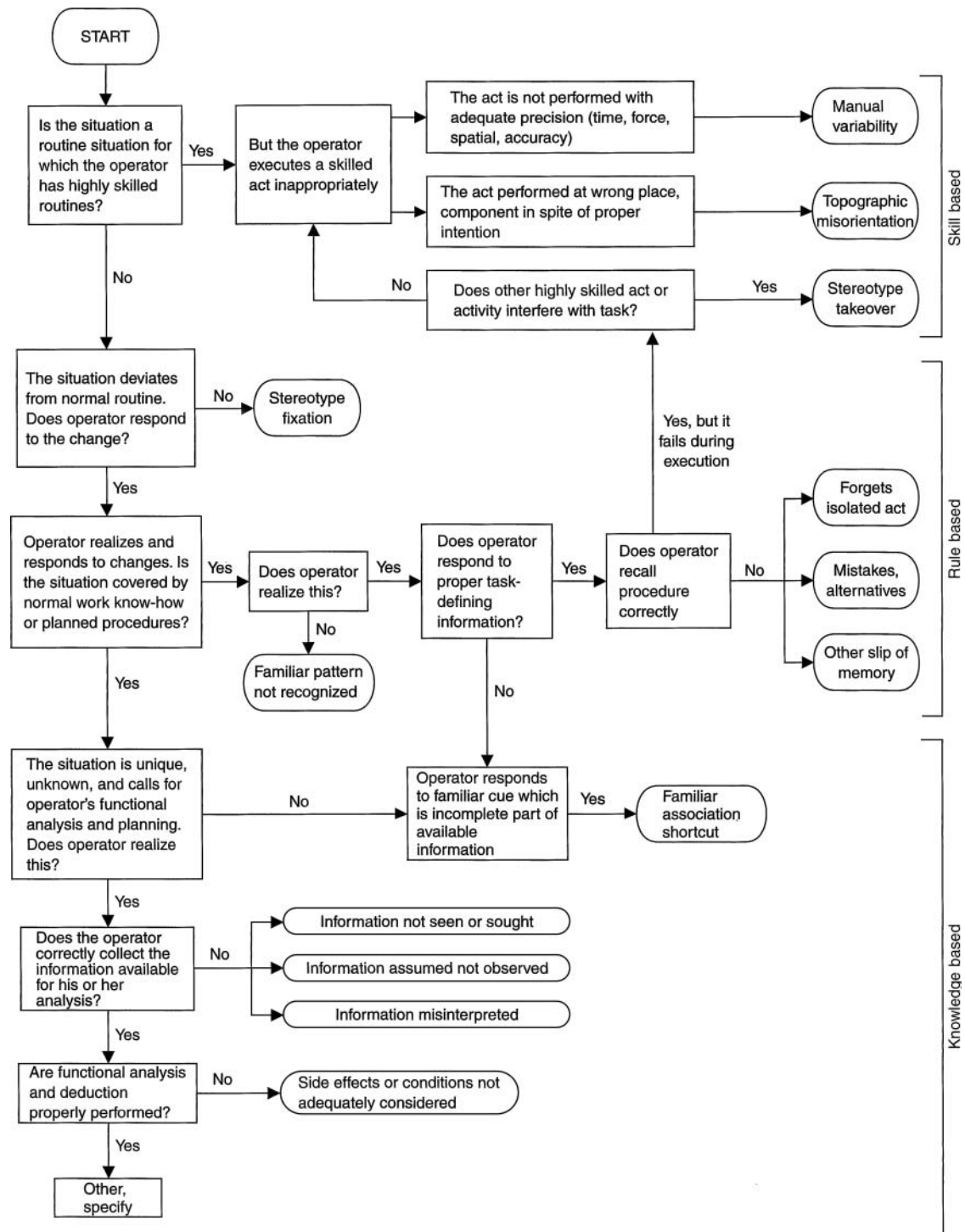
**Figure 4** Decision flow diagram for analyzing an event into one of 13 types of human error. (From Rasmussen, 1982. Copyright 1982 with permission from Elsevier.)

**Table 4  Human Error Modes Associated with Rasmussen's SRK Framework**

### Skill-Based Performance

| Inattention | Overattention |
| --- | --- |
| Double-capture slips | Omissions |
| Omissions following | Repetitions |
| interruptions | Reversals |
| Reduced intentionality | |
| Perceptual confusions | |
| Interference errors | |

### Rule-Based Performance

| Misapplication of Good Rules | Application of Bad Rules |
| --- | --- |
| First exceptions | Encoding deficiencies |
| Countersigns and | Action deficiencies |
| nonsigns | Wrong rules |
| Informational overload | Inelegant rules |
| Rule strength | Inadvisable rules |
| General rules | |
| Redundancy | |
| Rigidity | |

### Knowledge-Based Performance

| Selectivity | Problems with complexity |
| --- | --- |
| Workspace limitations | Problems with delayed |
| Out of sight, out of mind | feedback |
| Confirmation bias | Insufficient consideration |
| Overconfidence | of processes in time |
| Biased reviewing | Difficulties with |
| Illusory correlation | exponential |
| Halo effects | developments |
| Problems with causality | Thinking in causal series |
| | and not causal nets |
| | Thematic vagabonding |
| | Encysting |

*Source*: Reason (1990). Copyright © Cambridge University Press 1990. Reprinted with permission of Cambridge University Press.

TA can often be enhanced through the use of a variety of auxiliary tools. For example, the analyst may choose to employ checklists that cover a broad range of ergonomic considerations to determine if the human is being subjected to factors (such as illumination, noise, awkward postures, or poor interfaces) that can contribute to erroneous actions. These types of checklists can be expanded to include human fallibility considerations (Section 2.2) and contextual factors (Section 2.3) at various levels of detail.

However, prior to making any such embellishments, it is essential that the analyst identify an appropriate TA method for the particular problem or work domain as a number of different methods exist for performing TA (e.g., Kirwan and Ainsworth, 1992; Luczak, 1997; Shepherd, 2001; Chapter 13). Also, task analysts contending with complex systems will often need to consider various properties of the wider system or subsystem in which human activities take place (Sharit, 1997). As noted by Shepherd (2001), "Any task analysis method which purports to serve practical ends needs to be carried out beneath a general umbrella of systems thinking" (p. 11).

**Table 5  External Error Modes Classified According to Stages of Human Information Processing**

1. Activation/detection
   1.1  Fails to detect signal/cue
   1.2  Incomplete/partial detection
   1.3  Ignore signal
   1.4  Signal absent
   1.5  Fails to detect deterioration of situation
2. Observation/data collection
   2.1  Insufficient information gathered
   2.2  Confusing information gathered
   2.3  Monitoring/observation omitted
3. Identification of system state
   3.1  Plant-state identification failure
   3.2  Incomplete-state identification
   3.3  Incorrect-state identification
4. Interpretation
   4.1  Incorrect interpretation
   4.2  Incomplete interpretation
   4.3  Problem solving (other)
5. Evaluation
   5.1  Judgment error
   5.2  Problem-solving error (evaluation)
   5.3  Fails to define criteria
   5.4  Fails to carry out evaluation
6. Goal selection and task definition
   6.1  Fails to define goal/task
   6.2  Defines incomplete goal/task
   6.3  Defines incorrect or inappropriate goal/task
7. Procedure selection
   7.1  Selects wrong procedure
   7.2  Procedure inadequately formulated/shortcut invoked
   7.3  Procedure contains rule violation
   7.4  Fails to select or identify procedure
8. Procedure execution
   8.1  Too early/late
   8.2  Too much/little
   8.3  Wrong sequence
   8.4  Repeated action
   8.5  Substitution/intrusion error
   8.6  Orientation/misalignment error
   8.7  Right action on wrong object
   8.8  Wrong action on right object
   8.9  Check omitted
   8.10  Check fails/wrong check
   8.11  Check mistimed
   8.12  Communication error
   8.13  Act performed wrongly
   8.14  Part of act performed
   8.15  Forgets isolated act at end of task
   8.16  Accidental timing with other event/circumstance
   8.17  Latent error prevents execution
   8.18  Action omitted
   8.19  Information not obtained/transmitted
   8.20  Wrong information obtained/transmitted
   8.21  Other

*Source*: Kirwan (1994).
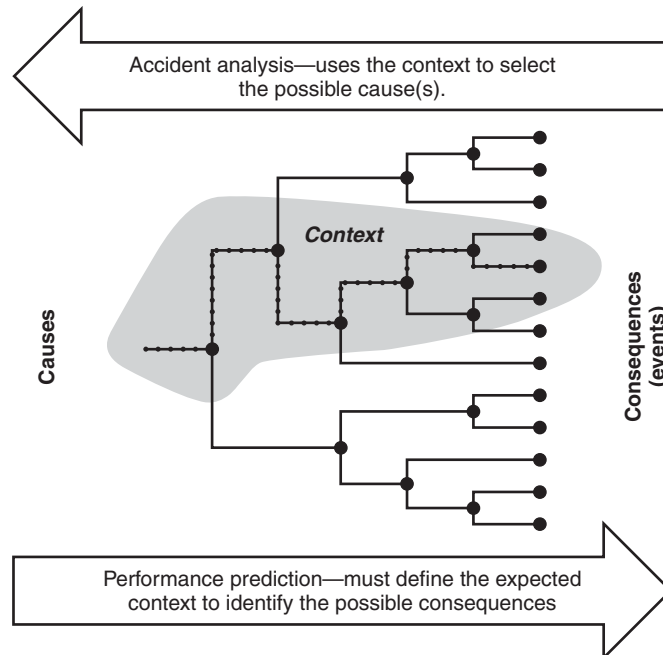
**Figure 5** Backward reasoning from events and context to analysis of causes is a much more constrained process than prediction of events through human actions in context. (From Hollnagel, 1993. Copyright 1993 with permission from Academic Press, Elsevier.)

In cognitive task analysis (CTA), the interest is in determining how the human conceptualizes tasks, recognizes critical information and patterns of cues, assesses situations, makes discriminations, and uses strategies for solving problems, forming judgments, and making decisions. Successful application of CTA for enhancing system performance will depend on a concurrent understanding of the cognitive processes underlying human performance in the work domain and the constraints on cognitive processing that the work domain imposes (Vicente, 1999). In developing new systems, meeting this objective may require multiple, coordinated approaches. As Potter et al. (1998) have noted: "No one approach can capture the richness required for a comprehensive, insightful CTA" (p. 395).

As with TA, many different CTA techniques are presently available (Hollnagel, 2003). TA and CTA, however, should not be viewed as mutually exclusive enterprises—in fact, the case could be made that TA methods that incorporate CTA represent "good" task analyses. With respect to the prediction of errors, generally TA should be capable of uncovering answers to the following questions: What kinds of actions by people are capable of resulting, by one's definition, in errors? What are the possible consequences of these errors? What kinds of barriers do these errors and their consequences call for?

Even when applied at relatively superficial levels, TA techniques are well suited for identifying mismatches between demands imposed by the work context and the human's capabilities for meeting these demands. At this level of analysis, windows of opportunity for error could still be readily exposed that, in and of themselves, can suggest countermeasures capable of reducing risk potential. For example, these analyses may determine that there is insufficient time to input information accurately into a computer-based documentation system; that the design of displays is likely to evoke control responses that are contraindicated; or that sources of information on which high-risk decisions are based contain incomplete or ambiguous information. This coarser approach to predicting errors or error-inducing conditions that derives from analyzing *demand-capability mismatches* can also highlight contextual and cognitive considerations that can form the basis for a more focused application of TA or CTA techniques.

In a type of TA known as a *hierarchical task analysis* (HTA), if the human–system interactions or operations underlying a goal cannot be usefully described or examined, then the goal is reexamined in terms of its subordinate goals and their accompanying plans—a process referred to as "redescription" (Shepherd, 2001). Table 6 depicts a portion of an HTA that was developed for analyzing the task of filling a storage tank with chlorine from a tank truck. The primary purpose of this HTA was to identify potential human errors that could contribute to a major flammable release resulting either from a spill during unloading of the truck or from a tank rupture. From this relatively simple HTA, identifying

**Table 6  Part of a Hierarchical Task Analysis Associated with Filling a Chlorine Tanker**

0.  Fill tanker with chlorine.
    *Plan*: Do tasks 1–5 in order.
1.  Park tanker and check documents (not analyzed).
2.  Prepare tanker for filling.
    *Plan*: Do 2.1 or 2.2 in any order, then do 2.3–2.5 in order.
    2.1  Verify tanker is empty.
        *Plan*: Do in order:
        2.1.1  Open test valve.
        2.1.2  Test for $Cl_2$.
        2.1.3  Close test valve.
    2.2  Check weight of tanker.
    2.3  Enter tanker target weight.
    2.4  Prepare fill line.
        *Plan*: Do in order:
        2.4.1  Vent and purge line.
        2.4.2  Ensure main $Cl_2$ valve is closed.
    2.5  Connect main $Cl_2$ fill line.
3.  Initiate and monitor tanker filling operation.
        *Plan*: Do in order:
    3.1  Initiate filling operation.
        *Plan*: Do in order:
        3.1.1  Open supply line valves.
        3.1.2  Ensure tanker is filling with chlorine.
    3.2  Monitor tanker-filling operation.
        *Plan*: Do 3.2.1, do 3.2.2 every 20 min;
        on initial weight alarm, do 3.2.3 and 3.2.4;
        on final weight alarm, do 3.2.5 and 3.2.6.
        3.2.1  Remain within earshot while tanker is filling.
        3.2.2  Check tanker while filling.
        3.2.3  Attend tanker during last filling of 2 or 3 tons.
        3.2.4  Cancel initial weight alarm and remain at controls.
        3.2.5  Cancel final weight alarm.
        3.2.6  Close supply valve A when target weight is reached.
4.  Terminate filling and release tanker.
    4.1  Stop filling operation.
        *Plan*: Do in order:
        4.1.1  Close supply valve B.
        4.1.2  Clear lines.
        4.1.3  Close tanker valve.
    4.2  Disconnect tanker.
        *Plan*: Repeat 4.2.1 five times, then do 4.2.2–4.2.4 in order.
        4.2.1  Vent and purge lines.
        4.2.2  Remove instrument air from valves.
        4.2.3  Secure blocking device on valves.
        4.2.4  Break tanker connections.
    4.3  Store hoses.
    4.4  Secure tanker.
        *Plan*: Do in order:
        4.4.1  Check valves for leakage.
        4.4.2  Secure log-in nuts.
        4.4.3  Close and secure dome.
    4.5  Secure panel (not analyzed).
5.  Document and report (not analyzed).

*Source*: CCPS (1994). Copyright 1994 by the American Institute of Chemical Engineers. Reproduced by permission of AIChE.

external error modes is a relatively straightforward matter.

For example, consider steps 2.3, 3.2.2, 4.1.2, and 4.2.1 in Table 6. Referring to the error classification scheme in Table 5, the manifestation of errors related to each of these actions would occur at the procedure execution stage (stage 8) of information processing:

- For step 2.3 (enter tanker target weight), it would be 8.20: Wrong information obtained, which would lead to entering an incorrect weight.
- For step 3.2.2 (check tanker while filling), it would be 8.9: Check omitted.
- For step 4.1.3 (close tanker valve), it would be 8.18: Action omitted.
- For step 4.2.1 (vent and purge lines), it would be 8.2: Operation incomplete.

Tabular formats are often used to accompany such stepwise task descriptions, allowing for the inclusion of a variety of complimentary assessments. For example, for each task step of a HTA, one column could be assigned to address the possible kinds of performance failures that could arise from these actions; additional columns could be used to document possible causes and consequences of these failures; and still other columns could be directed at error reduction recommendations, which could be further categorized into error mitigation or elimination strategies that resort to procedures, training, or hardware/software (CCPS, 1994) or the use of cognitive error detection strategies (Section 2.4.5).

The taxonomy shown in Table 5 can also be linked to more underlying psychological mechanisms. This would enable errors with identical or similar external manifestations to be distinguished and thus add considerable depth to the understanding of potential errors predicted from the TA. An example of such a scheme is the human error identification in systems technique (HEIST), which classifies external error modes according to the eight stages of human information processing listed in Table 5. The first column in a HEIST table consists of a code whose initial letter(s) refers to one of these eight stages. The next letter in the code refers to one of six general PSFs (Section 2.3): time (T), interface (I), training/experience/familiarity (E), procedures (P), task organization (O), and task complexity (C). These codes can then be linked to external error modes based on various underlying psychological error mechanisms (PEMs). Many of these mechanisms are consistent with the failure modes in Reason's error taxonomy (Table 4).

Table 7 presents an extract from a HEIST table containing a sample of items related to the first two of the eight stages of human information processing listed in Table 5: activation/detection (corresponding to codes beginning with "A") and observation/data collection (corresponding to codes beginning with "O"). More detailed explanations of some of the PEMs listed in the HEIST table are presented in Table 8. A complete HEIST table and the corresponding listing of PEMs can be found in Kirwan (1994).

The human reliability analysis method known as CREAM (Section 4.10) developed by Hollnagel (1998)

**Table 7  Some Items from HEIST Table Corresponding to First Two Stages in Table 5**

| Code | Error Identifier Prompt | External Error Mode | System Cause/Psychological Error Mechanism | Error Reduction Guidelines |
|---|---|---|---|---|
| AT | Does the signal occur at the appropriate time? Could it be delayed? | Action omitted or performed either too early or too late | Signal-timing deficiency, failure of prospective memory | Alter system configuration to present signal appropriately; repeat signal until action has occurred. |
| AI | Is the signal strong and in a prominent location? Could the signal be confused with another? | Action omitted or performed too late or wrong act performed | Signal detection failure | Prioritize signals; place signals in primary location; use diverse signals; use multiple-signal coding; give training in signal priorities; increase signal intensity. |
| AE | Does the operator understand the significance of the signal? | Action omitted or performed too late | Inadequate mental model | Training and procedures should be amended to ensure that significance is understood. |
| AO | Will the operator have a very high or low workload? | Action omitted or performed either too late or too early | Lapse of memory, other memory failure, signal detection failure | Improve task and crew organization; use a recurring signal; consider automation; utilize flexible crewing; enhance signal salience. |
| AC | Is the signal in conflict with the current diagnostic mindset? | Action omitted or wrong act performed | Confirmation bias, signal ignored | Procedures should emphasize disconfirming as well as confirmatory signals; carry out problem-solving training and team training; implement automation. |
| OT | Could the information or check occur at the wrong time? | Failure to act or action performed too late or too early or wrong act performed | Inadequate mental model/inexperience/crew coordination failure | Procedure and training should specify the priority and timing of checks; present key information centrally; utilize trend displays and predictor displays if possible; implement team training. |
| OI | Are any information sources ambiguous? | Action omitted or performed too late or too early or wrong act performed | Misinterpretation, mistakes alternatives | Use task-based displays; design symptom-based diagnostic aids; utilize diverse information sources; ensure clarity of information displayed; utilize alarm conditioning. |
| OE | Could the operator interrogate too many information sources for too long? | Action omitted or performed too late | Thematic vagabonding, risk recognition failure, inadequate mental model | Provide training in fault diagnosis; put procedural emphasis on required data collection time frames; implement high-level indicators (alarms) of system integrity deterioration. |
| OP | Could the operator forget one or more items in the procedures? | Action omitted or performed either too early or too late or wrong act performed | Forget isolated act, slip of memory, place-losing error | Ensure an ergonomic procedure design; utilize tick-off sheets, place keeping aids, etc.; provide team training to emphasize checking by other team member(s). |
| OO | Could information collected fail to be transmitted effectively across shift-handover boundaries? | Failure to act or action is wrong or performed too late or too early or an error of quality | Crew coordination failure | Develop robust shift-handover procedures; training; provide team training across shift boundaries; develop robust and auditable data-recording systems (logs). |
| OC | Does the scenario involve multiple events, thus causing a high level of complexity or a high workload? | Failure to act or wrong action performed or action performed either too early or too late | Cognitive overload | Provide emergency response training; use flexible crewing strategies; develop emergency operating procedures able to deal with multiple transients; generate decision/diagnostic support facilities. |

*Source:* Adapted from Kirwan (1994).

757

**Table 8  A Sample of Psychological Error Mechanism Descriptions for Some Items in Table 7 and Recommendations for Their Remediation**

*Vigilance failure:* lapse of attention. Ergonomic design of interface to allow provision of effective attention-gaining measures; supervision and checking; task-organization optimization, so that the operators are not inactive for long periods and are not isolated.

*Cognitive/stimulus overload:* too many signals present for the operator to cope with. Prioritization of signals (e.g., high-, medium-, and low-level alarms); overview displays; decision support systems; simplification of signals; flowchart procedures; simulator training; automation.

*Stereotype fixation:* operator fails to realize that situation has deviated from norm. Training and procedural emphasis on range of possible symptoms/causes; fault–symptom matrix as a job aid; decision support system; shift technical advisor/supervision.

*Signal discrimination failure:* operator fails to realize that the signal is different. Improved ergonomics in the interface design; enhanced training and procedural support in the area of signal differentiation; supervision checking.

*Confirmation bias:* operator only selects data that confirm given hypothesis and ignores other disconfirming data sources. Problem-solving training; team training; shift technical advisor (diverse, highly qualified operator who can ''stand back'' and consider alternative diagnoses), functional procedures: high-level information displays; simulator training.

*Thematic vagabonding:* operator flits from datum to datum, never actually collating it meaningfully. Problem-solving training; team training; simulator training; functional procedure specification for decision-timing requirements; high-level alarms for system integrity degradation.

*Encystment:* operator focuses exclusively on only one data source. Problem-solving training; team training (including training in the need to question decisions and in the ability of the team leader(s) to take constructive criticism); high-level information displays; simulator training; high-level alarms for system integrity degradation.

*Source* Adapted from Kirwan (1994).

has, at its core, a method for qualitative performance prediction that is highly dependent on TA. Fundamental to this approach is the distinction referred to in Section 2.3 between phenotypes (Section 3.1), which are the external error modes, and genotypes, which are the possible "causes" of these error modes. Hollnagel presents a large number of tables of genotypes; within each of these tables, the genotype is further resolved into "general consequents," which are, in turn, categorized into "specific consequents." When a general or specific consequent from one genotype can influence the consequents of one or more other genotypes, these initial consequents are considered antecedents. Ultimately, these influences can give rise to chains of antecedent-consequent links. Table 9 lists the consequents associated with the person-related genotype "observation," the technology-related genotype "equipment failure," and the organizational/environment-related genotype "communication."

One problem that can arise using this scheme is the combinatorial explosion of error prediction paths. Hollnagel argues that this potentially large solution space can be logically constrained if the context is sufficiently well known. Toward this end, he suggests using a relatively small set of common performance conditions (CPCs), which he believes contain the general determinants of performance, in order to produce a general context description (Table 10). Although these CPCs are intended to have minimal overlap, they are not considered to be mutually independent.

Using this scheme, the process of human performance or error prediction occurs as follows. First, an analysis of the operator control tasks using TA, as well as analysis of organizational and technical system considerations, is performed. Next, using the CPCs, the context is described. The CPCs serve to "prime" the various classification groups (e.g., Table 9), enabling the more logical or probable antecedent-consequent links, as well as the more likely error modes (Section 3.1), to be specified. The third step consists of specifying the initiating events. These are usually actions humans perform at the "sharp end" (Section 1.1) and are consistent with human actions that are of interest in probabilistic risk assessments (Section 4.1).

The fourth step uses the phenotype–genotype classification scheme to generate propagation paths that lead through the various "causes" of the sharp end's external error mode. The CPCs are used to constrain the propagation paths by allowing the analyst to consider only those consequents that are consistent with the situation; otherwise, the nonhierarchical ordering of the genotype classification groups can produce an excessive number of steps. Phenotypes will always be categorized as consequents as they are the endpoints of the paths.

More recently, an approach to human performance prediction has been proposed that consists of an integration of a number of human factors and system safety hazard analysis techniques (Sharit, 2008). The starting point of this methodology is a TA. The results of the TA become the "human components" of a failure modes and effects analysis (FMEA), a hazard evaluation technique (Kumamoto and Henley, 1996) that in its conventional implementation requires specifying the failure modes for each system component, assembly, or subsystem as well as the consequences and causes of these failure modes. The mapping from the steps of the TA to the possible human performance failures modes essentially results in a "human" FMEA (HFMEA). This process is aided by a classification system in the form of a checklist that considers four broad categories of behavior: perceptual processes (searching for and receiving information and identifying objects, actions, and events); mediational processes (information processing

**Table 9  General and Specific Consequents of Three Genotypes**

| | | |
|---|---|---|
| | ***Person-Related Genotype: "Observation"*** | |
| General Consequent | Specific Consequent | Definition/Explanation |
| Observation missed | Overlook cue/signal | A signal or an event that should have been the start of an action (sequence) is missed. |
| | Overlook measurement | A measurement or some information is missed, usually during a sequence of actions. |
| False observation | False reaction | A response is given to an incorrect stimulus or event, e.g., starting to drive when the light changes to red. |
| | False recognition | An event or some information is incorrectly recognized or mistaken for something else. |
| Wrong identification | Mistaken cue | A signal or a cue is misunderstood as something else. Unlike in a "false reaction," it does not immediately lead to an action. |
| | Partial identification | The identification of an event or some information is incomplete, e.g., as in jumping to a conclusion. |
| | Incorrect identification | The identification of an event/information is incorrect but, unlike in a "false recognition," is a more deliberate process. |
| | ***Technology-Related Genotype: "Equipment Failure"*** | |
| Equipment failure | Actuator stick/slip | An actuator or control either cannot be moved or moves too easily. |
| | Blocking | Something obstructs or is in the way of an action. |
| | Release | Uncontrolled release of matter or energy that causes other equipment to fail. |
| | Speed up/slow down | The speed of the process (e.g., a flow) changes significantly. |
| | No indicators | An equipment failure occurs without a clear signature. |
| Software fault | Performance slowdown | The performance of the system slows down. This can in particular be critical for command and control. |
| | Information delays | There are delays in the transmission of information, hence in the efficiency of communication, both within the system and between systems. |
| | Command queues | Commands or actions are not being carried out because the system is unstable, but are (presumably) stacked. |
| | Information not available | Information is not available due to software or other problems. |
| | ***Organization-Related Genotype: "Communication"*** | |
| Communication failure | Message not received | The message or the transmission of information did not reach the receiver. This could be due to incorrect address or failure of communication channels. |
| | Message misunderstood | The message was received, but it was misunderstood. The misunderstanding is, however, not deliberate. |
| Missing information | No information | Information is not being given when it was needed or requested, e.g., missing feedback. |
| | Incorrect information | The information being given is incorrect or incomplete. |
| | Misunderstanding | There is a misunderstanding between sender and receiver about the purpose, form, or structure of the communication. |

*Source:* Adapted from Hollnagel (1998) by permission of Elsevier.

and problem solving/decision making); communication processes; and motor execution processes.

A well-known disadvantage of FMEAs is their emphasis on single-point failures (e.g., a valve failing open), which increases the likelihood of failing to account for adverse system outcomes deriving from multiple coexisting hazards or failures (U.S. Department of Health and Human Services, 1998). This problem is overcome in the proposed methodology by combining the HFMEA with the hazard and operability (HAZOP) analysis method (CCPS, 1992), a hazard analysis technique that, through creative brainstorming, can enable further insight into possible human–system failures. HAZOP uses a very systematic and thorough approach to analyze points of a process or operation, referred to as "study nodes" or "process sections," by applying, at each point of the process being analyzed, guide words (such as "no," "more," "high," "reverse," "as well as," and "other than") to parameters (such as "flow," "pressure," "temperature," and "operation") in order to generate

**Table 10  Common Performance Conditions**

| CPC Name | Description | Level (Typical ''Values'' They Can Take On) | Expected Effect On Performance Reliability |
| --- | --- | --- | --- |
| Adequacy of organization | The quality of the roles and responsibilities of team members, additional support, communication systems, safety management system, instructions and guidelines, role of external agencies, etc. | Very efficient<br>Efficient<br>Inefficient<br>Deficient | Improved<br>Not significant<br>Reduced<br>Reduced |
| Working conditions | The nature of the physical working conditions such as ambient lighting, glare on screens, noise from alarms, interruptions from the task, etc. | Advantageous<br>Compatible<br>Incompatible | Improved<br>Not significant<br>Reduced |
| Adequacy of MMI and operational support | The man–machine Interface in general, including the information available on displays, workstations, and operational support provided by decision aids. | Supportive<br>Adequate<br>Tolerable<br>Inappropriate | Improved<br>Not significant<br>Not significant<br>Reduced |
| Availability of procedures/ plans | Procedures and plans include operating and emergency procedures, familiar patterns of response heuristics, routines, etc. | Appropriate<br>Acceptable<br>Inappropriate | Improved<br>Not significant<br>Reduced |
| Number of simultaneous goals | The number of tasks a person is required to pursue/attend to at the same time (i.e., evaluating the effects of actions, sampling new information, assessing multiple goals). | Fewer than capacity<br>Matching current capacity<br>More than capacity | Not significant<br>Not significant<br>Reduced |
| Available time | The time available to carry out a task; corresponds to how well the task execution is synchronized to the process dynamics. | Adequate<br>Temporarily inadequate<br>Continuously inadequate | Improved<br>Not significant<br>Reduced |
| Time of day (circadian rhythm) | The time of day/night describes the time at which the task is carried out, in particular whether or not the person is adjusted to the current time (circadian rhythm). | Daytime (adjusted)<br>Nighttime (unadjusted) | Not significant<br>Reduced |
| Adequacy of training and experience | The level and quality of training provided to operators as familiarization to new technology, refreshing old skills, and also the level of operational experience. | Adequate, high experience<br>Adequate, limited experience<br>Inadequate | Improved<br>Not significant<br>Reduced |
| Crew collaboration quality | The quality of the collaboration between crew members, including the overlap between the official and unofficial structure, the level of trust, and the general and social climate among crew members. | Very efficient<br>Efficient<br>Inefficient<br>Deficient | Improved<br>Not significant<br>Not significant<br>Reduced |

*Source*: Adapted from Hollnagel (1998) by permission of Elsevier.

deviations (such as "no flow" or "high temperature") that represent departures from the design intention. The key to integrating HAZOP with HFMEA is to derive "guide words" and "parameters" that are applicable to the TA.

The proposed methodology also incorporates two additional checklists: One that aids analysts in identifying relevant contextual factors and a second that provides a detailed listing of human tendencies and limitations. Using the first aid, the objective is to assemble, through some form of representation (e.g., through an unconstrained network approach as discussed in Section 2.3), various realistic scenarios that characterize the conditions under which the human performs the activities identified in the TA. Using the second aid, the analytical team would then need to determine which human tendencies or limitations are relevant to the contexts under examination and how these tendencies could result in errors or behaviors that undermine system performance.

Other brainstorming methods, such as what-if analysis, are suggested for analyzing dependencies, such as the impact of human performance failures upon other (impending) human behaviors, and the effects on the system of multiple human failures that may or may not be coupled. Inherent in the methodology is the consideration of barriers that can prevent or mitigate the adverse consequences of the human performance failures or that can promote new previously unforeseen risks.

## 4  HUMAN RELIABILITY ANALYSIS

### 4.1  Probabilistic Risk Assessment

Two sets of tools that analysts often resort to for assuring the safety of systems with hazard potential are (1) traditional safety analysis techniques (CCPS, 1992) such as FMEA (Section 3.2), which utilize primarily qualitative methods, and (2) quantitative risk assessment

procedures (Kumamoto and Henley, 1996), most notably probabilistic risk assessment (PRA). According to Apostolakis (2004), PRAs are not intended to replace other safety methods but rather should be viewed as an additional tool in safety analysis that is capable of *informing* safety-related decision making.

Early on in the development of PRA, analysts recognized that a realistic evaluation of the risks of system operations would require integrating human reliability—the *probability of human failures* in critical system interactions—with hardware and software reliability analysis. In PRA, the objectives of human reliability analysis (HRA) are to identify, represent (within the logic structure of the system or plant PRA), and quantify those human errors or failures for the purpose of determining their contribution to *predetermined* system failures.

In PRAs, it is these failures that are initially specified. For each such consequence or *end state*, disturbances to normal operation, referred to as *initiating events*, are then identified that are capable of leading to these end states. Finally, through plant or system models typically represented as *event trees* or *fault trees* (Section 4.1.1), the sequence of events linking initiating events to end states is developed. The assignment of probabilities to the events leading to accidents ultimately enables the accident scenarios to be ranked according to their risk potential.

The human errors or actions that are considered in a PRA study are often grouped into three categories. The first category consists of *preinitiator* human events. These are actions during normal operations such as faulty calibrations or misalignments that can cause equipment or systems to be unavailable when required. The second category consists of *initiator* human events, which are actions that either by themselves or in combination with equipment failures can lead to initiating events. The third category involves *postfault* human actions. These can include human actions during the accident that, due to the inadequate recognition of the situation or the selection of the wrong strategy, make the situation worse or actions, such as improper repair of equipment, that prevent recovery of the situation.

This categorization of human actions in PRAs highlights the subtle but very important distinction that should be made between human error and human failure, as in some contexts they can have very different meanings. For example, in the category of postfault human actions, following the execution of a particular recovery action there may be insufficient time, through no fault of the human, to perform a subsequent emergency operating procedure.

The catalyst for one of the first HRA methods to be proposed was the problem of nuclear weapons assembly by humans. Alan Swain approached this problem by resorting to detailed task analysis of the steps involved in the assembly process and seeking, through various means, estimates of the probabilities of human errors for each of these steps. This approach was referred to as the technique for human error rate prediction (THERP) and ultimately evolved into a systematic and highly elaborate HRA method that targeted the safety of nuclear power plant operations (Swain and Guttman, 1983).

The WASH 1400 study (1975) led by Norman Rasmussen is often cited as the first formal PRA (Spurgin, 2010). It was directed at investigating accidents resulting from single failures (e.g., a loss of coolant accident) in pressurized and boiling water reactors. This study relied in large part on THERP for deriving human error probabilities (HEPs). Further developments in HRA methods and ways in which they could be incorporated into PRAs involved a number of organizations within the United States such as the U.S. Nuclear Regulatory Commission (NRC), Oak Ridge National Laboratory, and the Electric Power Research Institute (EPRI). Other countries were also making major contributions to HRA (Spurgin, 2010), by either modifying proposed methods or developing new methods.

The use of PRAs in the nuclear industry has also influenced the use of quantitative risk assessment methods in other industries, most notably industries involved in chemical, waste repository, and space operations (Apostolakis, 2004). In addition, other agencies, such as the Environmental Protection Agency, the Food and Drug Administration, and state air and water quality agencies, have also come to embrace NRC-type policies and procedures (Kumamoto and Henley, 1996) and have established their own approaches to assessing risks from human error.

Over the years, HRA has evolved into a discipline that has come to mean different things to different people. This broader perspective to HRA encompasses conceptual and analytic tools needed for understanding how a system's complexity and dynamics can impact human actions and decisions; the appraisal of human errors that may arise within the context of system operations; and design interventions in the form of various barriers that can eliminate or mitigate these negative effects. Within this broader perspective, the choice still remains whether to pursue quantitative estimates of human error probabilities and their contribution to system risks.

An objective assignment of probabilities to human failures implies that HEP be defined as a ratio of the number of observed occurrences of the error to the number of opportunities for that error to occur. Thus, it can be argued that with the possible exception of routine skill-based activities it is questionable whether reliable estimates of HEPs are obtainable. This leaves open the prospect for further diluting the uncertainty that is already implicit to many quantitative system risk assessments such as PRAs.

However, what is often not given sufficient consideration is that the process itself of performing a PRA, irrespective of the precise quantitative figures that they are intended to produce, can provide a number of important benefits (Kumamoto and Henley, 1996; Apostolakis, 2004). Many of the tangible benefits derive from systematic and comprehensive qualitative HRA efforts and can become manifest in the form of improvements in operating procedures in maintenance, testing, and emergency procedures; the kinds of collaborations among workers that are most likely to have safety benefits through redundancy effects; the types of interfaces and

aiding devices that are most likely to improve efficiency during normal operations and response capabilities during emergencies; and clearer identification of areas that would benefit from training, especially in the human's ability to detect, diagnose, and respond to incidents.

Benefits of PRAs of a less tangible nature include improved plant or system knowledge among design, engineering, and operations personnel regarding overall plant design and operations, especially in relation to the complex interactions between subsystems. PRAs also provide a common understanding of issues, thus facilitating communication among various stakeholder groups. Finally, by virtue of their emphasis on quantifying uncertainty, PRAs can better expose the boundaries of expert knowledge concerning particular issues and thereby inform decisions regarding needed research in diverse disciplines ranging from physical phenomena to the behavioral and social sciences.

### 4.1.1  Fault Trees and Event Trees

The two primary hazard analysis techniques that have become associated with PRAs are *fault tree* (FT) analysis and *event tree* (ET) analysis. These techniques can be applied to larger scale system events, for example, as a plant model in a PRA that might include human tasks, or to specified human tasks in order to analyze these tasks in terms of their more elemental task components. The starting point for each of these methods is an undesirable event (e.g., an undesirable system event or an undesirable human task event), whose identification often relies on other hazard analysis techniques (CCPS, 1992) or methods based on expert judgment.

An ET corresponds to an *inductive analysis* that seeks to determine how this undesirable event can propagate into an accident. These trees are thus capable of depicting the various sequences of events that can unfold following the initiating event as well as the risks associated with each of these sequences. Figure 6 depicts a simplified event tree for a loss-of-coolant accident-initiating event in a typical nuclear power plant (Kumamoto and Henley, 1996). The initiating event is a coolant pipe break having a probability (or frequency of occurrence per time period) of $P_A$. The event tree depicts the alternative courses of events that might follow. First, the availability of electric power is considered, followed by the next-in-line system, which is the emergency core-cooling system, whose failure results in the meltdown of fuel and varying amounts of nuclear fission product release depending on the containment integrity.

Figure 7 depicts an ET for an offshore emergency shutdown scenario in a chemical processing scenario. Because it is the sequence of human actions in response to an initiating event that is being addressed, this type of ET is often referred to as an operator action event tree (OAET). In both cases, each branch represents either success (the upper branch) or failure (represented in the OAET as an HEP) in achieving the required actions specified along the top. The probability of each end state on the right is the product of the failure/error or success probabilities of each branch leading to that end state, and the overall probability of any specified failure end state is the sum of the probabilities of the corresponding individual failure end states. In the OAET, the dashed lines indicate paths through which recovery from previous errors can occur.

| A | B | C | D | E | Probability | State |
|---|---|---|---|---|---|---|
| Pipe break | Electric power | ECCS | Fission product removal | Containment integrity | | |
| | | | | Succeeds $\bar{P}_{E1} = 1 - P_{E1}$ | $P_A\bar{P}_B\bar{P}_{C1}\bar{P}_{D1}\bar{P}_{E1}$ | Very small release |
| | | | Succeeds $\bar{P}_{D1} = 1 - P_{D1}$ | Fails $P_{E1}$ | $P_A\bar{P}_B\bar{P}_{C1}\bar{P}_{D1}P_{E1}$ | Small release |
| | | Succeeds $\bar{P}_{C1} = 1 - P_{C1}$ | Fails $P_{D1}$ | Succeeds $\bar{P}_{E2} = 1 - P_{E2}$ | $P_A\bar{P}_B\bar{P}_{C1}P_{D1}\bar{P}_{E2}$ | Small release |
| | | | | Fails $P_{E2}$ | $P_A\bar{P}_B\bar{P}_{C1}P_{D1}P_{E2}$ | Medium release |
| | Succeeds $\bar{P}_B = 1 - P_B$ | Fails $P_{C1}$ | Succeeds $\bar{P}_{D2} = 1 - P_{D2}$ | | $P_A\bar{P}_BP_{C1}\bar{P}_{D2}$ | Large release |
| Initiating event $P_A$ | | | Fails $P_{D2}$ | | $P_A\bar{P}_BP_{C1}P_{D2}$ | Very large release |
| | Fails $P_B$ | | | | $P_AP_B$ | Very large release |

**Figure 6**  Simple event tree for a loss-of-coolant accident with two operator actions and two safety systems. (From Kumamoto and Henley, 1996. Copyright © 2004 by IEEE.)

| S1 begins ESD demand | CCR operator initiates ESD within 20 min | Supervisor initiates ESD within the same 20 min | Operator detects only partial ESD has occured (within 2 h) | Supervisor detects only partial ESD has occured (within 2 h) | CCR operator identifies correct equipment room to outside operator | Outside operator identifies failed activator and communicates these to CCR operator | CCR operator identifies manual valves and tells outside operator | Outside operator moves valves to correct position within the same 2 h | End state |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Success |
| | | | | | | | | HEP 1.8 | Failure |
| | | | | | | | HEP 1.7 | | Failure |
| | | | | | | HEP 1.6 | | | Failure |
| | | | | | HEP 1.5 | | | | Failure |
| | | | HEP 1.3 | | | | | | |
| | | | | HEP 1.4 | | | | | Failure |
| | HEP 1.1 | | | | | | | | |
| | | HEP 1.2 | | | | | | | Failure |

**Figure 7** An OAET: ESD, emergency shutdown procedure; CCR, chemical control room. (From Kirwan, 1994.)

In contrast to ETs, an FT represents a *deductive*, top-down decomposition of an undesirable event, such as a loss in electrical power or failure by a human to detect a critical event. In PRAs, FTs utilize Boolean logic models to depict the relationships among hardware, human, and environmental events that can lead to the undesirable *top event*, where HRA is relied upon for producing the HEP inputs. When FTs are used as a quantitative method, *basic events* (for which no further analysis of the cause is carried out) are assigned probabilities or occurrence rates, which are then propagated into a probability or rate measure associated with the top event (Dhillon and Singh, 1981). FTs are also extremely valuable as a qualitative analysis tool, as they can exploit the use of Boolean logic to identify the various combinations of events (referred to as *cut sets*) that could lead to the top event and thus suggest where interventions should be targeted.

The inductive and deductive capabilities of ETs and FTs can go hand-in-hand in PRAs. When combining ETs and FTs, each major column of the ET can represent a top event (i.e., an undesirable event) whose failure probability can be computed through the evaluation of a corresponding FT model. Figure 8 illustrates a simple ET consisting of two safety systems and the two FTs needed to provide probability estimates for the safety system columns in this ET.
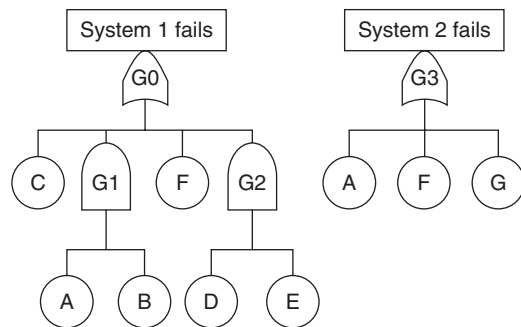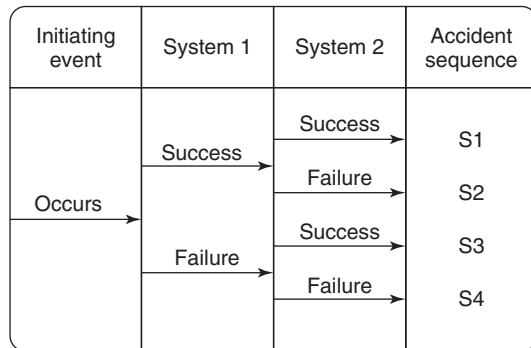


**Figure 8** Coupling of event trees and fault trees. The probabilities of failure associated with systems 1 and 2 in the event tree would be derived from the two corresponding fault tees. (From Kumamoto and Henley, 1996. Copyright © 2004 by IEEE.)

### 4.1.2 HRA Process

The recognition of HRA as a pivotal component of PRA does not necessarily ensure that HRA will be integrated effectively into PRA studies. Given the assumption that human reliability comprises somewhere between 60 and 80% of the total system risk, (e.g., Spurgin, 2010), it is thus imperative that HRA analysts not be excluded from and, ideally, have a substantial involvement in the PRA process.

Recommended practices for conducting HRA can be found in a number of publicly available sources. These include Institute for Electrical and Electronics Engineers (IEEE, 1997) standard 1082 for HRA, American Society for Mechanical Engineers (ASME, 2008) standard for probabilistic risk assessment (ASME STD-RA-S-2008), and the EPRI (1984) systematic human reliability procedure (SHARP; Hannaman and Spurgin, 1984).

It is important to emphasize that these recommended approaches to performing HRA do not imply a specific model for examining human interactions or a particular method for performing HRA and quantification of human errors. The specific needs of the organization will determine the nature of the PRA they wish to perform and thus dictate to some degree the specific HRA model requirements and needed data. However, the influence of the HRA analyst or team should not be discounted, as the biases these individuals have toward approaches to HRA and how these approaches will be implemented can determine, among other things, how contexts will be considered and how human behaviors will be modeled within these work contexts.

The 10-step HRA process proposed by Kirwan (1994) prominently highlights, in its earlier stages, the role of task analysis (Section 3.2) and human error analysis (Figure 9). It is in this respect that HRA can, in principle, be disconnected from PRA and serve objectives directed entirely to qualitative analysis of human error and error prediction (Section 3.2). This does not necessarily preclude quantification of human error, but neither does it imply that such quantification is necessary for identifying and adequately classifying risks to system operations stemming from human–system interactions.

### 4.2 Methods of HRA

In the ensuing sections, a number of proposed methods of HRA are discussed. Spurgin (2010) has characterized HRA methods into three classes: task related, time related, and context related. Another common classification scheme is to differentiate HRA methods in terms of being first or second generation. Some of the second-generation methods were intended in part to close the gaps of earlier methods (such as THERP) that were lacking in their consideration of human cognition in human error. Regardless of how one chooses to represent HRA methods, one fact that should not go unnoticed is that all methods rely, to some degree, on the use of expert judgment, whether it is to provide base estimates of human error probabilities, identify PSFs and determine their influence on human performance, or assess dependencies that might exist between people, tasks, or events.
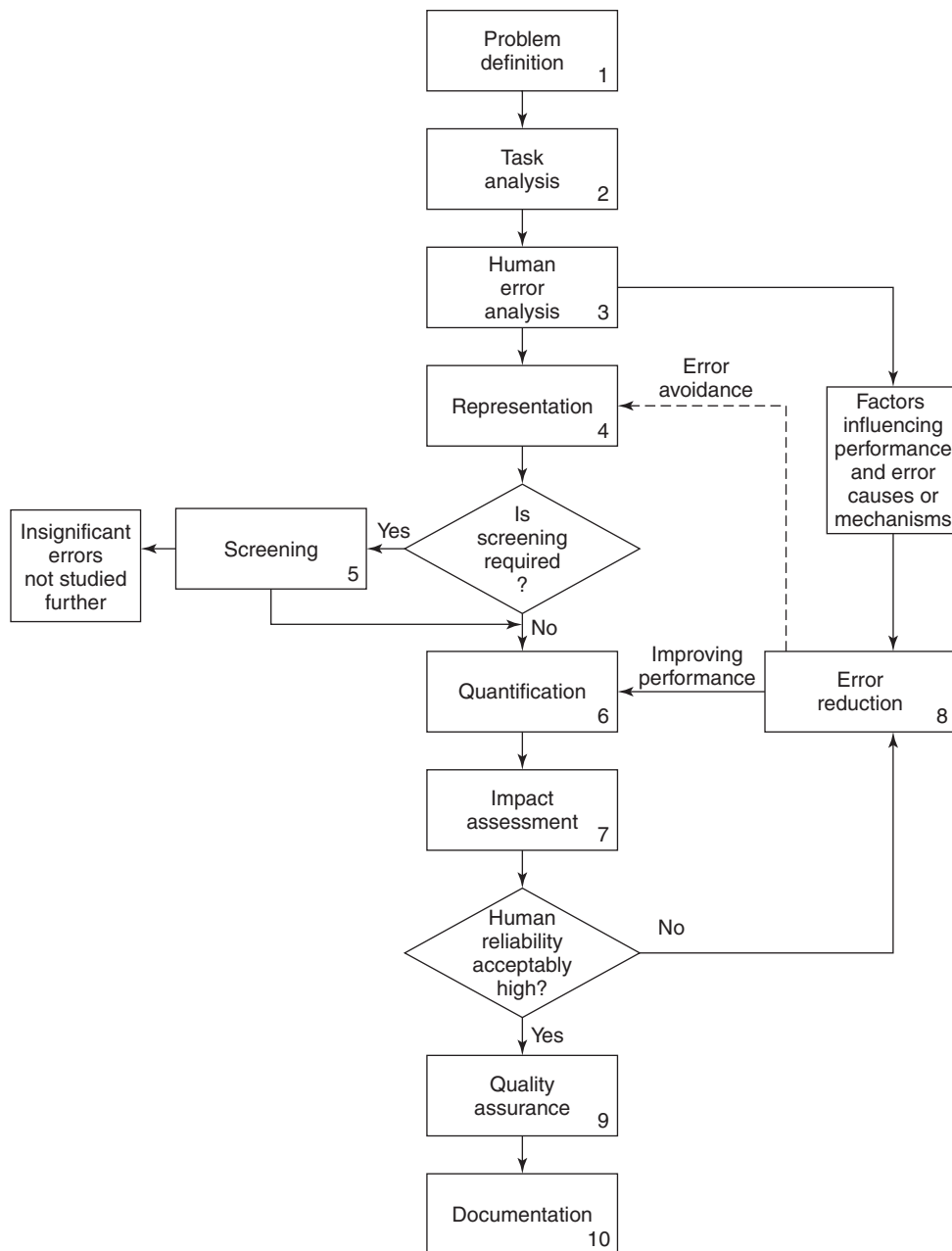
**Figure 9** The HRA process. (From Kirwan, 1994.)

Below, a sample of HRA methods is considered, beginning with THERP, which is still the most widely known method. Their discussion will hopefully underscore not only the historical unveiling of needs that motivated the development of alternative HRA methods but also the challenges that this discipline continually faces.

The coverage of these methods will be, by necessity, highly variable as there are many HRA methods that could potentially be considered. In no way is the degree of detail accorded to any HRA method intended to reflect the perceived importance of the method. Also, a number of highly respected methods are not covered at all, which, if anything, points to the challenge of doing this topic justice in a limited space. These methods include a technique for human event analysis (ATHEANA; Forester et al., 2007) and Method d′ Evaluation de la Realisation des Missions Operateur pour la Surete (MERMOS; Pesme et al., 2007).

## 4.3 THERP

The *technique for human error rate prediction*, generally referred to as THERP, is detailed in a work by Swain and Guttmann (1983) sponsored by the U.S. Nuclear Regulatory Commission. Its methodology is largely driven by decomposition and subsequent aggregation: Human tasks are first decomposed into clearly separable actions or subtasks; HEP estimates are then assigned to each of these actions; and, finally, these HEPs are aggregated to derive probabilities of task failure. These outputs could then be used as inputs for the analysis of system reliability (e.g., through the use of a system fault or event tree).

The procedural steps of THERP are outlined in Figure 10. Although these steps are depicted sequentially, in actuality there could be any of a number of feedback loops when carrying out this procedure. The first two steps involve establishing which work activities or events will require emphasis due to their risk potential and the human tasks associated with these activities or events. In steps 3–5, a series of qualitative assessments are performed. Walk-throughs and talk-throughs (e.g., informal interviews) are carried out to determine the "boundary conditions" under which the tasks are performed, such as time and skill requirements, alerting cues, and recovery factors.

Task analysis (Section 3.2) is then conducted to decompose each human task into a sequence of discrete activities. At this stage, it may be opportunistic for the analyst to repeat step 3, with the emphasis this time on encouraging workers to talk through hypothetical, yet realistic, work scenarios for the purpose of assessing the potential for human errors associated with the individual task activities. The analyst may also wish to pursue factors related to error detection and the potential for error recovery.

The results of these efforts are represented by an HRA event tree. In this tree, each relevant discrete task step or activity is characterized by two limbs representing either successful or unsuccessful performance. As indicated in the HRA event tree depicted in Figure 11, the probability that the failure occurs at a particular step in the task sequence is determined by multiplying the product of the probabilities of success on each of the preceding steps by the probability of failure on the step in question. Thus, in Figure 11, the probability that the failure occurs during the execution of step 2 of the task sequence is computed as $F_2 = 0.9898 \times (1 - 0.9845) = 0.0153$. The sum of $F_i, i = 1, \ldots, n$, represents the probability of failure in the performance of this task.

The next set of steps in THERP (steps 6–10) constitutes quantitative assessment procedures. First, HEPs are assigned to each of the limbs of the tree corresponding to incorrect performance. These probabilities, referred to as *nominal HEPs*, in theory are presumed to represent medians of lognormal probability distributions. Associated with each nominal HEP are *upper and lower uncertainty bounds* (UCBs), which reflect the variance associated with any given error distribution. The square root of the ratio of the upper to the lower UCB defines the *error factor* (the value selected for this factor will depend on the variability believed to be associated with the probability distribution for that error). Swain and Guttmann (1983) provide values of nominal HEPs and their corresponding error factors for a variety
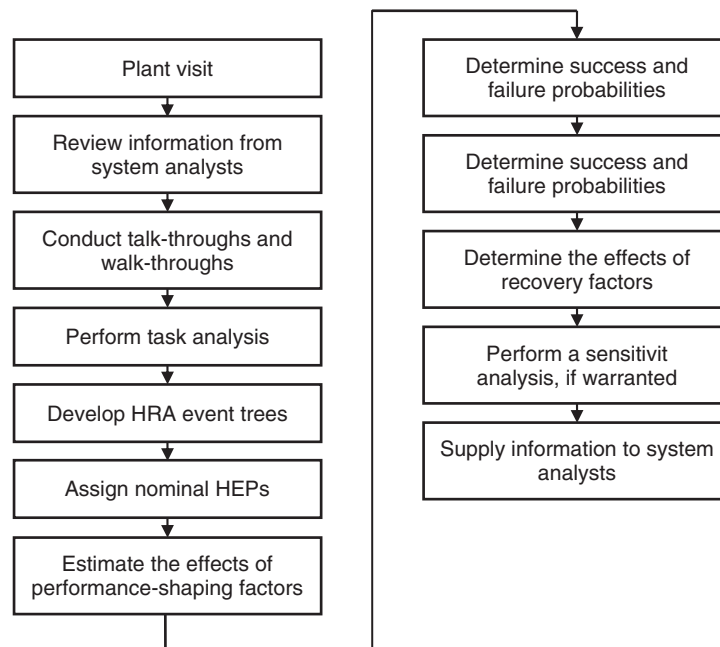


**Figure 10** Steps comprising THERP.

**Figure 11** HRA event tree corresponding to a nuclear power control room task that includes one recovery factor. (From Kumamoto and Henley, 1996. Copyright © 2004 by IEEE.)
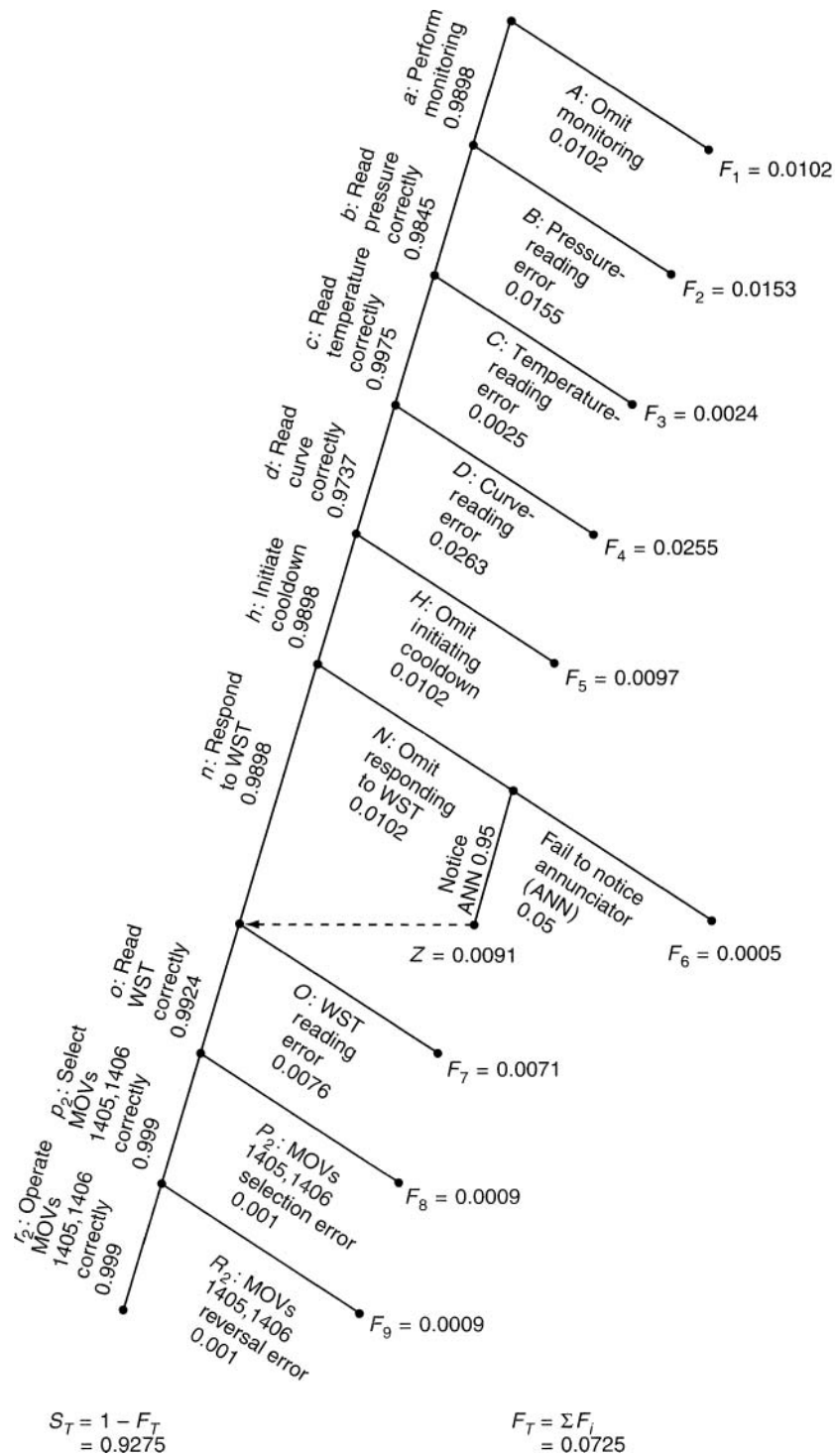
of nuclear power plant tasks. Naturally, as technologies evolve and procedures alter, the HEP values provided in such tables become less reliable.

For some tasks the nominal HEPs that are provided refer to *joint HEPs* because it is the performance of a team rather than that of an individual worker that is being evaluated. Generally, the absence of existing hard data from the operations of interest will require that nominal HEPs be derived from other sources, which include (1) expert judgment elicited through techniques such as direct numerical estimation or paired comparisons (Swain and Guttmann, 1983; Kirwan, 1994); (2) simulators (Gertman and Blackman, 1994); and (3) data from jobs similar in psychological content to the operations of interest.

To account for more specific individual, environmental, and task-related influences on performance, nominal HEPs are subjected to a series of refinements. First, nominal HEPs are modified based on the influence of PSFs, resulting in *basic HEPs* (BHEPs). In some cases, guidelines are provided in tables indicating the direction and extent of influence of particular PSFs on nominal HEPs. For example, adjustments that are to be made in nominal HEPs due to the influence of the PSF of stress are provided as a function of the characteristics of the task and the degree of worker experience.

Next, a nonlinear *dependency model* is incorporated which considers positive dependencies that exist between adjacent limbs of the tree, resulting in *conditional HEPs* (CHEPs). In a positive dependency model, failure on a subtask increases the probability of failure on the following subtask, and successful performance of a subtask decreases the probability of failure in performing the subsequent task element. Instances of negative dependence can be accounted for but require the discretion of the analyst. In the case of positive dependence, THERP provides equations for modifying BHEPs to CHEPs based on the extent to which the analyst believes dependencies exist. Five levels of dependency are considered in THERP: zero dependence, low dependence, medium dependence, high dependence, and complete dependence.

For example, assume the BHEP for task step B is $10^{-2}$ and a high dependence exists between task steps A and B. The CHEP of B given failure on step A would be given by the following equation for high dependence: CHEP $= (1 + \text{BHEP})/2 \sim 0.50$. Corresponding equations are given for computing CHEP under low- and medium-dependency conditions. For zero dependence the CHEP reduces to the BHEP ($10^{-2}$ in the example involving step B) and for complete dependence the CHEP would be 1 (failure on the prior task step assures failure on the subsequent step).

At this point, success and failure probabilities are computed for the entire task. Various approaches to these computations can be taken. The most straightforward approach is to multiply the individual CHEPs associated with each path on the tree leading to failure, sum these individual failure probabilities to arrive at the probability of failure for the total task, and then assign UCBs to this probability. More complex approaches to these computations take into account the variability associated with the combinations of events comprising the probability tree (Swain and Guttmann, 1983).

The final steps of THERP consider the ways in which errors can be recovered and the kinds of design interventions that can have the greatest impact on task success probability. Common recovery factors include the presence of annunciators that can alert the operator to the occurrence of an error, co-workers potentially capable of catching or discovering (in time) a fellow worker's errors, and various types of scheduled walkthrough inspections. As with conventional ETs, these recovery paths can easily be represented in HRA event trees (Figure 11). In the case of annunciators or inspectors, the relevant failure limb is extended into two additional limbs: one failure limb and one success limb. The probability that the human responds successfully to the annunciator or that the inspector spots the operator's error is then fed back into the success path of the original tree. In the case of recovery by fellow team members, BHEPs are modified to CHEPs by considering the degree of dependency between the operator and one or more fellow workers who are in a position to notice the error. The effects of recovery factors can be determined by repeating the computations for total task failure.

The analyst can also choose to perform *sensitivity analysis*. One approach to sensitivity analysis is to identify the most probable errors on the tree, propose design modifications corresponding to those task elements, estimate the degree to which the corresponding HEPs would become reduced by virtue of these modifications, and evaluate the effect of these design interventions on the computation of the total task failure probability. The final step in THERP is to incorporate the results of the HRA into system risk assessments such as PRAs.

## 4.4  HEART and NARA

The human error assessment and reduction technique (HEART) proposed by Williams (1988) was an HRA method that was directed at assessing tasks of a more holistic nature based on the assumption that human reliability is dependent upon the generic nature of the task to be performed. Thus the method was relatively easy to apply as, in comparison to THERP, it was not constrained to quantify large numbers of elemental subtasks.

In its emphasis on more holistically appraising the reliability of human task performance, HEART defines a limited set of "generic" tasks (GTs) describing nuclear power plant (NPP) activities from which the analyst can select from. Nominal HEPs (50th percentile) along with lower (5th percentile) and upper (90th percentile) bounds to these estimates are assigned to each of these tasks. For example, one of the generic tasks considered by HEART, together with its corresponding nominal HEP and associated lower and upper bounds is: "Shift or restore systems to a new or original state on a single attempt with supervision or procedures" (0.26; 0.14–0.42).

Although HEART, as in THERP, uses PSFs, referred to as error-producing conditions (EPCs), to modify HEPs, it applies a different approach to this process. In

its consideration of EPCs, HEART emphasizes the practical concern of reliability assessors with the potential for changes in the probability of failure of systems by an order of magnitude of 10. This "factor of 10" criterion is translated into a concern for identifying those EPCs that are likely to modify the probability of task failure by a factor of 3. In HEART's comprehensive listing of EPCs, each EPC is accompanied by an order of magnitude corresponding to the maximum amount by which the nominal HEP might change when considering the EPC at its worst relative to its best state. By providing a battery of remedial measures corresponding to each of the EPCs, HEART also offers a form of closure, by way of design considerations, to the issue of human contribution to system risk. Examples of five of the 38 EPCs along with their associated "orders of magnitude" are:

- A means of suppressing or overriding information or features which is too easily accessible ($\times 9$)
- A need to unlearn a technique and apply one which requires the application of an opposing philosophy ($\times 6$)
- No clear, direct, and timely confirmation of an intended action from the portion of the system over which control is to be exerted ($\times 4$)
- A mismatch between the educational achievement level of an individual and the requirements of the task ($\times 2$)
- No obvious way to keep track of progress during an activity ($\times 1.6$)

The process of computing HEPs in HEART first requires the HRA analyst to match a description of the situation for which a quantitative human error assessment is desired with one of the generic tasks. All relevant EPCs, especially those that satisfy the "factor of 3" criterion, are then identified. Next, the analyst must derive the weighting factor, $\mathrm{WF}_i$, associated with each $\mathrm{EPC}_i$, $i = 1, \ldots, n$, which requires assessing the proportion of the order of magnitude (APOM) associated with each EPC for the generic task being considered. The weighting factor is then defined as

$$\mathrm{WF}_i = [(\mathrm{EPC}_i \text{ order of magnitude} - 1) \times \mathrm{APOM} + 1)]$$

The HEP for the generic task, $\mathrm{GT}_{\mathrm{HEP}}$, is adjusted by multiplying this value by the product of all the weighting factors:

$$\mathrm{HEP} = \mathrm{GT}_{\mathrm{HEP}} \times \sum_{i=1}^{n} \mathrm{WF}_i.$$

NARA (nuclear action reliability assessment) represents a further development of HEART (Kirwan et al., 2005, cited in Spurgin, 2010). This method was partly motivated by concerns with the HEP values associated with the generic tasks in HEART as well as the vagueness of their description, which made the process of selecting generic tasks difficult. The primary differences between NARA and HEART are (1) the dependency on an improved database referred to as CORE-DATA

(Kirwan et al., 1999; Gibson et al., 1999) for HEP values; (2) the inclusion of a set of NARA tasks in place of the set of generic tasks in HEART; and (3) the incorporation of a human performance limit value to address concerns that enhancements in the reliability of human operations can, when human error terms are multiplied together, result in unreasonably low HEPs.

In NARA, the human tasks that are considered are categorized into one of four GT types: (1) type A comprises tasks related to task execution; (2) type B covers tasks related to ensuring correct plant status and availability of plant resources; (3) type C deals with responses to alarms and indicators; and (4) type D tasks involve communication behaviors. The tasks within these GT groups will often be linked so that responses to NARA tasks can come to define more complex tasks. For example, in the case of an accident the initial response may be to type C tasks, which could lead to situational assessment by the crew (type D tasks), and finally various types of execution (type A tasks), possibly following the availability of various systems and components (type B tasks).

Each task within each GT group has an associated HEP value, and EPCs are used, as in HEART, to modify the nominal HEP values. NARA, however, provides much more documentation on the use of EPCs, for example, in the form of anchor values and explanations of these values for each EPC, and guidance in determining the APOM values for each EPC.

## 4.5 SPAR-H

The standardized plant analysis risk human reliability analysis method (SPAR-H) was intended to be a relatively simple HRA method for estimating HEPs in support of plant-specific PRA models (Gertman et al., 2005). The method targets two task categories in NPP operations: action failures (e.g., operating equipment, starting pumps, conducting calibration or testing) and diagnosis activities (e.g., using knowledge and experience to understand existing conditions and making decisions). Although it differs from THERP in a number of its assumptions, THERP's underlying foundation is still very much apparent in SPAR-H. The method works as follows:

*Step* 1. Given an initiating event (e.g., partial loss of off-site power) and a description of the basic event being rated (e.g., the operator fails to restore one of the emergency diesel generators), the analyst must decide whether the basic event involves diagnosis, action, or both diagnosis and action. Guidance is provided to analysts for deciding between these three categories. In SPAR-H, the nominal HEP (NHEP) value assigned to a diagnosis failure is 0.01 and the NHEP assigned to an action failure is 0.001. These base failure rates are considered compatible with those from other HRA methods.

*Step* 2. SPAR-H considers eight PSFs. Each of these PSFs is described in terms of a number of operationally defined levels, including a nominal level.

Associated with each of these levels is a corresponding multiplier that determines the extent of the negative or positive effect the PSF has on the HEP. For some of the PSFs, the definitions of the levels depend on whether an action or a diagnosis is being considered. The eight PSFs are available time; stress/stressors; complexity; experience/training; procedures; ergonomics/human-machine interface; fitness for duty; and work processes. Some of these PSFs in and of themselves encompass a broad array of factors. For example, the PSF "complexity," which refers to how difficult the task is to perform in the given context, considers both task and environment-related factors. Task factors include requirements for a large number of actions, a large amount of communication, high degree of memorization, transitioning between multiple procedures, and mental calculations. The environment-related factors include the presence of multiple faults, misleading indicators, a large number of distractions, symptoms of one fault masking those of another, and ill-defined system interdependencies. It is important to note that in some schemes these could represent separate PSFs.

To illustrate how levels are operationally defined for a PSF, the four levels associated with the "procedures" PSF for an action task are:

- Not available: The procedure needed for the task in the event is not available.
- Task instructions, sections, or other needed information are not contained in the procedure.
- Available, but poor: A procedure is available but it contains wrong, inadequate, ambiguous, or other poor information.
- Nominal: Procedures are available and enhance performance.

In addition, all PSFs have an *insufficient information* level, which the analyst selects if there is insufficient information to enable a choice from among the other alternatives. The assignment of levels denotes ratings of PSFs that translate into multipliers that increase the nominal HEP (i.e., *negative ratings*) or decrease the nominal HEP (i.e., *positive ratings*). The idea that a PSF could reduce the nominal HEP is a departure from HRA methods such as THERP, but some other HRA methods also allow for this possibility (e.g., CREAM, Section 4.9).

*Step* 3. Although the eight PSFs are clearly non-orthogonal, with complex relationships assumed to exist between several of the PSFs, SPAR-H treats these influencing factors as if they were mutually independent. Consequently, to help prevent the analyst from "double counting" when assigning values to the PSFs for the purpose of modifying the nominal task HEP, SPAR-H provides a 64-cell table that contains the presumed degree of correlation among the PSFs based on qualitative rankings of low, medium, or high.

To obtain a composite PSF value, the ratings of the eight PSFs are multiplied by one another, regardless of whether the PSF influence is positive or negative. The HEP is then computed as the product of the composite PSF ($PSF_{composite}$) and the NHEP. Because of the independence assumption and the values ($>1$) that negative PSF ratings can assume, when three or more PSFs are assigned negative rankings, there is a relatively high probability that the resultant HEP would exceed 1.0. In most HRAs, the HEP is simply rounded down to 1.0. To decrease the possibility for HEP values exceeding 1.0, SPAR-H uses the following formula for adjusting the nominal HEP in order to compute the HEP, where NHEP equals 0.01 for diagnosis tasks and 0.001 for action tasks:

$$HEP = \frac{NHEP \times PSF_{composite}}{NHEP \times (PSF_{composite} - 1) + 1}$$

As an example, assume a diagnosis activity (at a nuclear power plant) is required and a review of the operating event revealed that the following PSF parameters were found to have influenced the crew's diagnosis of "loss of inventory": Procedures were misleading; displays were not updated in accordance with requirements; and the event was complex due to the existence of multiple simultaneous faults in other parts of the plant. Assuming these were the only influences contributing to the event, the assignment of the PSF levels and associated multipliers would be:

| PSF | Status | Multiplier |
|---|---|---|
| Procedures | Misleading | ×10 |
| Ergonomics | Poor | ×20 |
| Complexity | Moderately complex | ×2 |

The PSF composite score would be $10 \times 20 \times 2 = 400$. Without an adjustment on the NHEP, the HEP would be computed as NHEP × $PSF_{composite} = HEP = 4.0$. Use of the adjustment factor produces

$$HEP = \frac{0.01 \times 400}{0.01 \times (400 - 1) + 1} = 0.81$$

The adjustment factor can also be applied when a number of positive influences of PSFs are present. In this case, the multiplication factors associated with the "positive" levels of the PSFs would be less than 1.0. However, the SPAR-H PSFs are negatively skewed, so that they have a relatively larger range of influence for negative as compared to positive influences.

*Step* 4. In cases where a series of activities are performed, it is possible that failure on one activity (A) can influence the probability of error on the subsequent activity (B). In THERP, a dependency model consisting of five levels of dependency ranging from no dependency to complete dependency is used to account for such situations

(Section 4.3). SPAR-H also uses these dependency levels. In addition, SPAR-H makes use of a number of factors that can promote dependency between errors in activities performed in series (such as whether the crew is the same or different or whether the current task is being performed close in time to the prior task) to construct a dependency matrix. This matrix yields 16 dependency rules that map, correspondingly, to the four levels where some degree of dependency exists; a 17th rule is used to account for the case of no dependency. The modifications to nominal HEPs resulting from these levels of dependency follow the same procedure as in THERP.

*Step* 5. SPAR-H deals with the concept of uncertainty regarding the HEP estimate, which is the basis for producing lower and upper bounds on the error, very differently than THERP. Whereas THERP assumes HEPs derive from a lognormal probability distribution and uses error factors to derive the lower (5th percentile) and upper (95th percentile) bounds on the error estimate based on this distribution, SPAR-H does not assume a lognormal distribution of HEPs nor does it use error factors.

Instead, SPAR-H uses a "constrained noninformative prior" (CNI) distribution, where the constraint is that the prior distribution (i.e., "starting-point" distribution) has a user-specified mean (which is the product of the composite PSFs and the nominal HEP). The reasons for using this distribution are (1) it takes the form of a beta distribution for probability-type events, which is a distribution that has the flexibility to mimic normal, lognormal, and other types of distributions; (2) unlike THERP, it does not require uncertainty parameter information such as a standard deviation; and (3) it can produce small values at the lower end of the HEP distribution (e.g., $<1 \times 10^{-6}$) but will more properly represent expected error probability at the upper end.

Once the mean HEP is known (i.e., the product of the composite PSFs and the nominal HEP), the starting-point CNI distribution can be transformed to an approximate distribution that is based on the beta distribution. This requires deriving two parameters: $\alpha$ and $\beta$. Tables can be used to obtain values of $\alpha$ for a given value of the mean (which is the HEP), and then this value of $\alpha$ together with the mean can be used to compute $\beta$ using the formula $\alpha(1 - \text{HEP})/\text{HEP}$. Once values of $\alpha$ and $\beta$ are known, various mathematical analysis packages can be used to compute the 5th, 95th, or any percentile desired for the HEP.

*Step* 6. To ensure analyst consistency in using this method, SPAR-H provides designated worksheets to guide the analyst through the entire process required to generate the HEP.

## 4.6 Time-Related HRA Models

HRA models based on time–reliability curves, sometimes referred to as time–reliability correlations (TRCs),

are concerned with the time it takes for a crew to respond to an emergency or accident. The most well-known TRC was the human cognitive reliability (HCR) model (Hannaman et al., 1984) that was sponsored by EPRI based on data obtained from prior simulator studies.

Let $P(t)$ denote the nonresponse probability by a crew to a problem within a given time window $t$, where $t$ is estimated based on analysis of the event sequence following the stimulus. According to the HCR model, this probability can be estimated using a three-parameter Weibull distribution function, a type of distribution applied in equipment reliability models, of the form

$$P_t = e\left[-\left\{\frac{(t/T_{1/2}) - B}{A}\right\}^C\right]$$

where $T_{1/2}$ is the estimated median time to complete the action(s) and $A$, $B$, and $C$ are coefficients associated or "correlated" with the level of cognitive processing required by the crew. Specifically, their values, following Rasmussen's SRK model (Section 2.2.4), depend on whether task performance is occurring at the skill-, rule-, or knowledge-based level.

The variable $t/T_{1/2}$ represents "normalized time," which controls for contributions to crew response times that are unrelated to human activities (Figure 12). Obtaining this normalized time requires defining a "nominal" median response time, $T^n_{1/2}$, which is the time corresponding to a probability of 0.5 that the crew successfully carries out the required task(s) under nominal conditions. Nominal median response times are typically derived from simulator data and talk-throughs with operating crews. The actual (estimated) median response time $T_{1/2}$ is computed from the nominal median response time $T^n_{1/2}$ by

$$T_{1/2} = (1 + K_1)(1 + K_2)(1 + K_3)T^n_{1/2}$$

where $K_1$, $K_2$, and $K_3$ are coefficients whose values depend on PSFs. The HCR model thus assumes that PSFs impact the median response time rather than affect the type of cognitive processing, so that the relationships between the three types of curves remain preserved. The derivation of the estimated median response time from the nominal median response time is illustrated in the following example taken from Kumamoto and Henley (1996, p. 492).

Consider the task of detecting that a failure of an automatic plant shutdown system has occurred. The nominal median response time is 10 s. Assume average operator experience ($K_1 = 0.00$) under potential emergency conditions ($K_2 = 0.28$) with a good operator/plant interface in place ($K_3 = 0.00$). The actual median response time is then estimated to be

$$T_{1/2} = (1 + 0.00)(1 + 0.28)(1 + 0.00)(10) = 12.8 \text{ s}$$

Continuing with this example, assume that the initiating event was loss of feedwater to the heat exchanger that cools a reactor and, due to the failure of the automatic plant shutdown system, manual plant shutdown by the crew is called for. Suppose the crew must
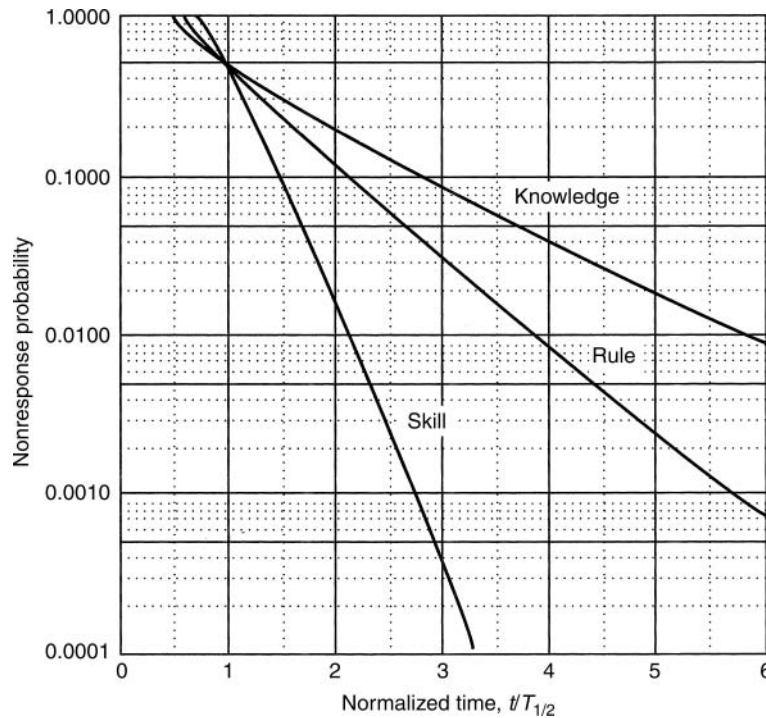
**Figure 12**  HCR model curves. (From Kumamoto and Henley, 1996. Copyright © 2004 by IEEE.)

complete the plant shutdown within 79 s from the start of the initiating event. This time window encompasses not only the time to detect the event, for which the nominal median response time was 10 s, but also diagnosis of and response to the event.

For this example, assume the nature of the instrumentation enables easy diagnosis by control room personnel of the loss-of-feedwater accident and the automatic shutdown system failure, resulting in a nominal median diagnosis time of 15 s. Also, assume errors due to slips (e.g., unintentional activation of an incorrect control) for this procedure are judged to be negligible given the operator–system interface design, so that the nominal median response time can be considered to be 0 s. The total nominal median response time for the shutdown procedure would then be $10 \, \text{s} + 15 \, \text{s} = 25 \, \text{s}$ and, using the $K$ values above, would result in an actual median response time $T_{1/2} = 1.28 \times 25 \, \text{s} = 32 \, \text{s}$. With the level of performance assumed to be at the skill-based level, the corresponding parameter values in the HCR model are $A = 0.407, B = 0.7$, and $C = 1.2$, resulting in a probability that the crew successfully responds to this initiating event within the 79-s window of

$$P[t \leq 79] = e\left[-\left(\frac{(79/32) - 0.7}{0.407}\right)^{1.2}\right] = 0.0029$$

With $K_2 = 0$ (i.e., an optimal stress level), the nonresponse probability would be reduced to 0.00017 per demand for manual shutdown.

The HCR model can also be used to obtain estimates of human or crew failures for more complex operations. Continuing with the example above, the analyst may assume successful plant shutdown occurs but may now be interested in assessing the risks associated with accident recovery, which requires removing heat from the reactor before damage to the reactor is incurred. Suppose there are two options (see below) for coping with the loss of feedwater accident and three different strategies for combining these two options (Kumamoto and Henley, 1996). The different strategies may not only have different time windows but may also involve different levels of cognitive processing as well as distinctive auxiliary human operations. These operations would have associated error probabilities whose values would need to be combined with the results from the HCR model to provide total probabilities of failure for any given strategy.

For example, one strategy, the "anticipatory" strategy, assumes that the crew has concluded that recovery of feedwater through the secondary heat removal system (option 1) is not feasible and decide to establish "feed-and-bleed" (option 2), which involves manually opening pressure-operated relief valves (PORVs) and activation of high-pressure injection (HPI) pumps to exhaust heat to the reactor containment housing. There is a 60-min time window available for establishing the feed-and-bleed operation before damage to the core occurs. Because this operation requires 1 min, the effective time window is reduced to 59 min. If one

assumes well-trained operators ($K_1 = -0.22$), a grave emergency stress level ($K_2 = 0.44$), and good operator interface ($K_3 = 0.00$); a knowledge-based level of performance ($A = 0.791$, $B = 0.5$, and $C = 0.8$); and a nominal median response time of 8 s, $P[t \leq 59] = 0.006$. Assuming the HEP for the manipulation of the PORVs and HPI is 0.001, the HEP for the feed-and-bleed operation is then $0.006 + 0.001 = 0.007$, implying a success probability of 0.993.

However, the success of the anticipatory strategy also hinges on following the feed-and-bleed operation with successful alignment of the heat removal valve system. Kumamoto and Henley (1996) provide a *human reliability fault tree* (in which all the events in the FT are human action events) that computed this top event failure to be 0.0005. Taking a human reliability event tree approach to computing the probability of failure of the anticipatory strategy, this probability has two *failure paths*: (1) failure to perform the feed-and-bleed operation (0.007) and (2) assuming successful performance of this operation (0.993), failure to perform alignment of the heat removal valve system, which is $0.993 \times 0.0005$, resulting in an anticipatory strategy failure of $0.007 + (0.993 \times 0.0005) = 0.0075$.

The validity of the HCR model has been questioned by Apostolakis et al. (1988), who raised the issue of whether the normalized response times for all tasks can be modeled by a Weibull or any other single distribution and the issue of identifying the correct curve due to the fact that many tasks cannot be characterized exclusively as skill, rule, or knowledge based. Following the development of the HCR model, EPRI sponsored a large simulator data collection project (Spurgin et al., 1990) that, in fact, did not confirm a number of the underlying hypotheses associated with HCR model performance.

## 4.7 SLIM

The HRA method SLIM refers to the success likelihood index (SLI) methodology developed by Embrey et al. (1984) for deriving HEPs for specified human actions in NPP operations, although the method is generally believed to be equally as applicable to other industries. SLIM allows the analyst to derive HEPs for relatively low-level actions that cannot be further decomposed as well as for more broadly defined holistic actions that encompass many of these lower level actions.

Underlying SLIM are two premises: (1) that the probability a human will carry out a particular task successfully depends on the combined effect of a number of PSFs and (2) that these PSFs can be identified and appropriately evaluated through expert judgment. For each action under consideration, SLIM requires that domain experts identify the relevant set of PSFs; assess the relative importance (or weights) of each of these PSFs with respect to the likelihood of some potential error mode associated with the action; and, independent of this assessment, rate how good or bad each PSF actually is within the context of task operations.

The first step in SLIM consists of identifying (through the use of experts) the potential error modes associated with human actions of interest and the PSFs most relevant to these error modes. The identification of all possible error modes is generally arrived at through in-depth analysis and discussions that could include task analysis and reviews of documentation concerning operating procedures.

Next, relative-importance weights for the PSFs are derived by asking each analyst to assign a weight of 100 to the most important PSF and then assign weights ranging from 0 to 100 to each of the remaining PSFs based on the importance of these PSFs relative to the one assigned the value of 100. Discussion concerning these weightings is encouraged in order to arrive at consensus weights. Normalized weights are then derived by dividing each weight by the sum of the weights for all the PSFs.

The expert judges then rate each PSF on each action or task, with the lowest scale value indicating that the PSF is as poor as it is likely to be under real operating conditions and the highest scale value indicating that the PSF is as good as it is likely to be in terms of promoting successful task performance. The range of possible SLI values is dictated by the ranges of values associated with the rating scale. As with the procedure for deriving weights, the individual ratings should be subjected to discussion in order to arrive at consensus ratings. The likelihood of success for each human action or task is determined by summing the product of the normalized weights and ratings for each PSF, resulting in numbers (SLIs) that represent a scale of success likelihood.

To illustrate the process by which SLIs are computed, four human actions from the task analysis of the chlorine tanker filling task (Table 6) taken from CCPS (1994) will be considered as indicated in Table 11. When identifying

**Table 11 PSF Ratings, Rescaled Ratings (in parentheses), and SLIs for Chlorine Tanker Filling Example**

| Human Actions | Performance-Shaping Factors | | | | SLIs |
|---|---|---|---|---|---|
| | Time Stress | Experience | Distractions | Procedures | |
| Close test valve (2.1.3) | 4 (0.63) | 8 (0.88) | 7 (0.25) | 6 (0.63) | 0.54 |
| Close tanker valve 4.1.3 | 8 (0.13) | 8 (0.88) | 5 (0.50) | 6 (0.63) | 0.41 |
| Secure locking nuts (4.4.2) | 8 (0.13) | 7 (0.75) | 4 (0.63) | 2 (0.13) | 0.34 |
| Secure blocking device (4.2.3) | 8 (0.13) | 8 (0.88) | 4 (0.63) | 2 (0.13) | 0.35 |
| PSF weights | 0.4 | 0.1 | 0.3 | 0.2 | |

*Source*: Adapted from CCPS (1994). Copyright 1994 by the American Institute of Chemical Engineers. Reproduced by permission of AICHE.

tasks or actions that will be subjected to analysis by SLIM, it is constructive to group activities that are likely to be influenced by the same PSFs, which is considered a legitimate assumption for this set of tasks.

In this example, the main PSFs which determine the likelihood of error are assumed to be time stress, level of operator experience, level of distractions, and quality of procedures. The consensus normalized weights arrived at for these four PSFs are 0.4, 0.1, 0.3, and 0.2, implying that for these tasks time stress is most influential and experience level has the least influence on errors.

Each task is then rated on each PSF. A numerical scale from 1 to 9 will be used, where 1 and 9 represent either best or worst conditions. For the PSFs time stress and distractions, ratings of 9 would represent high levels of stress and distractions and imply an increased likelihood of errors; ratings of 1 would be ideal for these PSFs. In contrast, high ratings for experience and procedures would imply decreased likelihood errors; in the case of these two PSFs, ratings of 1 would represent worst-case conditions. The ratings assigned to each of the four activities are given in Table 11.

To calculate the SLIs, the data in Table 11 are rescaled to take into account the fact that the ideal point (IP) is at different ends of the rating scale for some of the PSFs (either 1 or 9). Rescaling will also serve to convert the range of ratings from 1–9 to 0–1. The formula used to convert the original ratings to rescaled ratings is

$$1 - \frac{ABS\,(R - IP)}{8}$$

where ABS represents the absolute value operator, $R =$ original rating, and IP is the ideal value for the PSF being considered. When the rating is either 1 or 9, this formula converts the original rating to 0.0 or 1.0, as appropriate. The rescaled ratings are shown in parentheses next to the original ratings. Finally, an additive model is assumed whereby the SLI for each task $j$ in Table 11 is calculated using the expression

$$\mathrm{SLI}_j = \sum_{i=1}^{4} \sum_{j=1}^{4} R_{ij}\, \mathrm{PSFw}_i$$

where $\mathrm{PSFw}_j$ is the weight assigned to the $i$th PSF and $R_{ij}$ is the rescaled rating of the $j$th task on the $i$th PSF.

The SLIs represent a measure of the likelihood that the task operations will succeed or fail relative to one another and are useful in their own right. For example, if the actions under consideration represent alternative modes of response in an emergency scenario, the analyst may be interested in determining which types of responses are least or most likely to succeed. However, for the purpose of conducting PRAs, SLIM converts the SLIs to HEPs.

Converting the SLI scale to an HEP scale requires some form of calibration process. In practice, if a large number of tasks in the set being evaluated have known probabilities of error, for example, from internal or industrywide incident data, then the regression equation resulting from the best fitting regression line between the SLI values and their corresponding HEPs can be

used to compute HEPs for other operations in the group for which HEPs are not available. Typically, data sets that enable an empirical relationship between SLIs and HEPs to be computed are not available, requiring the assumption of some form of mathematical relationship. One such assumption is the following loglinear relationship (where logs to base 10 are used) between HEPs and SLIs:

$$\log\,[\mathrm{HEP}] = a \times \mathrm{SLI} + b$$

where $a$ and $b$ are constants. This assumption is partly based on experimental evidence that has indicated a loglinear relationship between factors affecting performance on maintenance tasks and actual performance on those tasks (CCPS, 1994).

To compute the constants in this equation, at least two tasks with known SLIs and HEPs must be available in the set of tasks being evaluated. Continuing with the chlorine tanker filling example, assume evidence was available for arriving at the following HEP estimates for two of the four tasks being evaluated:

- Probability of test valve being left open: $1 \times 10^{-4}$
- Probability of locking nuts not being secured: $1 \times 10^{-2}$

The substitution of these HEP values and their corresponding SLIs into the loglinear equation produces the calibration equation

$$\log\,[\mathrm{HEP}] = -2.303 \times \mathrm{SLI} + 3.166$$

from which the HEPs for the remaining two tasks in the set can be derived:

- Probability of not closing tanker valve: $1.8 \times 10^{-3}$
- Probability of not securing blocking device: $7.5 \times 10^{-3}$

As in THERP, the impact of design interventions can be examined through sensitivity analysis. However, in SLIM, the sensitivity analyses that are performed are based on evaluating the effects of the interventions on PSFs, which result in new SLIs and ultimately new HEPs that can be compared to previous values. In this way, what-if analyses can be used to explore potential design modifications for the purpose of determining which resource allocation strategies provide the greatest reductions in risk potential.

In the absence of HEP data, the calibration values would have to be generated by expert judgment. In these cases, for each task, each expert can be asked to make absolute judgments of the probability of failure associated with two boundary conditions corresponding to situations where the PSFs are as good and as bad as they could credibly be under real operating conditions. These judgments are facilitated through the use of a logarithmic probability scale and are assigned values of 100 and 0, respectively. The SLI computed for a given task is then used to interpolate between these lower

bound (LB) and upper bound (UB) probabilities, which are preferably obtained through consensus, resulting in the following estimate of the HEP for each task:

$$\text{HEP} = \text{LB}^{\text{SLI}/100} \times \text{UB}^{(1-\text{SLI})/100}$$

As PRAs typically require that measures of uncertainty accompany HEP estimates, this direct-estimation approach can also be used to derive these lower and upper uncertainty bounds for the HEP estimates derived by SLIM. In using this approach, the analyst must ensure that the question posed to the expert concerns identifying upper and lower bounds for the HEP such that the true HEP falls between these bounds with 95% certainty.

A user-friendly computer-interactive environment for implementing SLIM, referred to as multi-attribute utility decomposition (MAUD), has been developed which can help ensure that many of the assumptions that are critical to the theoretical underpinnings of SLIM are met. For example, MAUD can determine if the ratings for the various PSFs by a given analyst are independent of one another and whether the relative-importance weights elicited for the PSFs are consistent with the analyst's preferences. In addition, MAUD provides procedures for assisting the expert in identifying the relevant PSFs.

### 4.8 Holistic Decision Tree Method

The holistic decision tree (HDT) method developed by Spurgin (2010) was directed at determining how the context humans find themselves operating in during accident scenarios impacts their failure probability. A detailed example of its application for HRA in various International Space Station (ISS) accident scenarios is given in Spurgin (2010).

Determining the context under which personnel are operating during various ISS accident scenarios requires understanding the relationship between two groups of personnel associated with direct ISS operations: (1) astronauts/cosmonauts, who need to respond to accidents requiring rapid action, control experiments, engage in maintenance activities, and support flight controllers operating from the ground in detecting system anomalies, and (2) flight controllers, who are responsible for monitoring and controlling the ISS systems remotely and on occasion must engage astronauts in debugging activities as controllers are limited in the amount of information available to them.

Some of the concepts associated with the HDT method are related to SLIM and HEART/NARA, especially its emphasis on identifying PSFs as a basis for characterizing the contexts in which humans operate and in evaluating the quality of those PSFs for a given scenario. It also assumes, as in SLIM, a loglinear relationship between an HEP and PSFs. In the HDT method, PSFs are referred to as influence factors (IFs); the ratings of those IFs are referred to as quality values (QVs); and the descriptions on which these QVs are based are referred to as quality descriptors (QDs). As in SLIM, importance weights (i.e., relative rankings) are also determined for each of the IFs. The following steps summarize the process of applying the HDT method (Spurgin, 2010):

*Step* 1. First, a list of potential IFs needs to be identified, which will require HRA analysts becoming familiar with ISS operations. This process will entail detailed reviews of simulator training programs, astronaut operations, and flight controller operations as well as interviews with training staff, astronauts, and controllers. Witnessing training sessions covering simulated accidents is essential. In the ISS study, 43 IFs were initially identified which were ultimately reduced to 6 through interaction with ISS personnel.

*Step* 2. The list of IFs is sorted into scenario-dependent (IFs specific to a particular scenario) or global IFs (which are present in every scenario), as the assumption is that HEPs would be affected by both types of IFs. In the ISS study, all six IFs ultimately identified were global IFs; thus these same IFs were used for all the scenarios considered.

*Step* 3. IFs are then ranked in order of importance, and the most important ones are selected. In this example, the 6 (of the 43) IFs considered (by consensus) to be most important were (1) quality of communication; (2) quality of man–machine interface; (3) quality of procedures; (4) quality of training; (5) quality of command, control, and decision making (CC&DM); and (6) degree of workload. Although these IFs may be very broadly defined, for the purposes of evaluating their QVs, comprehensive yet concise definitions that are clearly linked to the scenarios being evaluated need to be provided for each of these IFs. Examples of these definitions are given by Spurgin (2010).

*Step* 4. Prior to rating the quality of the IFs, QDs need to be defined. In the HDT method, each IF has three possible quality levels, whose descriptions will depend on the IF. For example, in the ISS study, the descriptors for the CC&DM IF were "efficient," "adequate," and "deficient"; for the "quality of procedures" IF the descriptors were "supportive," "adequate," and "adverse"; and for the workload IF the descriptors were "more than capacity," "matching capacity," and "less than capacity." For any given IF, the QDs need to be explicitly defined. For example, for the CC&DM IF, the QD "Deficient" is defined as follows: "Collaboration between (team) members interferes with the ability to resolve problems and return to a desired state of the system."

*Step* 5. Importance weights are derived for each IF. In the HDT method, these weights are obtained through use of the *analytic hierarchy process* (AHP), a mathematical technique developed by Saaty (1980) that has been applied to a wide variety of decision problems. This method requires that each rater rank the relative importance (or preference) of each IF as compared to every other IF. When making these paired comparisons, each IF is given a value from 1 to 9; for example, for a given scenario the relative ranking of communication to procedures may be 6

to 3. The AHP method is amenable to aggregating paired-comparison data from groups of raters and provides estimates of the variability associated with each IF weight that is derived. Among its other advantages are its ability to quantify (and thus ensure) consistency in human judgments; provide empirical results in the absence of statistical assumptions regarding the distribution of human judgments; and its relative ease of administration. In the ISS study, importance weights were derived for five different scenarios for which human performance failure probabilities were of interest: docking, fire, coolant leak, loss of CC&DM, and extravehicular activity (astronauts working in space suits outside the ISS).

*Step* 6. Upper and lower anchor values for the scenario HEP are determined. The upper anchor value may often be assumed to be 1.0, which implies that if all IFs that impact human performance are as poor as possible, it is almost certain that humans will fail. The lower anchor is typically subject to greater variability; thus, each scenario is likely to be assigned a different distribution for the lower anchor value. Spurgin (2010) notes that for most scenarios the 5th percentile of this anchor is set at $10^{-4}$ and the 95th percentile is set at $10^{-3}$; however, for more severe cases, these lower and upper bounds can be set to $10^{-2}$ and 1.0, respectively.

*Step* 7. Using the QDs, each IF is rated for each scenario. In rating each IF, a QV of 1, 3, or 9 is assigned, where 1 represents a "good" quality description, 3 represents a "fair" quality description, and 9 represents a "poor" quality description. Thus a factor of 3 was used to represent ordinal-scale transitions from good to fair and from fair to poor.

*Step* 8. At this point, a decision tree (DT) can be constructed that, for a given scenario, captures the IFs, the importance weights of these IFs, and the three QVs assigned to these IFs. For the six IFs in this example there are a total number of $3^6 = 729$ different paths through this tree, and each path will result in a unique HEP for that scenario. To determine the HEP for a given scenario, the pathway corresponding to the set of QVs that were assigned is located.

*Step* 9. The distribution of HEP values for the different pathways in the DT is derived. For example, consider the portion of the HDT depicted in Figure 13 for a coolant loop leak scenario. The HEPs in the end branches of this tree are computed with the aid of a Microsoft Excel spreadsheet based on the relationship between the IF importance weights, the QVs, and the anchor values. The program provides the ability for quantifying the variance in each IF importance weight and the variances in the lower bound and upper bound anchor values. In Figure 13, the low (anchor) HEP is at the top end branch; increasingly higher values occur as one descends the

tree. In the HDT method, as in SLIM, the log of HEP as a function of IFs is computed, from which HEP values are readily calculated. The HDT method uses the upper and lower HEP anchor values as a basis for deriving HEPs, with the precise expressions used as follows:

$$\ln(\text{HEP}_i) = \ln(\text{HEP}_l) + \ln\left(\frac{\text{HEP}_h}{\text{HEP}_l}\right)\left[\frac{S_i - S_l}{S_h - S_l}\right]$$

$$S_i = \sum_{j=1}^{n}(QV_j)I_j$$

where

$$\sum_{j=1}^{n}I_j = 1$$

In these expressions, $\text{HEP}_i$ is the human error probability of the $i$th pathway through the HDT; $\text{HEP}_l$ is the low HEP anchor value; $\text{HEP}_h$ is the high HEP anchor value; $S_l$ is the lowest possible value of $S_i$ (which equals 1 in the current formulation of QVs for IFs); $S_h$ is the highest possible value of $S_i$ (which equals 9 in the current formulation of QVs for IFs); $QV_j$ is the quality descriptor value (1, 3, or 9 in the current formulation) corresponding to the $j$th IF; and $I_j$ is the importance weight of the $j$th IF.

Overall, the HDT method, like SLIM, is a very flexible method that can be easily adapted to many different types of applications. Also, like SLIM, the impact of changes to influencing factors, which reflects changes in the contexts in which humans operate, can be easily explored and used to assess cost–benefit trade-offs associated with proposed design interventions. However, like SLIM, its success depends on the rigor and skill employed in collecting relevant information regarding operations and the impact of contextual factors on these operations and on generating and managing expert judgments from qualified personnel.

## 4.9  CREAM

The HRA method known as CREAM (cognitive reliability and error analysis method) was developed by Hollnagel (1998). CREAM distinguishes between two methods: a *basic method* and an *extended method*. Both methods result in estimates of the probability of performing an action (either a task as a whole or a segment of a task) incorrectly; in the basic method this estimate is referred to as a general action failure probability and in the extended method this estimate is referred to as a specific action failure probability.

Prior to presenting the basic method, an additional concept that is fundamental to CREAM needs to be introduced, namely, the notion of *control mode*. Hollnagel (1998) has suggested four control modes that are considered important for performance prediction: scrambled control, opportunistic control, tactical control, and strategic control. These different levels of control are influenced by the context as perceived by the person

**HDT for ISS PRA Scenario**  | 3 | Coolant loop leak

Columns (diagonal headers): Communications | Procedures | Training | Man–machine interface | Workload | Command, control & decision making | Weighted sums | HEP Result

IF weights ====>

| 0.19 | 0.21 | 0.16 | 0.17 | 0.13 | 0.14 |
|------|------|------|------|------|------|

| Branch path | Weighted sums | HEP Result |
|-------------|---------------|------------|
| Enter  1  1  1  1  1  1 | 1.00E+00 | 4.04E-04 |
| 3 | 1.28E+00 | 4.90E-04 |
| 9 | 2.12E+00 | 8.73E-04 |
| 3  1 | 1.26E+00 | 4.83E-04 |
| 3 | 1.54E+00 | 5.86E-04 |
| 9 | 2.83E+00 | 1.04E-03 |
| 9  1 | 2.04E+00 | 8.26E-04 |
| 3 | 2.32E+00 | 1.00E-03 |
| 9 | 3.16E+00 | 1.79E-03 |
| 3  1  1 | 1.34E+00 | 5.10E-04 |
| 3 | 1.62E+00 | 6.19E-04 |
| 9 | 2.46E+00 | 1.10E-03 |
| 3  1 | 1.60E+00 | 6.10E-04 |
| 3 | 1.88E+00 | 7.40E-04 |
| 9 | 2.72E+00 | 1.32E-03 |
| 9  1 | 2.38E+00 | 1.04E-03 |
| 3 | 2.66E+00 | 1.27E-03 |
| 9 | 3.50E+00 | 2.26E-03 |
| 9  1  1 | 2.36E+00 | 1.03E-03 |
| 3 | 2.64E+00 | 1.25E-03 |
| 9 | 3.48E+00 | 2.23E-03 |
| 3  1 | 2.62E+00 | 1.23E-03 |
| 3 | 2.90E+00 | 1.49E-03 |
| 9 | 3.74E+00 | 2.67E-03 |
| 9  1 | 3.40E+00 | 2.11E-03 |
| 3 | 3.68E+00 | 2.56E-03 |
| 9 | 4.52E+00 | 4.56E-03 |
| 3  1  1  1 | 1.32E+00 | 5.03E-04 |
| 3 | 1.60E+00 | 6.10E-04 |
| 9 | 2.44E+00 | 1.09E-03 |
| 3  1 | 1.58E+00 | 6.02E-04 |
| 3 | 1.86E+00 | 7.30E-04 |
| 9 | 2.70E+00 | 1.30E-03 |
| 9  1 | 2.36E+00 | 1.03E-03 |
| 3 | 2.64E+00 | 1.25E-03 |
| 9 | 3.48E+00 | 2.23E-03 |
| 3  1  1 | 1.66E+00 | 6.36E-04 |
| 3 | 1.94E+00 | 7.71E-04 |
| 9 | 2.78E+00 | 1.38E-03 |
| 3  1 | 1.92E+00 | 7.61E-04 |
| 3 | 2.20E+00 | 9.23E-04 |
| 9 | 3.04E+00 | 1.65E-03 |
| 9  1 | 2.70E+00 | 1.30E-03 |
| 3 | 2.98E+00 | 1.58E-03 |
| 9 | 3.82E+00 | 2.82E-03 |
| 9  1  1 | 2.68E+00 | 1.28E-03 |
| 3 | 2.96E+00 | 1.56E-03 |
| 9 | 3.80E+00 | 2.78E-03 |
| 3  1 | 2.94E+00 | 1.54E-03 |
| 3 | 3.22E+00 | 1.86E-03 |
| 9 | 4.06E+00 | 3.32E-03 |
| 9  1 | 3.72E+00 | 2.63E-03 |
| 3 | 4.00E+00 | 3.19E-03 |
| 9 | 4.84E+00 | 5.69E-03 |
| 9  1  1  1 | 2.28E+00 | 9.75E-04 |
| 3 | 2.56E+00 | 1.18E-03 |

**Quality factors**

| | IF 1 (Com) | IF 2 (Proc) | IF 3 (Trng) | IF 4 (MMI) | IF 5 (Work) |
|------|------|------|------|------|------|
| Rating==> | G | F | F | F | F |
| Good | 1 | 1 | 1 | 1 | 1 |
| Fair | 3 | 3 | 3 | 3 | 3 |
| Poor | 9 | 9 | 9 | 9 | 9 |

**The order of IF in the tree does not matter in this model because using path-dependent factors.**

Indicates input cell

Low HEP anchor   4.04E-04
High HEP anchor   1.00E+00

The IF weight sum to unity. They are used as weighting factors on
The quality factors are a factor that measures the increase or decre
**A quality factor larger than 1 indicates an increase in HEP.**
Model is normalized such that if all influence factors are "good," then
HEP is achieved, and if all influence factors are "poor," then the maxi
HEP is achieved.

| Legend | | |
|--------|--|--|
| low= | Low HEP anchor | Mark scenario 2 in OR |
| high= | High HEP anchor | Mark scenario 3 in GR |
| lowws= | Lowest weighted sum | Mark scenario 4 in GR |
| highws= | Highest weighted sum | Mark scenario 5 in BLI |

**Figure 13**  Representation of a portion of a holistic decision tree for a coolant loop leak scenario. (From Spurgin, 2010. Copyright 2010 with permission from Taylor & Francis, 2010.)

(e.g., the person's knowledge and experience concerning dependencies between actions) and by expectations about how the situation is going to develop. The distinctions between these four control modes are briefly described as follows:

- *Scrambled Control*. In this mode, there is little or no thinking involved in choosing what to do; human actions are thus unpredictable or haphazard. This usually occurs when task demands are excessive, the situation is unfamiliar and changes in unexpected ways, and there is a loss of situation awareness.

- *Opportunistic Control*. The person's next action is based on the salient features of the current context as opposed to more stable intentions or goals. There is little planning or anticipation.

- *Tactical Control*. The person's performance is based on planning and is thus driven to some extent by rules or procedures. The planning, however, is limited in scope.

● *Strategic Control*. The person considers the global context, using a wider time horizon, and takes into account higher level performance goals.

Initially, as in any HRA method, a task or scenario that will be the subject of the analysis by CREAM needs to be identified. Consistent with most PRA studies, this information is presumably available from lists of failures that can be expected to occur or is based on the requirements from the industry's regulatory body. Following identification of the scenario to be analyzed, the steps comprising the *basic method* of CREAM are as follows.

*Step* 1. The first step involves performing a task analysis (Section 3.1). The TA needs to be sufficiently descriptive to enable the determination of the most relevant cognitive demands imposed by each part of the task and the impact of context, as reflected in CPCs (Section 3.2), on the prediction of performance.

*Step* 2. The next step involves assessing the CPCs. Instead of using an additive weighted sum of CPCs, which assumes independence among the CPCs, CREAM derives a combined CPC score that takes into account dependencies among the CPCs. For example, referring to the CPCs in Table 10, both the "number of simultaneous goals" (the number of simultaneous tasks the person has to attend to at the same time) and "available time" are assumed to depend on the "working conditions." Specifically, improvement in working conditions would result in a *reduction* in the number of simultaneous goals and an *increase* in the available time CPC. Suppose "working conditions" are assessed as "compatible," implying a "not significant" effect on performance reliability (refer to columns 3 and 4 in Table 10).

Then, depending on the effects of other CPCs on "working conditions" (e.g., if they reduce or improve "working conditions"), the assessment of "working conditions" may either remain as "compatible," or be changed to "improved" or "reduced." Using this dependency approach, the synergistic effects of CPCs are taken into account and ultimately (qualitatively) reassessed in terms of their expected effects on performance reliability.

*Step* 3. A combined CPC score is derived by counting the number of times a CPC is expected to (1) reduce performance reliability; (2) have no significant effect; or (3) improve performance reliability. The combined CPC score is expressed as the triplet

$$\left[ \sum\nolimits_{\text{reduced}} \quad \sum\nolimits_{\text{not significant}} \quad \sum\nolimits_{\text{improved}} \right]$$

Not all values are possible when deriving a combined CPC score from these counts. For example, as indicated in Table 10, neither the "number of simultaneous goals" nor the "time of day" can result in an improvement on performance reliability. In the end, there are a total of 52 different combined CPC scores (Figure 14). Among these 52 scores, the triplet [9, 0, 0] describes the least desirable situation (because all 9 CPCs have a "reduced" effect on performance reliability) and the triplet [0, 2, 7] describes the most desirable situation (because the best effect on performance reliability of two of the CPCs is "not significant").

*Step* 4. The final step in the basic method of CREAM is to map this combined CPC score to a general action failure probability. This is accomplished
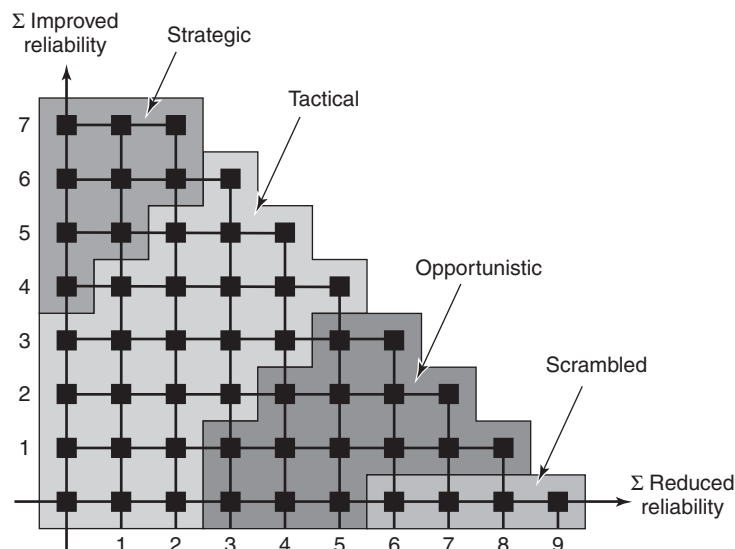


**Figure 14** Relations between CPC score and control modes. (From Hollnagel, 1998. Copyright 1998 with permission from Elsevier.)

by invoking the concept of "control mode" and creating a plot (Figure 14) that serves a function similar to that of the *risk assessment matrix*, a tool used to conduct subjective risk assessments in hazard analysis (U.S. Department of Defense, 1993). Depending on the region within the plot where 1 of the 52 values of the combined CPC score falls, the human is assumed to be performing in one of the four control modes. For example, the scrambled control mode is represented by the four cases where $\sum_{improved} = 0$ and $\sum_{reduced} = >5$.

While there may be a number of different ways to map the four control modes to corresponding human reliability intervals, Hollnagel (1998) offers one particular set of such intervals. For example, for the strategic control mode, the interval comprising the probability ($p$) of an action failure is [0.5E-5 < $p$ < 1.0E-2], whereas for the scrambled control mode this interval would be [1.0E-1 < $p$ < 1.0E-0], implying the possibility for a probability of failure as high as 1.0.

CREAM's *extended method*, like its basic method, is also centered on the principle that actions occur in a context. However, it offers the additional refinement of producing specific action failure probabilities. Thus, different actions or task segments that may fall into the same control mode region would, in principle, have different failure probabilities. The extended method shares a number of the same features as the basic method, in particular the initial emphasis on task analysis and the evaluation of CPCs. However, to generate more specific action failure probabilities, it incorporates the following layers of refinements:

*Step* 1. The first step is to characterize the task segments or steps of the overall task in terms of the cognitive activities they involve. The goal is to determine if the task depends on a specific set of cognitive activities, where the following list of cognitive activities is considered: coordinate, communicate, compare, diagnose, evaluate, execute, identify, maintain, monitor, observe, plan, record, regulate, scan, and verify. The method recognizes that this list of cognitive activities is not necessarily complete or correct. It also acknowledges the important role that judgment plays in selecting one or more of these cognitive activities to characterize a task step and recommends documenting the reasons for these assignments.

*Step* 2. A *cognitive demands profile* is then created by mapping each of the cognitive activities into four broad cognitive functions. These four functions are observation, interpretation, planning, and execution (which correspond to the three stages of information processing depicted in Figure 2). For example, the cognitive activity "evaluation" is described in terms of the cognitive functions "interpretation" and "planning"; the cognitive activity "coordinate" refers to the cognitive functions "planning" and "execution"; and

the cognitive activity "monitor" refers to the cognitive functions "observation" and "interpretation." Although some cognitive activities (e.g., "diagnose" and "evaluate") may both refer to the same cognitive functions ("interpretation" and "planning"), they are considered distinct because they refer to different task activities during performance. For example, during diagnosis, the emphasis may be on reasoning whereas during evaluation the emphasis may be on assessing a situation through an inspection operation.

Following the description of each cognitive activity in terms of its associated cognitive functions, a cognitive demands profile can be constructed by counting the number of times each of the four cognitive functions occurs. This can be done for each of the task segments or for the task as a whole. A cognitive demands profile plot can then be generated, for example, by listing each task segment on the $x$ axis, and the corresponding relative percentages to which each of the cognitive functions is demanded by that segment.

*Step* 3. Once a profile of cognitive functions associated with the task segments has been constructed, the next step is to identify the likely failures associated with each of these four cognitive functions. In principle, the basis for determining these failures should derive from the complete list of phenotypes (Section 3.1) and genotypes (e.g., Table 9), but for practical purposes, a subset of this list can be used. Thus, for each of the four cognitive functions, a number of potential failures are considered. For example, for the cognitive function "observation," three observation errors are taken into account: observation of wrong object, wrong identification made, and observation not made. Similarly, a subset of interpretation (3), planning (2), and execution (5) errors are considered corresponding to the other three cognitive functions, resulting in a total of 13 types of *cognitive function failures*.

Clearly, if a different set of cognitive functions is identified for use in this HRA model, then a set of cognitive function failures corresponding to those cognitive functions would need to be selected. In any case, given the knowledge of the task and of the CPCs under which the task is being performed, for each task segment the analyst must assess the likely failures that can occur. Note that the distribution of cognitive function failures for each task segment may look very different than the cognitive demands profile distribution, largely because of the impact that performance conditions (i.e., context) are believed to be having. Thus, a task segment may have a larger percentage of cognitive functions associated with observation than interpretation, but following the assessment of cognitive function failures may show a larger number of interpretation failures.

*Step* 4. Following the assignment of likely cognitive function failures to each task segment, the next step involves computing a cognitive failure

probability for each type of error that can occur. These cognitive failure probabilities (CFPs) are analogous to HEPs. Using a variety of different sources, including Swain and Guttmann (1983) and Gertman and Blackman (1994), nominal values as well as corresponding lower (0.05) and upper (0.95) bounds are assigned to these CFPs. For example, for the observation error "wrong identification made," the nominal CFP given is 7.0E-2, and the corresponding lower and upper bound estimates are [2.0E-2, 1.7E-2].

*Step* 5. Next, the effects of CPCs on the nominal CFPs are assessed. The computation of the CPC score that was part of the basic method of CREAM is used to determine which of the four control modes is governing performance of the task segment or task. The nominal CFP is then adjusted based on weighting factors associated with each of these control modes. For the scrambled, opportunistic, tactical, and strategic control modes, the four corresponding weighting factors that are specified are $[2.3E + 01, 7.5E + 00, 1.9E + 00, 9.4E-01]$. These adjustments imply, for example, multiplying the CFP value by 23 if the control mode is determined to be "scrambled" and multiplying the CFP by 0.94 if the control mode is determined to be "strategic."

*Step* 6. If the analyst wishes to reduce the uncertainty associated with adjusting nominal CFPs based on the control mode that is governing performance, a more complex approach can be used. This approach requires that couplings between the nine CPCs and the four cognitive functions (observation, interpretation, planning, and execution) be established by assigning, to each CPC, a "weak," medium, or "strong" influence on each cognitive function. These influences are inherent to the CPCs. For example, the CPC "availability of procedures" would be expected to have a strong influence on the cognitive function "planning," as planning what to do would depend on what alternatives are available, which are presumably described in the procedures. However, this CPC would be expected to have a weak influence on "interpretation" (presumably because procedures do not provide such elaboration). Using similar logic, the CPC "working conditions" would be expected to have a weak influence on "planning" but a medium influence on "observation."

The nominal CFPs and their corresponding lower and upper bounds are then adjusted by weighting factors that are derived as follows. First, the CPC table (Table 10) is consulted to determine whether each CPC is expected to have an effect on performance reliability (if the effect is assessed to be "not significant," then the weighting factor is 1, implying no modification of the nominal CFP). If the CPC is expected to have an effect on performance reliability, then the couplings that were established between the CPCs and the four cognitive functions are used to moderate those effects accordingly; in the

case where the coupling between a CPC and a cognitive function was deemed "weak," then a weight of 1 is assigned.

Ultimately, based on various sources of knowledge, weighting factors are assigned to each of the four cognitive functions for each CPC level, and these weights are used to adjust the original nominal CFPs of the 13 types of failures that were classified according to the type of cognitive function required. For example, for the error type "wrong identification," which is one of the three error types classified under the cognitive function "Observation," consider the CPC "working conditions." Further, consider the three levels of this CPC: advantageous, compatible, and incompatible. For the cognitive function "observation," the weighting factors that would be used to adjust the nominal CFP for the "wrong identification" error type, which is 7.0E-2, are 0.8, 1.0, and 2.0, respectively. Thus, if the CPC is indeed evaluated to be advantageous, the nominal CFP would be adjusted down from 0.07 to $0.8 \times 7.0E-2 = 0.056$, whereas if the CPC is evaluated to be incompatible, the CFP would be adjusted up to $2 \times 7.0E-2 = 0.14$. The lower and upper bounds would be adjusted accordingly. No adjustment would be made if the CPC is evaluated to be compatible.

*Step* 7. Continuing with the example above, in reality, the other eight CPCs could also have an effect on the cognitive function of "observation" and thus on the "wrong identification" observation error. Referring to Table 10, assume the evaluations of the nine CPCs, from top to bottom, were as follows: inefficient, compatible, tolerable, inappropriate, matching current capacity, adequate, daytime, inadequate, and efficient. The corresponding weighting factors would be [1.0, 1.0, 1.0, 2.0, 1.0, 0.5, 1.0, 2.0, 1.0]. The total effect of the influence from the CPCs for this error type is determined by multiplying all the weights, which results in a value of 2.0. The nominal CFP of 7.0E-2 would then be multiplied by 2, resulting in an overall adjusted CFP of $14.0E-2 = 0.14$. If a task is comprised of a number of task segments, the one or more errors that could occur in each segment would be determined in the same way.

*Step* 8. The final step in the extended method of CREAM involves incorporation of the adjusted CFPs into a PRA. This requires providing a single quantitative estimate of human error for the task. If the method was applied to an entire task, the resulting CFP would be used. However, if the method was applied to a number of task segments comprising a task, for example, a sequence of task steps that could be described through a HTA, then the task CFP required for input to the PRA would be based on the component CFPs.

In a fault tree representation of the HTA, if a task requires that a number of component substeps *all* be performed correctly, then any

substep performed incorrectly would lead to failure; under these disjunctive (i.e., logical OR) conditions, the error probability for the step can be taken as the maximum of the individual substep CFPs. If, however, a task step requires only one of a number of component substeps to be performed correctly for the task step to be successful, then only if all the substeps are performed incorrectly would the task step fail; under these conjunctive (i.e., logical AND) conditions, the error probability for the step can be taken as the product of the individual substep CFPs.

## 4.10 HRA Methods: Concluding Remarks

### 4.10.1 Benchmarking HRA Methods

There are a number of ways in which HRA methods can be evaluated and compared—that is, "benchmarked." In a recent article, lessons learned from benchmarking studies involving a variety of other types of methods, as well as issues associated with HRA benchmarking studies, were reviewed for the purpose of ensuring that important considerations were accounted for in planning future HRA benchmarking studies (Boring et al., 2010).

Validation in HRA benchmarking studies is often based on some objective performance measure, such as the probability of human error, against which the corresponding estimates of the HRA methods can be compared. However, even such comparisons can be problematic as different HRA methods may have different degrees of fit to the task or scenario chosen for analysis. Emphasis thus also needs to be given to the diversity of "product" areas—the different kinds of tasks, scenarios, or ways in which the HRA method can analyze a situation—for which these methods are best suited in order to more fully evaluate their capabilities.

In addition to the focus on end-state probabilities generated by the different methods, evaluations in benchmarking studies should also be directed at the qualitative processes that led to those probabilities, including assumptions underlying the method, how PSFs are used, how tasks or scenarios are decomposed, and how dependencies are considered. Comparisons of HRA methods based on other qualitative considerations (e.g., the degree of HRA expertise needed or resources required to use the method), while inherently subjective, can still reveal strengths and weaknesses that can greatly influence the appropriateness of an HRA method for a particular problem (Bell and Holyroyd, 2009).

The inconsistency among analysts in how scenarios or tasks are decomposed for analysis is a particular concern in HRA benchmarking studies and may be partially accountable for low interrater reliability in HEP calculations among analysts using the same HRA method (Boring et al., 2010). Benchmarking studies thus could benefit from frameworks for comparing qualitative aspects of the analysis as well as from uncertainty information (in the form of lower and upper uncertainty bounds on HEPs) to allow comparisons of the range of the HEPs computed.

In Kirwan's (1996) quantitative validation study of three HRA methods, 10 different analysts assessed each of the methods for 30 human error scenarios derived from the CORE database (Sections 4.4 and 4.10.3). Although a generally strong degree of consistency was found between methods and across analysts, no one method was sufficiently comprehensive or flexible to cover a wide range of human performance scenarios, despite the exclusion of scenarios requiring knowledge-based performance (Section 2.2.4) or diagnostic tasks performed by operating crews. Comparing HRA methods on such tasks and on scenarios in domains other than nuclear power remains a challenge for HRA benchmarking studies (Boring et al., 2010).

### 4.10.2 Issue of Dependencies

A challenging problem for all HRAs is the identification of dependencies in human–system interactions and computing their effects on performance failures. Spurgin (2010) discusses the use of the beta factor as a means for accounting for dependencies, where

$$P[\text{B}|\text{A}] = \beta P[\text{B}]$$

In this expression, $P[\text{B}|\text{A}]$ is the probability of B given that activity A has occurred, $P[\text{B}]$ is the probability of activity B independent of A, and $\beta$ is the dependency factor. One method for determining $\beta$ in HRA studies is by using an event tree (ET) to model the influence of dependencies in any particular sequence of human activities that might occur in an accident scenario. In such an ET, the columns would correspond to the various types of dependency variables that would be considered to impact activity B following activity A.

Examples of such dependency variables are cognitive connections between tasks, time available for actions to be taken, relationships among various crew members and support staff, and work stress due to workload. Each of these variables may have a number of discrete levels; for example, the time available variable may be classified as long, medium, and short, with the branch leading through a short amount of time resulting in a much larger beta factor. The various paths through *dependency ETs* would correspond to qualitatively different sets of dependency influences. Accordingly, the end branches of these paths would be designated by different beta values, with higher beta values resulting in increased HEPs associated with activity B.

The relationships between the levels of these input dependency variables and the designated end-branch dependence levels are, however, assumed to be based on expert judgment. To reduce the uncertainty associated with experts providing direct judgments on the dependence levels, a dependence assessment method based on fuzzy logic has been proposed (Podofillini et al., 2010). Using the same five levels of dependency as THERP, this approach assigns a number of different linguistic labels to dependency input variables (e.g., none, low, medium, high, very high) that can span ranges of values that can overlap with one another and, through an expert elicitation process, also provides anchor points to represent prototype conditions of the input variables for particular tasks. Judgments on these input variables can be given as point values on or between anchors or as

an interval range of values. These judgments are then assigned degrees of membership in fuzzy sets (based on trapezoidal membership functions), which represent the degrees to which the judgments match each of the linguistic labels. The expert's knowledge is represented as a set of rules by which the relationship between different values of the input variables and output (dependency level) variables is characterized. The fuzzy logic procedure used in this approach provides different degrees of activation of these rules and ultimately the degrees of belief in terms of the possibility for the different dependency levels (the output fuzzy set). For PRAs, a "defuzzification" procedure would be needed to convert this output set to probability values.

Generally, HRA analysts are free to select or modify whatever guidelines, such as those offered in Swain and Guttman (1983), and procedures to model dependencies. Handling dependencies remains, not unlike other aspects of HRA methods, as much art as science.

### 4.10.3  Deriving HEP Estimates

A fundamental issue that is troubling for many HRA methods, especially those that are based on assigning HEP values to tasks or elemental task activities, is the derivation of such HEP estimates. Ideally, HEP data should derive from the relevant operating experience or at least from similar industrial experiences. However, as Kirwan (1994) notes, there are a number of problems associated with collecting this type of quantitative HEP data. For example, many workers will be reluctant to report errors due to the threat of reprisals. Also, errors that do not lead to a violation of a company's technical specifications or that are recovered almost immediately will probably not be reported. In addition, data on errors associated with very low probability events, as in the execution of recovery procedures following an accident, may not be sufficiently available to produce reliable estimates and thus often require simulator studies for their generation.

Another problem is that error reports are usually confined to the observable manifestations of an error (the external error modes). Without knowledge of the underlying cognitive processes or psychological mechanisms, errors that are in fact dissimilar (Table 1) may be aggregated. This would not only corrupt the HEP data but could also compromise error reduction strategies.

Kirwan (1999) has reported on the construction of an HEP database in the United Kingdom referred to as CORE-DATA (computerized operator reliability and error database) for supporting HRA activities (in fact, as was noted in Section 4.4, the HRA method NARA relies on this database). While CORE-DATA currently contains a large number of HEPs, its long-term objective was to apply its data to new industrial contexts through the development of extrapolation rules. Other large-scale projects intended for obtaining HEP data in support of HRA are the human error repository and analysis project sponsored by the U.S. Nuclear Regulatory Commission (Halbert et al., 2006) for establishing empirical relationships between contextual factors and human performance failures and the International HRA Empirical Study that is being performed by a group of international organizations jointly with the Organisation for Economic Co-operation and Development Halden Reactor Project (Lois et al., 2008), in which simulator data are being used to validate the predictions of HRA methods (Boring et al., 2010).

In HRA, what seems undeniable is that much depends on the use of expert judgment, whether it is to identify relevant human interactions; provide a lower, nominal, or upper bound estimate of a human failure probability; identify contextual factors such as common performance conditions that could influence performance in a given scenario; generate importance weights and quality ratings for those factors; resolve the effects of dependencies among human activities and factors defining work contexts; or provide guidance on how to extrapolate human error data to new contexts. Ultimately, some form of expert judgment remains the underlying critical aspect governing all HRA methods.

## 5  MANAGING HUMAN ERROR

This Section is confined to a few select topics that have important implications for human error and its management. Specifically, this Section overviews some issues related to designer error, the role of automation in human error, human error in maintenance operations, and the use of incident-reporting systems.

### 5.1  Designer Error

Designer errors generally arise from two sources: inadequate or incorrect knowledge about the application area (i.e., a failure for designers to anticipate important scenarios) and the inability to anticipate how the product will influence user performance (i.e., insufficient understanding by designers). The vulnerability of designers to these sources of performance failure is not surprising when one considers that designers' conceptualizations typically are nothing more than initial hypotheses concerning the collaborative relationship between their technological products and human users. Accordingly, their beliefs regarding this relationship need to be gradually shaped by data that are based on actual human interaction with these technologies, including the transformations in work experiences that these interactions produce (Dekker, 2005).

Although designers have a reasonable number of choices available to them that can translate into different technical, social, and emotional experiences for users, like users they themselves are under the influence of sociocultural (Evan and Manion, 2002) and organizational factors. For example, the reward structure of the organization, an emphasis on rapid completion of projects, and the insulation of designers from the consequences of their design decisions can induce designers to give less consideration to factors related to ease of operation and even safety (Perrow, 1983).

According to Perrow (1999), a major deficiency in the design process is the inability of both designers and managers to appreciate human fallibility by failing to take into account relevant information that could be supplied by human factors and ergonomics specialists. While this concern is given serious consideration in

user-centered design practices (Nielsen, 1995), in some highly technical systems, where designers may still be viewing their products as closed systems governed by perfect logic, this issue may still exist.

### 5.1.1 User Adaptation to New Technologies

Much of our core human factors knowledge concerning human adaptation to new technology in complex systems has been derived from experiences in the nuclear power and aviation industries. These industries were forced to address the consequences of imposing on their workers major transformations in the way that system data were presented. In nuclear power control rooms, the banks of hardwired displays were replaced by one or a few computer-based display screens, and in cockpits the analog single-function single displays were replaced by sophisticated software-driven electronic integrated displays.

These changes drastically altered the human's visual–spatial landscape and offered a wide variety of schemes for representing, integrating, and customizing data. For those experienced operators who were used to having the entire "data world" available to them at a glance, the mental models and strategies that they had developed and relied on were not likely to be as successful when applied to these newly designed environments and perhaps even predisposed them to committing errors to a greater extent than their less experienced counterparts.

In complex work domains such as health care that require the human to cope with a potentially enormous number of different task contexts, anticipating the user's adaptation to new technology can become so difficult for designers that they themselves, like the practitioners who will use their products, can be expected to resort to the tendency to minimize cognitive effort (Section 2.2.5). Instead of designing systems with operational contexts in mind, one cognitively less taxing solution is to identify and make available all possible information that the user may require, but to place the burden on the user to search for, extract, or configure the information as the situation demands.

These designer strategies are often manifest as technological mediums that exhibit the *keyhole property*, whereby the size of the available "viewports" is very small relative to the number of data displays that potentially could be examined (Woods and Watts, 1997). Unfortunately, this approach to design makes it more likely that the user can "get lost in the large space of possibilities" and makes it difficult to find the right data at the right time as activities change and unfold.

An example of this problem was demonstrated in a study by Cook and Woods (1996) that examined adapting to new technology in the domain of cardiac anesthesia. In this study, physiological monitoring equipment dedicated to cardiothoracic surgery was upgraded to a computer system that integrated the functions of four devices onto a single display. By virtue of the keyhole property, the new technology created new interface management tasks to contend with that derived, in part, from the need to access highly interrelated data serially. New interface management tasks also included the need

to declutter displays periodically to avoid obscuring data channels that required monitoring. This requirement resulted from collapsing into a single device the data world that was previously made available through a multi-instrument configuration.

To cope with these potentially overloading situations, physicians were observed to tailor both the computer-based system (*system tailoring*) and their own cognitive strategies (*task tailoring*). For example, to tailor their tasks, they planned their interactions with the device to coincide with self-paced periods of low criticality and developed stereotypical routines to avoid getting lost in the complex menu structures rather than risk exploiting the system's flexibility. In the face of circumstances incompatible with task-tailoring strategies, the physicians had no choice but to confront the complexity of the device, thus diverting information-processing resources from the patient management function (Cook and Woods, 1996).

This *irony of automation*, whereby the burden of interacting with the technology tends to occur during those situations when the human can least afford to divert attentional resources, is also found in aviation. For example, automation in cockpits can potentially reduce workload by allowing complete flight paths to be programmed through keyboards. Changes in the flight path, however, require that pilots divert their attention to the numerous keystrokes that need to be input to the keyboard, and these changes tend to occur during takeoff or descent—the phases of flight containing the highest risk and that can least accommodate increases in pilot workload (Strauch, 2002).

Task tailoring reflects a fundamental human adaptive process. Thus, humans should be expected to shape new technology to bridge gaps in their knowledge of the technology and fulfill task demands. The concern with task tailoring is that it can create new cognitive burdens, especially when the human is most vulnerable to demands on attention, and mask the real effects of technology change in terms of its capability for providing new opportunities for human error (Dekker, 2005).

The provision of such new windows of opportunity for error was illustrated in a study by Cao and Taylor (2004) on the effects of introducing a remote robotic surgical system for laparoscopic surgery on communication among the operating room (OR) team members. In their study, communication was analyzed using a framework referred to as *common ground*, which represents a person's knowledge or assumptions about what other people in the communication setting know (Clark and Schaefer, 1989). The introduction of new technology into the OR provides numerous ways in which common ground, and thus patient safety, can become compromised. For example, roles may change, people become less familiar with their roles, the procedures for using the new technology are less familiar, and expectations for responses from communication partners become more uncertain. Misunderstandings can propagate through team members in unpredictable ways, ultimately leading to new forms of errors.

In this case, what these researchers found was that the physical barrier necessitated by the introduction of

the surgical robot had an unanticipated effect on the work context (Section 2.4). For example, the surgeon, now removed from the surgical site, had to rely almost exclusively on video images from this remote site. Consequently, instead of receiving a full range of sensory information from the visual, auditory, haptic, and olfactory senses, the surgeon had to contend with a "restricted field of view and limited depth information from a frequently poor vantage point" (Cao and Taylor, 2004, p. 310). These changes potentially overload the surgeon's visual system and also create more opportunities for decision-making errors due to gaps in the information that is being received (Section 2.2.3). Moreover, in addition to the need for obtaining information on patient status and the progress of the procedure, the surgeon has to cope with information-processing demands deriving from the need to access information about the status of the robotic manipulator. Ensuring effective coordination of the robotic surgical procedure actually entailed that the surgeon verbally distribute more information to the OR team members than with conventional laparoscopic surgery.

Overall, the communication patterns were found to be haphazard, which increased the team member's uncertainty concerning what information and when information should be distributed or requested. This has the potential for increasing human error resulting from miscommunication or lack of communication. Cao and Taylor suggested training to attain common ground, possibly through the use of rules or an information visualization system that could facilitate the development of a *shared mental model* among the team members (Stout et al., 1999).

## 5.2 Automation and Human Error

Innovations in technology will always occur and will bring with them new ways of performing tasks and doing work. Whether the technology completely eliminates the need for the human to perform a task or results in new ways of performing tasks through automation of selective task functions, the human's tasks will probably become reconfigured (Chapter 59). As demonstrated in the previous section, the human is especially vulnerable when adapting to new technology. During this period, knowledge concerning the technology and the impact it may have when integrated into task activities is relatively unsophisticated, and biases from previous work routines are still influential.

### 5.2.1 Levels of Automation

Automating tasks or system functions by replacing the human's sensing, planning, decision-making, or manual control activities with computer-based technology often requires making *allocation of function* decisions—that is, deciding which functions to assign to the human and which to delegate to automatic control (Sharit, 1997). Because these decisions can have an impact on the propensity for human error, the *level of automation* to be incorporated into the system needs to be carefully considered (Parasuraman et al., 2000; Kaber and Endsley,

2004). Higher levels of automation imply that automation will assume greater autonomy in decision making and control.

The primary concern with technology-centered systems is that they deprive themselves of the potential benefits that can be gained by virtue of the human being actively involved in system operations. These benefits can derive from the human's ability to anticipate, search for, and discern relevant data based on the current context; make generalizations and inferences based on past experience; and modify activities based on changing constraints. Determining the optimal level of automation, however, is a daunting task for the designer. While levels of automation somewhere between the lowest and highest levels may be the most effective way to exploit the combined capabilities of both the automation and the human, identifying an ideal level of automation is complicated by the need to also account for the consequences of human error and system failures (Moray et al., 2000).

In view of evidence that unreliable "decision automation" (e.g., automation that has provided imperfect advice) can more adversely impact human performance than unreliable "information automation" (e.g., automation that provides incorrect status information), it has been suggested, particularly in systems with high-risk potential, that the level of automation associated with decision automation be set to allow for human input into the decision-making process (Parasuraman and Wickens, 2008). This can be accomplished, for example, by allowing for the automation of information analysis (an activity that, like decision making, places demands on working memory) but allocating to the human the responsibility for the generation of the *values* associated with the different courses of action (Sections 2.2.1 and 2.2.3). The reduced vulnerability of human performance to unreliable information automation as compared to unreliable decision automation may lie in the fact that the "data world" (i.e., the raw input data) is still potentially available to the human under information automation.

### 5.2.2 Intent Errors in Use of Automation

In characterizing usage of automation, a distinction has been made between *appraisal errors* and *intent errors* (Beck et al, 2002) as a basis for *disuse* and *misuse* of automation (Parasuraman and Riley, 1997). Appraisal errors refer to errors that occur when the perceived utilities of the automated and nonautomated alternatives are inconsistent with the actual utilities of these options. In contrast, intent errors occur when the human intentionally chooses the option that lowers the likelihood of task success, despite knowledge of whether the automated or nonautomated alternative is more likely to produce the more favorable outcome. An intent error of particular interest is when humans refuse to use an automated device that they know would increase the likelihood of a successful outcome. For example, a human supervisory controller may choose to manually schedule a sequence of machining operations in place of using a scheduling aid that has proven utility for those decision-making scenarios.

One explanation for this type of intent error is the perception of the automation as a competitor or threat; this phenomenon is known as the *John Henry effect*. The hypothesis that personal investment in unaided (i.e., nonautomated) performance would increase the likelihood of the John Henry effect was tested by Beck et al. (2009) in an experimental study that manipulated both the participant's degree of personal investment and the reliability of the automated device in a target detection task. The findings supported the hypothesis that when the automation was more reliable than the human high personal investment would lead to its increased disuse, and when the automation was less reliable than the human it would lead to its lower misuse, relative to those participants with less personal involvement.

John Henry effects can be expressed in many ways. For example, an experienced worker who feels threatened by newly introduced automation may convince other co-workers not to use the device, in effect creating a recalcitrant work culture (Section 6). Some strategies for countering John Henry effects include demonstrating to workers the advantages of using the aid in particular scenarios and construing the automation as a partner or collaborator rather than as an adversary.

### 5.2.3 Automation and Loss of Skill

Well-designed automation can lead to a number of indirect benefits related to human performance. For example, automation in manufacturing operations that offloads the operator from many control tasks enables the human controller to focus on the generation of strategies for improving system performance. Reckless design strategies, however, that automate functions based solely on technical feasibility can often lead to a number of problems (Bainbridge, 1987). For instance, manual and cognitive skills that are no longer used due to the presence of automation will deteriorate, jeopardizing the system during times when human intervention is required. Situations requiring rapid diagnosis that rely on the human having available or being able to rapidly construct an appropriate mental model will thus impose higher working memory demands on humans who are no longer actively involved in system operations. The human may also need to allocate significant attention to monitoring the automation, which is a task humans do not perform well.

These problems are due largely to the capability for automation to insulate the human from the process and are best handled through training that emphasizes ample hands-on simulation exercises encompassing varied scenarios. The important lesson learned is that "disinvolvement can create more work rather than less, and produce a greater error potential" (Dekker, 2005, p. 165). This tenet was highlighted in a recent article concerning errors involving air traffic controllers and pilots that have led to a sudden increase in near collisions of airliners (*Wall Street Journal*, 2010). In some cases, pilots had to make last-second changes in direction following warnings by cockpit alarms of an impending crash. Although collision warning systems have, together with other advances in cockpit safety equipment, contributed to the decrease in major airline crashes over the last decade, as stated by U.S. Transportation Department Inspector General Mary Schiavo, one consequence of the availability of these systems, which essentially constitute a type of symbolic barrier system (Table 3), is that "it's easy for pilots to lose their edge." As was discussed in Section 2.4, the perception of these barrier systems by humans can alter the context in ways that can increase the human's predisposition for performance failures.

### 5.2.4 Mode Errors and Automation Surprises

Automation can be "clumsy" for the human to interact with, making it difficult to program, monitor, or verify, especially during periods of high workload. A possible consequence of clumsy automation is that it "tunes out small errors and creates opportunities for larger ones" (Weiner, 1985) by virtue of its complex connections to and control of important systems.

Automation has also been associated with *mode errors*, a type of mistake in which the human acts based on the assumption that the system is in a particular mode of operation (either because the available data support this premise or because the human instructed the system to adopt that mode), when in fact it is in a different mode. In these situations, unanticipated consequences may result if the system remains capable of accommodating the human's actions.

Generally, when the logic governing the automation is complex and not fully understood by the human, the actions taken by automatic systems may appear confusing. In these situations, the human's tendency for partial matching and biased assessments (Section 2.2) could lead to the use of an inappropriate rule for explaining the behavior of the system—a mistake that, in the face of properly functioning automation, could have adverse consequences. These forms of human–automation interaction have been examined in detail in flight deck operations in the cockpit and have been termed *automation surprises* (Woods et al., 1997).

Training that allows the human to explore the various functions of the automation under a wide range of system or device states can help reduce some of these problems. However, it is also essential that designers work with users of automation to ensure that the user is informed about what the automation is doing and the basis for why it is doing it. In the past, slips and mistakes by flight crews tended to be errors of commission. With automation, errors of omission have become more common, whereby problems are not perceived and corrective interventions are not made in a timely fashion.

### 5.2.5 Mistrust of and Overreliance on Automation

When the performance of automatic systems or subsystems is perceived to be unreliable or uncertain, *mistrust of automation* can develop (Lee and Moray, 1994; Rasmussen et al., 1994). As Lee and See (2004) have pointed out, many parallels exist between the trust that we gain in other people and the trust we acquire in complex technology, and as in our interactions with other people, we tend to rely on automation we trust and reject automation we do not trust.

Mistrust of automation can provide new opportunities for errors, as when the human decides to assume manual control of a system or decision-making responsibilities that may be ill-advised under the prevailing conditions. Mistrust of automation can also lead to its disuse, which impedes the development of knowledge concerning the system's capabilities and thus further increases the tendency for mistrust and human error. To help promote appropriate trust in automation, Lee and See suggest that the algorithms governing the automation be made more transparent to the user; that the interface provide information regarding the capabilities of the automation in a format that is easily understandable; and that training address the varieties of situations that can affect the capabilities of the automation.

Harboring high trust in imperfect automation could also lead to human performance failures as a result of the *complacency* that could arise from *overreliance on automation*. A particularly dangerous situation is when the automation encounters inputs or situations unanticipated in its design but which the human believes the automation was programmed to handle.

In situations involving monitoring information sources for critical state changes, overreliance on the automation to perform these functions could lead to the human diverting resources of attention to other concurrent tasks. One way to counter such overreliance on automation is through *adaptive automation* (Sharit, 1997; Parasuraman and Wickens, 2008), which returns the automated task to human control when the (adaptive) automated system detects phases when human workload is low. Such a reallocation strategy, when implemented sporadically, could also serve to refresh and thus reinforce the human's mental model of automated task behavior.

System-driven adaptation, however, whether it is initiated for the purpose of countering complacency during low-workload phases or for off-loading the human during high-workload phases, adds an element of unpredictability to the overall human–system interactive process. The alternative solution of shifting the control of the adaptive process to the human may, on the other hand, impose an excessive decision-making load. Not surprisingly, implementing effective adaptive automation designs in complex work domains remains a challenging area.

### 5.3   Human Error in Maintenance

To function effectively, almost all systems require maintenance. Frequent scheduled (i.e., preventive) maintenance, however, can be costly, and organizations often seek to balance these costs against the risks of equipment failures. Lost in this equation, however, is a possible "irony of maintenance"—an increased frequency in scheduled maintenance may actually increase system risk by providing more opportunities for human interaction with the system (Reason, 1997). This increase in risk is more likely if assembly rather than disassembly operations are called for, as the comparatively fewer constraints associated with assembly operations make these activities much more susceptible to various errors,

such as identifying the wrong component, applying inappropriate force, or omitting an assembly step (Lehto and Buck, 2008).

Maintenance environments are notorious for breakdowns in communication, often in the form of implicit assumptions or ambiguity in instructions that go unconfirmed (Reason and Hobbs, 2003). When operations extend over shifts and involve unfamiliar people, these breakdowns in communication can propagate into catastrophic accidents, as was the case in the explosion aboard the Piper Alpha oil and gas platform in the North Sea (Reason and Hobbs, 2003) and the crash of ValueJet flight 592 (Strauch, 2002).

Incoming shift workers are particularly vulnerable to errors following the commencement of their task activities, especially if maintenance personnel in the outgoing shift fail to brief incoming shift workers adequately concerning the operational context about to be confronted (Sharit, 1998). In these cases, incoming shift workers may be placed in the difficult position of needing to invest considerable attention almost immediately in order to avoid an incident or accident.

Many preventive maintenance activities initially involve searching for flaws prior to applying corrective procedures, and these search processes are often subject to various expectancies that could lead to errors. For example, if faults or flaws are seldom encountered, the likelihood of missing such targets will increase; if they are encountered frequently, properly functioning equipment may be disassembled. Maintenance workers are also often required to work in restricted spaces that are error inducing by virtue of the physical and cognitive constraints that these work conditions can impose (Reynolds-Mozrall et al., 2000).

Flawed partnerships between maintenance workers and troubleshooting equipment can also give rise to errors. As with other types of aiding devices, troubleshooting aids can compensate for human limitations and extend human capabilities when designed appropriately. However, these devices are often opaque and may be misused or disregarded (Parasuraman and Riley, 1997). For instance, if the logic underlying the software of an expert troubleshooting system is inaccessible, the user may not trust the recommendations or explanations given by the device (Section 5.2) and therefore choose not to replace a component that the device has identified as faulty.

Errors resulting from interruptions are particularly prevalent in maintenance environments. Interruptions due to the need to assist a co-worker or following the discovery that the work procedure called for the wrong tool or equipment generally require the worker to leave the scene of operations. In these kinds of situations, the most likely type of error is an omission. In fact, memory lapses probably constitute the most common errors in maintenance, suggesting the need for incorporating good reminders (Reason, 1997). Reason and Hobbs (2003) emphasize the need for mental readiness and mental rehearsal as ways that maintenance workers can inoculate themselves against errors that could arise from interruptions, time pressure, communication, and unfamiliar situations that may arise.

Written work procedures are pervasive in maintenance operations, and numerous problems with the design of these procedures may exist that can predispose their users to errors (Drury, 1998). Violations of these procedures are also relatively common, and management has been known to consider such violations as causes and contributors of adverse events. This belief, however, is both simplistic and unrealistic, and may be partly due to the fact that work procedures are generally based on normative models of work operations. The actual contexts under which real work takes place are often very different from those that the designers of the procedures have envisioned or were willing to acknowledge. To the followers of the procedures, who must negotiate their tasks while being subjected to limited resources, conflicting goals, and pressures from various sources, the cognitive process of transforming procedures into actions is likely to expose incomplete and ambiguous specifications that, at best, appear only loosely related to the actual circumstances (Dekker, 2005).

A worker's ability to adapt (and thereby violate) these procedures successfully may, in fact, be lauded by management and garner respect from fellow workers. However, if these violations happen to become linked to accidents, management would most likely refute their knowledge or tacit approval of these informal activities and retreat steadfastly to the official doctrine—that safety will be compromised if workers do not follow procedures. Dekker suggests that organizations monitor (Section 5.4) and understand the basis for the gaps between procedures and practice and develop ways of supporting the cognitive skill of applying procedures successfully across different situations by enhancing workers' judgments of when and how to adapt.

## 5.4 Incident-Reporting Systems

Information systems such as incident-reporting systems (IRSs) can allow extensive data to be collected on incidents, accidents, and human errors. *Incidents* comprise events that are not often easy to define. They may include actions, including human errors, responsible for the creation of hazardous conditions. They may also include *near misses*, which are sometimes referred to as close calls.

Capturing information on near misses is particularly advantageous as, depending on the work domain, near misses may occur hundreds of times more often than adverse events. The contexts surrounding near misses, however, should be similar to and thus highly predictive of accidents. The reporting of near misses, especially in the form of short event descriptions or detailed anecdotal reports, could then provide a potentially rich set of data that could be used as a basis for proactive interventions.

The role of management is critical to the successful development and implementation of an IRS (CCPS, 1994). Management not only allocates the resources for developing and maintaining the system but can also influence the formation of work cultures that may be resistive to the deployment of IRSs. In particular, organizations that have instituted "blame cultures" (Reason, 1997) are unlikely to advocate IRSs that emphasize underlying causes of errors, and workers in

these organizations are unlikely to volunteer information to these systems.

Often, the data that is collected or its interpretation will reflect management's attitudes concerning human error causation. The adoption of a system-based perspective on human error would imply the need for an information system that emphasizes the collection of data on possible causal factors, including organizational and management policies responsible for creating the latent conditions for errors. A system-based perspective on human error is also conducive to a dynamic approach to data collection: If the methodology is proving inadequate in accounting for or anticipating human error, it will probably be modified.

Worker acceptance of an IRS that relies on voluntary reporting entails that the organization meet three requirements: exact a minimal use of blame; ensure freedom from the threat of reprisals; and provide feedback indicating that the system is being used to affect positive changes that can benefit all stakeholders. Accordingly, workers would probably not report the occurrence of accidental damage to an unforgiving management and would discontinue voluntarily offering information on near misses if insights gained from intervention strategies are not shared (CCPS, 1994). It is therefore essential that reporters of information perceive IRSs as error management or learning tools and not as disciplinary instruments.

In addition to these fundamental requirements, two other issues need to be considered. First, consistent with user-centered design principles (Nielsen, 1995), potential users of the system should be involved in its design and implementation. Second, effective training is critical to the system's usefulness and usability. When human errors, near misses, or incidents occur, the people who are responsible for their reporting and investigation need to be capable of addressing in detail all considerations related to human fallibility, context, and barriers that affect the incident. Thus, training may be required for recognizing that an incident has in fact occurred and for providing full descriptions of the event.

Analysts also would need training, specifically, on applying the system's tools, including the use of any modeling frameworks for analyzing causality of human error, and on interpreting the results of these application tools. They would also need training on generating summary reports and recommendations and on making modifications to the system's database and inferential tools if the input data imply the need for such adjustments.

Data for input into IRSs can be of two types: quantitative data, which are more readily coded and classified, and qualitative data in the form of free-text descriptions. Kjellén (2000) has specified the basic requirements for a safety information system in terms of data collection, distribution and presentation of information, and overall information system attributes. To meet data collection requirements, the input data need to be reliable (i.e., if the analysis were to be repeated, it should produce similar results) and accurate and provide adequate coverage (e.g., on organizational and human factors issues) needed for exercising efficient control.

Foremost in the distribution and presentation of information is the need for *relevant* information. Relevance will depend on how the system will be used. For example, if the objective is to analyze statistics on accidents in order to assess trends, a limited set of data on each accident or near miss would be sufficient and the nature of these data can often be specified in advance. However, suppose that the user is interested in querying the system regarding the degree to which new technology and communication issues have been joint factors in incidents involving errors of omission. In this case, the relevance will be decided by the coverage. Generally, the inability to derive satisfactory answers to specific questions will signal the need for modifications of the system.

### 5.4.1 Aviation Safety Reporting System

The Aviation Safety Reporting System (ASRS) was developed in 1976 by the Federal Aviation Administration (FAA) in conjunction with NASA. Many significant improvements in aviation practices have since been attributed to the ASRS, and these improvements have largely accounted for the promotion and development of IRSs in other work domains, most notably, the health care industry, which has been struggling with what has been termed an epidemic of adverse events stemming from medical errors (Kohn et al., 1999).

The ASRS's mission is threefold: to identify deficiencies and discrepancies in the National Aviation System (NAS); to support policy formulation and planning for the NAS; and to collect human performance data and strengthen research in the aviation domain. All pilots, air traffic controllers, flight attendants, mechanics, ground personnel, and other personnel associated with aviation operations can submit confidential reports if they have been involved in or observed any incident or situation that could have a potential effect on aviation safety. The ASRS database can be queried by accessing its Internet site (http://asrs.arc.nasa.gov).

ASRS reports are processed in two stages by groups of analysts composed of experienced pilots and air traffic controllers. In the first stage, each report is read by at least two analysts who identify incidents and situations requiring immediate attention. Alerting messages are then drafted and sent to the appropriate group. In the second stage, analysts classify the reports and assess causes of the incident. Their analyses and the information contained in the reports are then incorporated into the ASRS database. The database consists of the narratives submitted by each reporter and coded information that is used for information retrieval and statistical analysis procedures.

Several provisions exist for disseminating ASRS outputs. These include alerting messages that are sent out in response to immediate and hazardous situations; the *CALLBACK* safety bulletin, which is a monthly publication containing excerpts of incident report narratives and added comments; and the ASRS *Directline*, which is published to meet the needs of airline operators and flight crews. In addition, in response to database search requests, ASRS staff communicates with the FAA and the National Transportation Safety Board (NTSB) on

an institutional level in support of various tasks, such as accident investigations, and conducts and publishes research related primarily to human performance issues.

### 5.4.2 Some Issues with IRSs

Some IRSs, by virtue of their inability to cope with the vast number of incidents in their databases, have apparently become "victims of their own success" (Johnson, 2002). The FAA's ASRS and the Food and Drug Administration's MedWatch Reporting System (designed to gather data on regulated, marketed medical products, including prescription drugs, specialized nutritional products, and medical devices) both contain enormous numbers of incidents. Because their database technologies were not designed to manage this magnitude of data, users who query these systems are having trouble extracting useful information and often fail to identify important cases.

This is particularly true of the many IRSs that rely on *relational database* technology. In these systems, each incident is stored as a record, and incident identifiers are used to link similar records in response to user queries. Relational database techniques, however, do not adapt well to changes in either the nature of incident reporting or the models of incident causation.

Another concern is that different organizations in the same industry tend to classify events differently, which reduces the benefits of drawing on the experiences of IRSs across different organizations. It can also be extremely difficult for people who were not involved in the coding and classification process to develop appropriate queries (Johnson, 2002).

Problems with IRSs can also arise when large numbers of reports on minor incidents are stored. These database systems may then begin to drift toward reporting information on quasi-incidents and precursors of quasi-incidents, which may not necessarily provide the IRS with increased predictive capability. As stated by Amalberti (2001), " The result is a bloated and costly reporting system with not necessarily better predictability, but where everything can be found; this system is chronically diverted from its true calling (safety) to serve literary or technical causes. When a specific point needs to be proved, it is (always) possible to find confirming elements in these extra-large databases" (p. 113).

A much more fundamental problem with IRSs is the difficulty in assuring anonymity to reporters of information, especially in smaller organizations. Although most IRSs are confidential, anonymity is more conducive to obtaining disclosures of incidents. Unfortunately, anonymity precludes the possibility for follow-up interviews, which are often necessary for clarifying reported information (Reason, 1997).

Being able to follow up interviews, however, does not always resolve problems contained in reports. Gaps in time between the submission of a report and the elicitation of additional contextual information can result in important details being forgotten or confused, especially if one considers the many forms of bias that can affect eyewitness testimony (Johnson, 2002). Biases that can affect reporters of incidents can also affect the teams of people (i.e., analysts) that large-scale IRSs often employ

to analyze and classify the reports. For example, there is evidence that persons who have received previous training in human factors are more likely to diagnose human factors issues in incident reports than persons who have not received this type of training (Lekberg, 1997).

IRSs that employ classification schemes for incidents that are based on detailed taxonomies can also generate confusion, and thus variability, among analysts. Difficulty in discriminating between the various terms in the taxonomy may result in low recall systems, whereby some analysts fail to identify potentially similar incidents. Generally, limitations in analysts' abilities to interpret causal events reduce the capability for organizations to draw important conclusions from incidents, whereas analyst bias can lead to organizations using IRSs for supporting existing preconceptions concerning human error and safety.

The FAA's Aviation Safety Action Program (ASAP), a voluntary carrier-specific safety program that grew out of the success of the FAA's ASRS (Section 5.4.1), exemplifies the challenges in developing a classification scheme capable of identifying underlying causes of errors. In this program, pilots can submit short text descriptions of incidents that occurred during line operations. Although extracting diagnostic information from ASAP's text narratives can be an arduous task, it could be greatly facilitated if pilots were able to classify causal contributors of incidents when filing these reports. Baker and Krokos (2007) detail the development of such a classification system, referred to as ACCERS (Aviation Causal Contributors for Event Reporting Systems), in a series of studies involving pilots who were used to both establish as well as validate this system's taxonomic structure. An initial set of about 300 causal contributors were ultimately transformed into a hierarchical taxonomy consisting of seven causal categories (e.g., policies or procedures, human error, human factors, and organizational factors) and 73 causal factors that were assigned to one of these seven categories (e.g., conflicting policies and procedures, misapplication of flight controls, proficiency/overreliance on automation, and airline's safety culture).

Despite results which suggested that ACCERS reasonably satisfied three important evaluation criteria in taxonomy development—internal validity, external validity, and perceived usefulness—a number of problems existed that highlight the confusion that taxonomies can bring about. For example, pilots had difficulty differentiating between the human error and human factors categories, possibly due to confounding the error "outcome" with the "performance itself." Also, interrater agreement was relatively low, especially at the factor level (i.e., selecting factor-level causal contributors to the incident in the ASAP report), suggesting the need for training to ensure greater consistency in appraising the meaning of the causal factors.

Issues associated with error or incident reporting can also be highly work-domain specific. For example, the presumed considerable underreporting of medical incidents and accidents in the health care industry is likely to be due to a number of relatively unique barriers to reporting that this industry faces (Holden and Karsh,

2007). One issue is that many medical providers, by virtue of the nature of their work, may not be willing to invest the effort in documenting incidents or filing reports. Even electronic IRSs that may make it seem relatively easy to document errors or incidents (e.g., through drop-down menus) may still demand that the reporter collect or otherwise track down supportive information, which may require leaving one's work area at the risk of a patient's safety.

Many medical providers may not even be aware of the existence of a medical IRS. For example, they may not have been present when these systems were introduced or when training on them was given or were somehow not informed of their existence. The existence or persistence of any of these kinds of situations is symptomatic of managerial failure to provide adequate commitment to the reporting system. Another consideration is the transient nature of many complex medical environments. For example, some medical residents or part-time nurses, for reasons related to fear or distrust of physicians in higher positions of authority or because they do not perceive themselves as stakeholders in the organization, may not feel as compelled to file incident reports. Many medical providers, including nurses and technicians, may not even have an understanding of what constitutes an "error" or "incident" and may require training to educate them on the wide range of situations that should be reported and, depending on the IRS, how these situations should be classified. More generally, blame cultures are likely to be more prevalent in medical environments, where a fear of reprimand, being held liable, or the stigma associated with admissions of negligence or fallibility (Holden and Karsh, 2007) is still well established in many workers. In fact, in some electronic IRSs the wording of the disclaimer regarding the nature of protection the reporting system provides the worker may be sufficient reason for some workers not to use the system.

Finally, a very different type of concern arises when IRSs are used as a basis for quantitative human error applications. In these situations, the voluntary nature of the reporting may invalidate the data that are used for deriving estimates of human error probabilities (Thomas and Helmreich, 2002). From a probabilistic risk assessment (Section 4) and risk management perspective, this issue can undermine decisions regarding allocating resources for resolving human errors: Which errors do you attempt to remediate if it is unclear how often the errors are occurring?

### 5.4.3 Establishing Resiliency through IRSs

A kind of information that would be advantageous to catalog but that is extremely challenging to capture by the current state-of-the-art in incident reporting concerns the various adaptations by an organization's constituents to the external pressures and conflicting goals to which they are continuously subjected (Dekker, 2005). Instead of the more salient events that signal reporting in conventional IRSs, these adaptations, as might occur when a worker confronts increasingly scarce resources while under pressure to meet higher production standards,

can give rise to potentially risky conditions—a process that can be characterized as *drifting into failure*.

If the adaptive responses by the worker to these demands gradually become absorbed into the organization's definition of normal work operations, work contexts that may be linked to system failures are unlikely to be reported and thus remain concealed. The intricate, incremental, and transparent nature of the adaptive processes underlying these drifts may be manifest at various levels of an organization. Left unchecked, the aggregation of these drifts seals an organization's fate by effectively excluding the possibility for proactive risk management solutions. In the case of the accident in Bhopal (Casey, 1993), these drifts were personified at all levels of the responsible organization.

Although reporting systems such as IRSs can, in theory, monitor and detect these types of drifts into failure, to do so these systems may need to be driven by new models of organizational dynamics and armed with new levels of intelligence. Overall, devising, managing, and effectively utilizing a reporting system capable of capturing an organization's adaptive capacity relative to the dynamic challenges to that capacity is consistent with the goal of creating a *resilient organization* (Dekker, 2005, 2006).

Presently, however, we have few models or frameworks to guide this process. To establish resiliency, this type of reporting enterprise would need to be capable of identifying the kinds of disruptions to its goals that can be absorbed without fundamental breakdowns in its performance or structure; when and how closely the system appears to be operating near its performance boundary; details related to the behavior of the system when it nears such a boundary; the types of organizational contexts, including management policies, that can resolve various challenges to system stability such as dealing with changing priorities, allocating responsibility to automation, or pressure to trade off production with safety concerns; and how adaptive responses by workers to these challenges, in turn, influence management policies and strategies (Woods, 2006). Getting the relevant data underlying these issues, let alone determining how this data should be exploited, remains a challenging problem.

Finally, while the focus in safety has largely been on models of failure, reflecting attempts to "confirm" our theories about how human error and failure events can result in accidents, in contrast we have little understanding of how normal work leads to stable system performance. This knowledge is prerequisite for determining how drifts become established and the kinds of system instability they can produce, especially when such drifts are built on a succession of incremental departures from previously established norms.

Identifying such drifts is further complicated by the reality that such incremental departures by one or more workers in response to system demands may produce simultaneous adaptive incremental responses by many other system constituents, including suppliers, managers, and even regulators, which can mask the initial behavioral departures. Collectively, these challenges are encapsulated by Dekker (2006) as follows: "a true model of drift may be out of reach altogether since it may be fundamentally immeasurable" (p. 85).

## 6  ORGANIZATIONAL CULTURE AND RESILIENCE

There are numerous factors with regard to the culture of an organization that are relevant to the topics of human error, risk, and safety. For example, Strauch (2002) identified two factors that he considered cultural antecedents to erroneous performance in organizations: identification with the group and acceptance of authority. In Hofstede's (1991) analysis of the influence of company cultures on behaviors among individuals, these factors were termed *individualism–collectivism* (the extent to which people identify with the group) and *power distance* (the extent to which people accept authority).

Whereas individually oriented people place personal goals ahead of organizational goals, collectivist-oriented persons tend to identify with the company (or work group), so that more of the responsibility for errors that they commit would be deflected onto the company. These distinctions thus may underlie attitudes that could affect the degree to which workers mentally prepare themselves for potential errors.

*Power distance* refers to the differences in power that employees perceive between themselves and subordinates and superiors. In cultures with high power distance, subordinates are less likely to point out or comment to others about errors committed by superiors as compared to workers in company cultures with low power distance. Cultures in which workers tend to defer to authority can also suppress the organization's capability for learning. For example, workers may be less willing to make suggestions that can improve training programs or operational procedures (Section 5.4).

Hofstede identified a third cultural factor, called *uncertainty avoidance*, which refers to the willingness or ability to deal with uncertainty; this factor also has implications for human error. For example, workers in cultures that are low in uncertainty avoidance are probably more likely to invoke performance at the knowledge-based level (Section 2.2.4) in response to novel or unanticipated situations for which rules are not available.

Another distinction related to organizational culture, especially in reference to industries engaged in high-risk operations, is whether an organization can be considered a *high-reliability organization* (HRO). Attributes generally associated with HROs include anticipating errors and encouraging safety at the expense of production; having effective error-reporting mechanisms without fear of reprisals; and maintaining channels of communication across all levels of the company's operations (Rochlin et al., 1987; Roberts, 1990; Bierly and Spender, 1995; Weick et al., 1999). In contrast, questionable hiring practices, poor economic incentives, inflexible and outmoded training programs, the absence of IRSs and meaningful accident investigation mechanisms, managerial instability, and the promotion of atmospheres that

discourage communication between superiors and subordinates represent attributes reflective of poor organizational cultures.

Through policies that prescribe a proactive safety culture, the mindset of HROs makes it possible to avert many basic human and system performance failures that plague numerous organizations. For example, HROs typically have policies in place that serve to ensure that various groups of workers interface with one another; relevant information, tools, and other specialized resources are available when needed; and problems do not arise due to inadequate staffing.

It can be argued that the attributes that often define an HRO also promote resiliency (Section 5.4.3). Organizations with "fortress mentalities" that lack a "culture of conscious inquiry" are antithetical to HROs; such organizations are more likely to miss potential risks that are unfolding and less likely to identify critical information needed to cope with the complexity that these situations carry (Westrum, 2006).

Building on work by Reason (1997) and Reason et al. (1998), Wreathall (2006) has identified the following seven organizational cultural themes which characterize the processes by which organizations become resilient in terms of both safety and production:

- *Top-Level Commitment*. Top-level management is attuned to human performance concerns and provides continuous and extensive follow-through to actions that address these concerns.
- *Just Culture*. As emphasized in Section 5.4, the perceived absence of a just culture will lessen the willingness of workers to report problems, ultimately diminishing the effectiveness of proactive risk management strategies.
- *Learning Culture*. Section 5.4 also alluded to the importance of well-designed and well-managed IRSs as a basis for enabling an organization to learn. However, this theme also encompasses the need to shed or otherwise avoid cultural attributes that can suppress organizational learning. An example of such an attribute is what Cook and Woods (2006) refer to as "distancing through differencing," whereby an organization may discount or distance itself from incidents or accidents that occur in other organizations with similar operations through various rationalizations that impede the possibility for learning.
- *Awareness*. This theme emphasizes the ongoing ability to extract insights from data gathered through reporting systems that can be used to gauge and rethink risk management models.
- *Preparedness*. This theme reflects a mindset of an organization that is continually anticipating mechanisms of failure (including human performance failures) and problems (including how improvements and other changes might induce new paths to failure), even when there has not been a recent history of accidents, and prepares for these potential problems (e.g., by ensuring the availability of needed resources for serious anomalous events).

- *Flexibility*. Organizations that embrace a learning culture are more likely to accord their supervisors with the flexibility to make adaptive responses in the face of routine and major crises that involve making difficult trade-off decisions.
- *Opacity*. The "open" culture that characterizes HROs, which allows interactions of individuals at all levels and encourages cross-monitoring and the open articulation of safety concerns without reprisals, provides such organizations with the buffering capacity to move toward safety boundaries without jeopardizing the safety or productivity of its operations.

To these themes one should add the willingness of management to temporarily relax the efficiency goal for the safety goal when circumstances dictate the need for doing so (Sheridan, 2008). Such circumstances appeared to be apparent in the case of the Deepwater Horizon accident (Section 6.2).

In Section 2.2.5, a number of common rules people apply were offered to exemplify the manifestation of the concept of the efficiency–thoroughness trade-off (ETTO) proposed by Hollnagel (2004). The manifestation of ETTO rules at the organizational level provides yet another basis upon which company cultures can be distinguished in terms of their propensity for inducing performance failures. Hollnagel (2004) offers the following examples of ETTO rules at the level of the organization:

- *Negative Reporting*. This rule drives organizations to report only deviations from normal states; the organization's "cognitive effort" is minimized by interpreting a lack of information as a confirmation that everything is safe.
- *Reduction of Uncertainty*. Overall physical and cognitive resources are saved through elimination of independent checks.
- *Management Double Standards*. This is personified in the classic situation whereby efficiency, in the form of meeting deadlines and productivity, is "pushed," often tacitly, on its workers, at the expense of the thoroughness that would be needed to ensure the safety standards that the organization purportedly, in its official doctrine, covets.

Another telltale sign that an organization's culture may be lacking in resilience, especially in its ability to balance the pressures of production with concerns for safety, resides in the nature of its maintenance operations. This was apparent in the crash of ValueJet flight 592 into the Florida Everglades in 1996 just minutes after takeoff. The crash occurred following an intense fire in the airplane's cargo compartment that made its way into the cabin and overcame the crew (Strauch, 2002). Unexpended and unprotected canisters of oxygen generators, which can inadvertently generate oxygen and heat and consequently ignite adjacent materials, had somehow managed to become placed onto the aircraft.

Although most of the errors that were uncovered by the investigation were associated with maintenance technicians at SabreTech—the maintenance facility contracted by ValueJet to overhaul several of its aircraft—these errors were attributed to practices at SabreTech that reflected organizational failures. For example, although the work cards (which specified the required steps for performing maintenance tasks) indicated either disabling the canisters with locking caps or expending them, these procedures were not carried out. Contributing to the failure to carry out these procedures was the unavailability of the locking caps needed to secure the unexpended oxygen generators. In addition, maintenance workers incorrectly tagged the canisters. Instead of applying red tags, which would have correctly identified the removed canisters as condemned or rejected components (the canisters were in fact expired), they applied green tags, which signified the need for further repairs or testing. Workers in shipping and receiving, who were ultimately responsible for placing the canisters on the airplane, thus assumed the canisters were to be retained. Had the correctly colored tags been attached to the components, these personnel would likely have realized that the canisters were of no value and thus were not to be returned to the airline.

There was also a lack of communication across shifts concerning the hazards associated with the oxygen generators, which was facilitated by the absence of procedures for briefing incoming and outgoing shift workers concerning hazardous materials and for tracking tasks performed during shifts. Deficiencies in training were also cited as a contributory cause of the accident. Although SabreTech provided instruction on various policies and procedures (e.g., involving inspection and hazardous material handling), contractor personnel, who comprised the majority of the company's technicians who worked on the canisters, received no training.

The finding that the majority of the technicians that removed oxygen canisters from ValueJet airplanes as part of the overhaul of these aircraft were not SabreTech personnel is particularly relevant to this discussion as this work arrangement can easily produce an inadequately informed organizational culture. It is also not surprising that management would be insensitive to the implications of outsourcing on worker communication and task performance, and focus instead on the cost reduction benefits. As Peters and Peters (2006) note: "Outsourcing can be a brain drain, a quality system nightmare, and an error producer unless rigorously and appropriately managed" (p. 152).

Finally, any discussion on organizational culture, especially within the context of risk management, would be remiss not to include the idea of a *safety culture* (Reason, 1997; Vicente, 2004; Glendon et al., 2006). A number of the elements required for the emergence of a safety culture within an organization have already been discussed with regard to IRSs (Section 5.4). Reason cautions, however, that having all the necessary ingredients of a safety culture does not necessarily establish a safety culture, and the perception by an organization that it has achieved a respectable or first-rate safety culture is almost a sure sign that they

are mistaken. This warning is consistent with one of the tenets of resiliency: as stated by Pariès (2006), "the core of a good safety culture is a self-defeating prophecy."

### 6.1 *Columbia* Accident

The physical cause of the *Columbia* space shuttle accident in 2003 was a breach in the thermal protection system on the leading edge of *Columbia*'s left wing about 82 s after the launch. This breach was caused by a piece of insulating foam that separated from the external tank in an area where the orbiter attaches to the external tank. However, the *Columbia* Accident Investigation Board's (2003) report stated that "NASA's organizational culture had as much to do with this accident as foam did," that "only significant structural changes to NASA's organizational culture will enable it to succeed," and that NASA's current organization "has not demonstrated the characteristics of a learning organization" (p. 12).

To some extent NASA's culture was shaped by compromises with political administrations that were required to gain approval for the space shuttle program. These compromises imposed competing budgetary and mission requirements that resulted in a "remarkably capable and resilient vehicle," but one that was "less than optimal for manned flights" and "that never met any of its original requirements for reliability, cost, ease of turnaround, maintainability, or, regrettably, safety" (p. 11).

The organizational failures are almost too numerous to document: unwillingness to trade off scheduling and production pressures for safety; shifting management systems and a lack of integrated management across program elements; reliance on past success as a basis for engineering practice rather than on dependable engineering data and rigorous testing; the existence of organizational barriers that compromised communication of critical safety information and discouraged differences of opinion; and the emergence of an informal command and decision-making apparatus that operated outside the organization's norms. According to the *Columbia* Accident Investigation Board, deficiencies in communication, both up and down the shuttle program's hierarchy, were a foundation for the *Columbia* accident.

These failures were largely responsible for missed opportunities, blocked or ineffective communication, and flawed analysis by management during *Columbia*'s final flight that hindered the possibility of a challenging but conceivable rescue of the crew by launching *Atlantis*, another space shuttle craft, to rendezvous with *Columbia*. The accident investigation board concluded: "Some Space Shuttle Program managers failed to fulfill the implicit contract to do whatever is possible to ensure the safety of the crew. In fact, their management techniques unknowingly imposed barriers that kept at bay both engineering concerns and dissenting views, and ultimately helped create 'blind spots' that prevented them from seeing the danger the foam strike posed" (p. 170). Essentially, the position adopted by managers concerning whether the debris strike created a safety-of-flight issue placed the burden on engineers to prove that the system was unsafe.

Numerous deficiencies were also found with the Problem Reporting and Corrective Action database, a critical information system that provided data on any nonconformances. In addition to being too time consuming and cumbersome, it was also incomplete. For example, only foam strikes that were considered in-flight anomalies were added to this database, which masked the extent of this problem.

What is particularly disturbing was the failure of the shuttle program to detect the foam trend and appreciate the danger that it presented. Shuttle managers discarded warning signs from previous foam strikes and normalized their occurrences. In so doing, they desensitized the program to the dangers of foam strikes and compromised the flight readiness process. Although many workers at NASA knew of the problem, in the absence of an effective mechanism for communicating these "incidents" (Section 5.4), proactive approaches for identifying and mitigating risks were unlikely to be in place. In particular, a proactive perspective to risk identification and management could have resulted in a better understanding of the risk of thermal protection damage from foam strikes; tests being performed on the resilience of the reinforced carbon–carbon panels; and either the elimination of external tank foam loss or its mitigation through the use of redundant layers of protection.

## 6.2 Deepwater Horizon Accident

On April 20, 2010, an explosion occurred on the Deepwater Horizon, a massive oil exploration rig located about 50 miles south of the Louisiana coast in the Gulf of Mexico. The rig was owned by the drilling company Transocean, the world's largest offshore drilling contractor, and leased to the energy company British Petroleum (BP). This accident resulted from a blowout—an uncontrolled or sudden release of oil or natural gases—of an oil well located about a mile below the surface of the sea. The explosion and ensuing inferno resulted in 11 deaths and at least 17 injuries.

The rig sank to the bottom of the sea and in the process ruptured the pipes that carried oil to the surface, ultimately leading to the worst oil spill in U.S. (and probably world) history and causing significant economic and environmental damage to the Gulf region. Because the pressure at the well site is more than a ton per square inch, recovery from the failure needed to be performed remotely. At the conclusion of the writing of this chapter, which occurred on the 80th day of the oil spill, BP had finally appeared, following a series of highly publicized failed attempts, to successfully cap the leak streaming from the blown well. The evidence accumulated by this time seemed to indicate that it was a complex combination of factors that led to this accident.

In these rigs the first line of defense is the blowout preventer (BOP), a stack of equipment about 40 ft high that contains hydraulic valves designed to automatically seal the wellhead during an emergency. However, workers on the Deepwater Horizon were not able to activate this equipment. The failure of the fail-safe BOP has garnered a tremendous amount of attention as it represents the ultimate (and single) defense against onrushing oil and gas when an oil rig loses control of a well.

The key device in the BOP is the blind shear ram. In the event of a blowout, the blind shear ram utilizes two blades to slice through the drill pipe and seal the well. However, if one of the small shuttle valves leading to the blind shear ram becomes jammed or leaks, the ram's blades may not budge, and there is evidence that there was leakage of hydraulic fluid in one or more of the shuttle valves when the crew on the rig activated the blind shear ram (*New York Times*, 2010a).

This vulnerability to the fail-safe system was known within the oil industry and prompted offshore drillers, including Transocean, to add a layer of redundancy by equipping their BOPs with two blind shear rams. In fact, at the time of the Deepwater Horizon accident 11 of Transocean's 14 rigs in the Gulf had two blind shear rams, as did every (other) oil rig under contract with BP (*New York Times*, 2010a). However, neither Transocean nor BP appeared to take the necessary steps to outfit Deepwater Horizon's BOP with two blind shear rams. Transocean stated that it was BP's responsibility, based on various factors such as water depth and seismic data, for deciding on the BOP. BP's position was that both companies needed to be involved in making such a determination, as the decision entailed consideration of contractor preferences and operator requirements.

The problem with assuring the reliability of these devices appears to extend across the entire oil industry and includes the whole process by which federally mandated tests on BOPs are run and evaluated. One study that examined the performance of blind shear rams in BOPs on 14 new rigs found that 7 had not even been checked to determine if their shear rams would function in deep water, and of the remaining 7 only 3 were found to be capable of shearing pipe at their maximum rated water depths. Yet, despite this lack of preparedness in the last line of defense against a blowout, and even as the oil industry moves into deeper water, BP and other oil companies financed a study in early 2010 aimed at arguing against conducting BOP pressure tests every 14 days in favor of having these tests performed every 35 days, which would result in an estimated annual savings of $193 million in lost productivity (*New York Times*, 2010a).

Irrespective of whether these required government tests indeed provide reasonable guarantees of safety, the federal Minerals Management Services (MMS), which at the time served under the U.S. Department of the Interior, issued permits to drill in deepwater without assurances that these companies' BOPs could shear pipe and seal a well at depths of 5000 ft. These regulatory shortcomings, which came to light following the Deepwater Horizon accident, led Ken Salazar, Secretary of the Interior, to announce plans for the reorganization of the MMS "by separating safety oversight from the division that collects royalties from oil and gas companies" (*New York Times*, 2010b). (The MMS has since been renamed Bureau of Ocean Energy Management, Regulation and Enforcement.)

There were also a number of indicators in the months and weeks prior to the Deepwater Horizon

accident that the risks of drilling might exceed acceptable risk boundaries. The crew of the Deepwater Horizon encountered difficulty maintaining control of the well against "kicks" (sudden releases of surging gas) and had problems with stuck drilling pipes and broken tools, costing BP millions of dollars in rig rental fees as they fell behind schedule. Immediately before the explosion there were warning signs that a blowout was impending, based on preliminary evidence of equipment readings suggesting that gas was bubbling into the well (*New York Times*, 2010c). In fact, in the month before the explosion BP officials conceded to federal regulators that there were issues controlling the well, and on at least three occasions BP records indicated that the BOP was leaking fluid, which limits its ability to operate effectively. Although regulators (the MMS) were informed by BP officials of these struggles with well control, they ultimately conceded to a request to delay their federally mandated BOP test, which is supposed to occur every two weeks, until problems were resolved. When the BOP was tested again, it was tested at a pressure level 35% below the levels used on the device before the delay and continued to be tested at this lower pressure level until the explosion.

In April, prior to the accident, according to testimony at hearings concerning the accident and documents made available to investigators (*New York Times* 2010a, 2010b), BP took what many industry experts felt were highly questionable shortcuts in preparing to seal the oil well, including using a type of casing that was the riskier (but more cost-effective in the long term) of two options. With this option, if the cement around the casing pipe does not seal properly, high-pressure gases could leak all the way to the wellhead, where only a single seal would serve as a barrier. In fact, hours before the explosion, gases were found to be leaking through the cement that had been set by an oil services contractor (*New York Times*, 2010d).

BP has blamed the various companies involved in the sealing operation, including Transocean's oil rig workers, who BP claimed did not pump sufficient water to fully replace the thick buffer liquid between the water and the mud. This buffer liquid may have clogged the pipe that was used for the critical negative pressure tests needed to determine if the well was properly sealed. The resulting (and satisfying) pressure test reading of zero may have reflected a *false assumption error* arising from the failure to consider a side effect—in this case, that the reading was due to the pipe being plugged and not due to the absence of pressure in the well. Following the misinterpretation of these pressure tests, the rig workers began replacing drilling mud in the pipe to the seabed with water. The blowout and ensuing explosion occurred about 2 h later (*New York Times*, 2010a).

Because BP had hoped to use the Deepwater Horizon to drill in another field by March 8, there may have been incentives for them to proceed quickly, trading off thoroughness for efficiency (Section 6). By the day of the accident, BP was 43 days behind schedule, and based on the cost of $533,000 per day that BP was paying to lease the rig, this delay had incurred a substantial financial cost. However, accusations during hearings that many of the decisions by BP officials were intended to save BP money and time at the risk of catastrophe were denied by BP's chief executive officer (CEO), Tony Hayward, who repeatedly defended many of these decisions in testimony to the U.S. House of Representatives Energy and Commerce committee by indicating that they were approved by the MMS.

The changes in safety culture at BP that presumably came about under the leadership of Tony Hayward (who was appointed CEO in 2007), though laudatory, appeared to address mostly lower level system issues, such as exhorting workers to grasp banisters (*New York Times*, 2010e). BP's safety culture at the larger system level was apparently already set in place under Hayward's predecessor, John Browne, who had a reputation for pursuing potentially lucrative and technologically riskier ventures. Along the way there was BP's oil refinery explosion in Texas City, Texas, in 2005 in which 15 people died and 170 were injured. Organizational and safety deficiencies at all levels of BP were deemed the cause of the accident; subsequently, OSHA found more than 300 safety violations, resulting in a then record of $21 million in fines. OSHA inspectors revisited the plant in 2009 and discovered more than 700 safety violations and proposed an $87.4 million fine, mostly because of failures to correct past failures. A year after the Texas City explosion, BP was responsible for the worst spill on Alaska's North Slope, where oil leaked from a network of pipelines.

BP's near sinking of its offshore Thunder Horse platform was caused by a check valve that had been installed backward. Following costly repairs to fix the damage to that rig, more significant welding problems were discovered in the form of cracks and breaks in the pipes comprising the underwater manifold that connects numerous wells and helps carry oil back to the platform. It turns out that the construction of this production platform was severely rushed and, once at sea, hundreds of employees worked to complete it under severe time constraints while living in cramped chaotic conditions in temporary encampments aboard ships. Overall, this history of near misses, accidents, and problems did not appear to translate into lessons learned in the case of the Deepwater Horizon.

The label "organizational error" has sometimes been applied to companies that have experienced highly adverse or catastrophic outcomes that are linked to risky decisions influenced by financial incentives, scheduling setbacks, or other pressures. While some may object to this label, in reality this type of "error" is similar to that which might be committed by, for example, a physician who chooses to process more patients at the risk of increased carelessness in identifying and assessing critical patient information. With organizational error, however, it is group dynamics that play an important role, which can lead to flawed assessments of system vulnerabilities as these assessments can be easily biased by higher order goals such as the need for meeting deadlines (Haines, 2008). These assessments are also susceptible to behaviors such as coercion and intimidation that can prevail in group decision-making scenarios (Lehto and Buck, 2008).

As the many factors potentially related to the Deepwater Horizon accident are further examined for their authenticity and more details come to light, which decisions and incidents, or combinations thereof, that may have led to the accident will likely continue to be scrutinized. Lax federal regulation, pressure from shareholders, and the technological challenges of deepwater drilling will surely form the core of factors that played a role, but so will the role of the safety culture. The federal government may also need to rethink its strategies. The recent decision by the administration to open up more challenging offshore areas to drilling in the interest of increasing domestic oil production provides the incentive for aggressive oil companies to pursue riskier operations using "ultra-deep" platforms far more sophisticated than the Deepwater Horizon. Such government policies, however, put everyone at risk if there is no simultaneous effort to ensure appropriate regulatory oversight.

## 7 FINAL REMARKS

Human error remains a vast and intriguing topic. Some of the relatively recent interest in understanding and even predicting human error has been motivated by the possibility of finding its markings in the brain. For example, evidence from neuroimaging studies has linked an *error negativity*, an event-related brain potential, to the detection by individuals of action slips, errors of choice, and other errors (Nieuwenhuis et al., 2001; Holroyd and Coles, 2002), possibly signifying the existence of a neurophysiological basis for a preconscious action-monitoring system.

However, suggestions that these kinds of findings may offer possibilities for predicting human errors in real-time operations (Parasuraman, 2003) are probably overstated. Event-related brain potentials may provide insight into attentional preparedness and awareness of response conflicts, but the complex interplay of factors responsible for human error (Section 2.1) takes these discoveries out of contention as explanatory devices for most meaningful types of errors. Moreover, the practical utility of such findings is highly questionable given the complexity, and thus uncertainty associated with the actual environmental conditions in which humans operate as well as the uncertainty inherent in psychophysiological measures and their subsequent analyses (Cummings, 2010).

Often, one hears of the need for eliminating human error. This goal, however, is not always desirable. The realization that errors have been committed can play a critical role in human adaptability, creativity, and the manifestation of expertise. The elimination of human error is also inconceivable if only because human fallibility will always exist. Even if our attention and memory capabilities could be vastly extended, either through normal evolutionary processes or technological tampering, the probable effect would be the design and production of new and more complex systems that, in turn, would lead to more complex human activities with new and unanticipated opportunities for human error.

In no way, however, should such suppositions deter the goals of human error prediction, assessment, and reduction, especially in complex high-risk systems. As a start, system hardware and software need to be made more reliable; better partnerships between humans and automation need to be established; barriers that are effective in providing detection and absorption of errors without adversely affecting contextual and cognitive constraints need to be put in place; and IRSs that enable organizations to learn and anticipate, especially when errors become less frequent and thus deprive analysts with the opportunity for preparing and coping with their effects, need to become more ubiquitous.

Organizations also need to consider the adoption of strategies and processes for implementing features that have come to be associated with *high-reliability organizations* (Section 6). In particular, emphasis needs to be given to the development of cultures of reliability that anticipate and plan for unexpected events, try to monitor and understand the gap between work procedures and practice (Dekker, 2005), and place value in organizational learning.

The qualitative role of HRA in PRAs (Section 4) also needs to be strengthened. It is not hard to imagine a third generation of approaches to HRA that focuses more on ways of analyzing human performance in varying contexts and can more effectively assess the contribution of a wide variety of human–system interactive behaviors to the creation of hazardous conditions and system risks. These advances in HRA would depend on continued developments in methods for describing work contexts and determining the perceptions and assessments that workers might make in response to these contexts.

Where relevant, these methods also need to be integrated into the conceptual, development, and testing stages of the product and system design process. This would enable designers to become better informed about the potential effects of design decisions, thus bridging the gap between the knowledge and intentions of the designer and the needs and goals of the user.

Problems associated with performance failures in work operations have traditionally been "dumped" on training departments. Instead of using training to compensate for these problems, it should be given a proactive role through the use of methods that emphasize management of task activities under uncertainty and time constraints and the development of cognitive strategies for error detection (Kontogiannis and Malakis, 2009); give consideration to the kinds of cues that are necessary for developing situation awareness (Endsley et al., 2003) and for interpreting common-cause and common-mode system failures; and utilize simulation to provide workers with extensive exposure to a wide variety of contexts. By including provisions in training for imparting mental preparedness, people will be better able to anticipate the anomalies they might encounter and thus the errors they might make (Reason and Hobbs, 2003).

However, perhaps the greatest challenge in reducing human error is managing these error management processes (Reason and Hobbs, 2003)—defense strategies need to be aggregated coherently (Amalberti, 2001). Too

often these types of error reduction enterprises, innovative as they may be, remain isolated or hidden from each other.

## REFERENCES

Alloy, L. B., and Tabachnik, N. (1984), "Assessment of Covariation by Humans and by Animals: The Joint Influence of Prior Expectations and Current Situational Information," *Psychological Review*, Vol. 91, pp. 112–149.

Amalberti, R. (2001), "The Paradoxes of Almost Totally Safe Transportation System," *Safety Science*, Vol. 3, pp. 109–126.

American Society for Mechanical Engineers (ASME) (2008), "Standard for Level 1/Large Early Release Frequency Probabilistic Risk Assessment for Nuclear Power Plant Applications," RA-S-2008, ASME, New York.

Apostolakis, G. E. (2004), "How Useful Is Quantitative Risk Assessment?" *Risk Analysis*, Vol. 24, pp. 515–520.

Apostolakis, G. E., Bier, V. M., and Mosleh, A. (1988), "A Critique of Recent Models for Human Error Rate Assessment," *Reliability Engineering and System Safety*, Vol. 22, pp. 201–217.

Bainbridge, L. (1987), "Ironies of Automation," in *New Technology and Human Error*, J. Rasmussen, K. Duncan, and J. Leplat, Eds., Wiley, New York, pp. 273–276.

Baker, D. P., and Krokos, K. J. (2007), "Development and Validation of Aviation Causal Contributors for Error Reporting Systems (ACCERS)", *Human Factors*, Vol. 49, pp. 185–199.

Beck, H. P., Dzindolet, M. T., and Pierce, L. G. (2002), "Applying a Decision-Making Model to Understand Misuse, Disuse, and Appropriate Automation Use," in *Advances in Human Factors and Cognitive Engineering: Automation*, Vol. 2, E. Salas, Ed., JAI, Amsterdam, pp. 37–78.

Beck, H. P., Bates McKinney, J., Dzindolet, M. T., and Pierce, L. G. (2009), "Effects of Human-Machine Competition on Intent Errors in a Target Detection Task," *Human Factors*, Vol. 51, pp. 477–486.

Bell, J., and Holyroyd, J. (2009), "Review of Human Reliability Assessment Methods," RR679, Health and Safety Executive, Bootle, UK., available on http://www.hse.gov.uk/research/rrhtm/rr679.htm

Bierly, P. E., and Spender, J. C. (1995), "Culture and High Reliability Organizations: The Case of the Nuclear Submarine," *Journal of Management*, Vol. 21, pp. 639–656.

Boring, R. L., Hendrickson, S. M. L., Forester, J. A., Tran, T. Q., and Lois, E. (2010), "Issues in Benchmarking Human Reliability Analysis Methods: A Literature Review," *Reliability Engineering and System Safety*, Vol. 95, pp. 591–605.

Cao, C. G. L., and Milgram, P. (2000), "Disorientation in Minimal Access Surgery: A Case Study," in *Proceedings of the IEA 2000/HFES 2000 Congress*, Vol. 4, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 169–172.

Cao, C. G. L., and Taylor, H. (2004), "Effects of New Technology on the Operating Room Team," in *Work with Computing Systems, 2004*, H. M. Khalid, M. G. Helander, and A. W. Yeo, Eds., Damai Sciences, Kuala Lumpur, Malaysia, pp. 309–312.

Casey, S. (1993), *Set Phasers on Stun and Other True Tales of Design, Technology, and Human Error*, Aegean Park Press, Santa Barbara, CA.

Casey, S. (2006), *The Atomic Chef and Other True Tales of Design, Technology, and Human Error*, Aegean Park Press, Santa Barbara, CA.

Center for Chemical Process Safety (CCPS) (1992), *Guidelines for Hazard Evaluation Procedures, with Worked Examples*, 2nd ed., CCPS, American Institute of Chemical Engineers, New York.

Center for Chemical Process Safety (CCPS) (1994), *Guidelines for Preventing Human Error in Process Safety*, CCPS, American Institute of Chemical Engineers, New York.

Christoffersen, K., and Woods, D. D. (1999), "How Complex Human–Machine Systems Fail: Putting 'Human Error' in Context," in *The Occupational Ergonomics Handbook*, W. Karwowski and W. S. Marras, Eds., CRC Press, Boca Raton, FL, pp. 585–600.

Clark, H. H., and Schaefer, E. F. (1989), "Contributing to Discourse," *Cognitive Science*, Vol. 13, pp. 259–294.

Clemens, R. (1996), *Making Hard Decisions*, Duxbury, Pacific Grove, CA.

Columbia Accident Investigation Board (2003), *Report Volume 1*, U.S. Government Printing Office, Washington, DC.

Cook, R. I., and Woods, D. D. (1996), "Adapting to New Technology in the Operating Room," *Human Factors*, Vol. 38, pp. 593–611.

Cook, R. I., and Woods, D. D. (2006), "Distancing through Differencing: An Obstacle to Organizational Learning Following Accidents," in *Resilience Engineering: Concepts and Precepts*, E. Hollnagel, D. D. Woods, and N. Leveson, Eds., Ashgate, Aldershot, England, pp. 329–338.

Cook, R. I., Render, M., and Woods, D. D. (2000), "Gaps in the Continuity of Care and Progress on Patient Safety," *British Medical Journal*, Vol. 320, pp. 791–794.

Cullen, D. J., Bates, D. W., Small, S. D., Cooper, J. B., Nemeskal, A. R., and Leape, L. L. (1995), "The Incident Reporting System Does Not Detect Adverse Drug Events: A Problem for Quality Improvement," *Joint Commission Journal on Quality Improvement*, Vol. 21, pp. 541–548.

Cummings, M. L. (2010), "Technological Impedances to Augmented Cognition," *Ergonomics in Design*, Vol. 18, pp. 25–27.

Czaja, S. J., and Sharit, J., Eds. (2009), *Aging and Work: Issues and Implications in a Changing Landscape*, Johns Hopkins University Press, Baltimore, MD.

Dekker, S. W. A. (2001), "The Disembodiment of Data in the Analysis of Human Factors Accidents," *Human Factors and Aerospace Safety*, Vol. 1, pp. 39–58.

Dekker, S. W. A. (2005), *Ten Questions about Human Error: A New View of Human Factors and System Safety*, Lawrence Erlbaum Associates, Mahwah, NJ.

Dekker, S. (2006), "Resilience Engineering: Chronicling the Emergence of Confused Consensus," in *Resilience Engineering: Concepts and Precepts*, E. Hollnagel, D. D. Woods, and N. Leveson, Eds., Ashgate, Aldershot, England, pp. 77–92.

Dervin, B. (1998), "Sense-Making Theory and Practice: An Overview of User Interests in Knowledge Seeking and Use," *Journal of Knowledge Management*, Vol. 2, pp. 36–46.

Dey, A. K. (2001), "Understanding and Using Context," *Personal and Ubiquitous Computing*, Vol. 5, pp. 4–7.

Dhillon, B. S., and Singh, C. (1981), *Engineering Reliability: New Technologies and Applications*, Wiley, New York.

Drury, C. G. (1998), "Human Factors in Aviation Maintenance," in *Handbook of Aviation Human Factors*, D. J. Garland, J. A. Wise, and V. D. Hopkin, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 591–606.

Embrey, D. E., Humphreys, P., Rosa, E. A., Kirwan, B., and Rea, K. (1984), *SLIM–MAUD: An Approach to Assessing Human Error Probabilities Using Structured Expert Judgment*, NUREG/CR-3518, U.S. Nuclear Regulatory Commission, Washington, DC.

Endsley, M. R. (1995), "Toward a Theory of Situation Awareness," *Human Factors*, Vol. 37, pp. 32–64.

Endsley, M. R., Bolté, B., and Jones, D. G. (2003), *Designing for Situation Awareness: An Approach to User-Centred Design*, CRC Press, Boca Raton, FL.

Electric Power Research Institute (EPRI) (1992), "Approach to the Analysis of Operator Actions in Probabilistic Risk Assessment," TR-100259, EPRI, Palo Alto, CA.

Evan, W. M., and Manion, M. (2002), *Minding the Machines: Preventing Technological Disasters*, Prentice-Hall, Upper Saddle River, NJ.

Fischhoff, B. (1975), "Hindsight–Foresight: The Effect of Outcome Knowledge on Judgment under Uncertainty," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 1, pp. 278–299.

Forester, T., Kolaczkowski, A., Cooper, S., Bley, D., and Lois, E. (2007), *ATHEANA User's Guide*, NUREG-1880, U.S. Nuclear Regulatory Commission, Washington, DC.

Fraser, J. M., Smith, P. J., and Smith, J. W. (1992), "A Catalog of Errors," *International Journal of Man–Machine Systems*, Vol. 37, pp. 265–307.

Gertman, D. I., and Blackman, H. S. (1994), *Human Reliability and Safety Analysis Data Handbook*, Wiley, New York.

Gertman, D. I., Blackman, H. S., Marble, J. L., Byers, J. C., and Smith, C. L. (2005), *The SPAR-H Human Reliability Analysis Method*, NUREG/CR-68:INL/EXT-05-00509, U.S. Nuclear Regulatory Commission, Washington, DC.

Glendon, A. I., Clarke, S. G., and McKenna, E. F. (2006), *Human Safety and Risk Management*, 2nd ed., CRC Press, Boca Raton, FL.

Grosjean, V., and Terrier, P. (1999), "Temporal Awareness: Pivotal in Performance?" *Ergonomics*, Vol. 42, pp. 443–456.

Haines, Y. Y. (2008), "Systems-Based Risk Analysis," in *Global Catastrophic Risks*, N. Bostrom and M. M. Ćirković, Eds., Oxford University Press, New York.

Halbert, B., Boring, R., Gertman, D., Dudenboeffer, D., Whaley, A., Marble, J., Joe, J., and Lois, E. (2006), "Human Event Repository and Analysis (HERA) System, Overview," NUREG/CR-6903, Vol. I, U.S. Nuclear Regulatory Commission, Washington, DC.

Hannaman, G. W., and Spurgin, A. J. (1984), "Systematic Human Action Reliability Procedure," EPRI NP-3583, Electric Power Research Institute, Palo Alto, CA.

Hannaman, G. W., Spurgin, A. J., and Lukic, Y. D. (1984), "Human Cognitive Reliability Model for PRA Analysis," Draft Report NUS-4531, EPRI Project RP2170-3, Electric Power Research Institute, Palo Alto, CA.

Helander, M. G. (1997), "The Human Factors Profession," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York, pp. 3–16.

Hofstede, G. (1991), *Cultures and Organizations: Software of the Mind*, McGraw-Hill, New York.

Holden, R. J., and Karsh, B. (2007), "A Review of Medical Error Reporting System Design Considerations and a Proposed Cross-Level Systems Research Framework," *Human Factors*, Vol. 49, pp. 257–276.

Hollnagel, E. (1993), *Human Reliability Analysis: Context and Control*, Academic, London.

Hollnagel, E. (1998), *Cognitive Reliability and Error Analysis Method*, Elsevier Science, New York.

Hollnagel, E., Ed. (2003), *Handbook of Cognitive Task Design*, Lawrence Erlbaum Associates, Mahwah, NJ.

Hollnagel, E. (2004), *Barriers and Accident Prevention*, Ashgate, Aldershot, England.

Holroyd, C. B., and Coles, M. G. H. (2002), "The Neural Basis of Human Error Processing: Reinforcement Learning, Dopamine, and the Error-Related Negativity," *Psychological Review*, Vol. 109, pp. 679–709.

Institute of Electrical and Electronics Engineers (IEEE) (1997), "Guide for Incorporating Human Action Reliability Analysis for Nuclear Power Generating Stations," IEEE Standard 1082, IEEE, Piscataway, NJ.

Johnson, C. (2002), "Software Tools to Support Incident Reporting in Safety-Critical Systems," *Safety Science*, Vol. 40, pp. 765–780.

Johnson, W. G. (1980), *MORT Safety Assurance Systems*, Marcel Dekker, New York.

Kaber, D. B., and Endsley, M. R. (2004), "The Effects of Level of Automation and Adaptive Automation on Human Performance, Situation Awareness and Workload in a Dynamic Control Task," *Theoretical Issues in Ergonomics Science*, Vol. 4, pp. 113–153.

Kaye, K. (1999), "Automated Flying Harbors Hidden Perils," *South Florida Sun-Sentinel*, September 27.

Kirwan, B. (1994), *A Guide to Practical Human Reliability Assessment*, Taylor & Francis, London.

Kirwan, B. (1997), "The Validation of Three Human Reliability Quantification Techniques—THERP, HEART, and JHEDI: Part II—Results of Validation Exercise," *Applied Ergonomics*, Vol. 28, pp. 27–39.

Kirwan, B. (1999), "Some Developments in Human Reliability Assessment," in *The Occupational Ergonomics Handbook*, W. Karwowski and W. S. Marras, Eds., CRC Press, Boca Raton, FL, pp. 643–666.

Kirwan, B., and Ainsworth, L. K. (1992), *Guide to Task Analysis*, Taylor & Francis, London.

Kjellén, U. (2000), *Prevention of Accidents through Experience Feedback*, Taylor & Francis, London.

Knox, N. W., and Eicher, R. W. (1983), *MORT User's Manual*, Department of Energy, 76/45-4, EG&G Idaho, Idaho Falls, ID.

Kohn, L. T., Corrigan, J. M., and Donaldson, M. S., Eds. (1999), *To Err Is Human: Building a Safer Health System*, National Academy Press, Washington, DC.

Kontogiannis, T., and Lucas, D. (1990), "Operator Performance under High Stress: An Evaluation of Modes, Case Studies and Countermeasures," Report No. R90/03, prepared for the Nuclear Power Engineering Test Center, Tokyo, Human Reliability Associates, Dalton, Wigan, Lancashire, UK.

Kontogiannis, T., and Malakis, S. (2009), "A Proactive Approach to Human Error Detection and Identification in Aviation and Air Traffic Control," *Safety Science*, Vol. 47, pp. 693–706.

Koppel, R., Metlay, J. P., Cohen, A., Abaluck, B., Localio, A. R., Kimmel, S., and Strom, B. L. (2005), "Role of Computerized Physician Order Entry Systems in Facilitating Medication Errors," *Journal of the American Medical Association*, Vol. 293, pp. 1197–1203.

Kumamoto, H., and Henley, E. J. (1996), *Probabilistic Risk Assessment and Management for Engineers and Scientists*, 2nd ed., IEEE Press, Piscataway, NJ.

Leape, L. L., Brennan, T. A., Laird, N. M., Lawthers, A. G., Localio, A. R., Barnes, B. A., Hebert, L., Newhouse, J. P., Weiler, P. C., and Hiatt, H. H. (1991), "The Nature of Adverse Events in Hospitalized Patients: Results from the Harvard Medical Practice Study II," *New England Journal of Medicine*, Vol. 324, pp. 377–384.

Lee, J. D., and Moray, N. (1994), "Trust, Self-Confidence, and Operators' Adaptation to Automation," *International Journal of Human–Computer Studies*, Vol. 40, pp. 153–184.

Lee, J. D., and See, K. A. (2004), "Trust in Automation: Designing for Appropriate Reliance," *Human Factors*, Vol. 46, pp. 50–80.

Lehto, M. R., and Buck, J. R. (2008), *Introduction to Human Factors and Ergonomics for Engineers*, Taylor & Francis, New York.

Lehto, M. R., and Nah, F. (1997), "Decision-Making Models and Decision Support," in *Handbook of Human Factors and Ergonomics*, 3rd ed., G. Salvendy, Ed., Wiley, New York, pp. 191–242.

Lekberg, A. (1997), "Different Approaches to Accident Investigation: How the Analyst Makes the Difference," in *Proceedings of the 15th International Systems Safety Conference*, B. Moriarty, Ed., International Systems Safety Society, Sterling, VA, pp. 178–193.

Leveson, N. G. (2002), *A New Approach to System Safety Engineering*, Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA.

Levy, J., Gopher, D., and Donchin, Y. (2002), "An Analysis of Work Activity in the Operating Room: Applying Psychological Theory to Lower the Likelihood of Human Error," in *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1457–1461.

Lois, E., Dang, V. N., Forester, J., Broberg, H., Massaiu, S., Hildebrandt, M., Braarud, P. O., Parry, G., Julius, J., Boring, R.L., Mänistö, I., and Bye, A. (2009), "International HRA Empirical Study—Phase I Report, NUREG/IA-0216", Vol 1, U.S. Nuclear Regulatory Commission, Washington, DC.

Luczak, H. (1997), "Task Analysis," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York, pp. 340–416.

Moray, N., Inagaki, T., and Itoh, M. (2000), "Adaptive Automation, Trust, and Self-Confidence in Fault Management of Time-Critical Tasks," *Journal of Experimental Psychology: Applied*, Vol. 6, pp. 44–58.

*New York Times* (2010a), "Regulators Failed to Address Risks in Oil Rig Fail-Safe Device," available: www.nytimes.com, accessed July 21, 2010.

*New York Times* (2010b), "U.S. Said to Allow Drilling without Needed Permits," available: www.nytimes.com, accessed May 14, 2010.

*New York Times* (2010c), "Documents Show Early Worries about Safety of Rig," available: www.nytimes.com, accessed May 30, 2010.

*New York Times* (2010d), "BP Used Riskier Method to Seal Oil Well before Blast," available: www.nytimes.com, accessed May 27, 2010.

*New York Times* (2010e), "In BP's Record, a History of Boldness and Costly Blunders," available: www.nytimes.com, accessed July 13, 2010.

Nickerson, R. S. (1998), "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology*, Vol. 2, pp. 175–220.

Nickerson, R. S. (2004), *Cognition and Chance: The Psychology of Probabilistic Reasoning*, Lawrence Erlbaum Associates, Mahwah, NJ.

Nielsen, J. (1995), *Usability Engineering*, Academic, San Diego, CA.

Nieuwenhuis, S. N., Ridderinkhof, K. R., Blom, J., Band, G. P. H., and Kok, A. (2001), "Error Related Brain Potentials Are Differentially Related to Awareness of Response Errors: Evidence from an Antisaccade Task," *Psychophysiology*, Vol. 38, pp. 752–760.

Norman, D. A. (1981), "Categorization of Action Slips," *Psychological Review*, Vol. 88, pp. 1–15.

Parasuraman, R. (2003), "Neuroergonomics: Research and Practice," *Theoretical Issues in Ergonomics Science*, Vol. 4, pp. 5–20.

Parasuraman, R., and Riley, V. (1997), "Humans and Automation: Use, Misuse, Disuse, and Abuse," *Human Factors*, Vol. 39, pp. 230–253.

Parasuraman, R., and Wickens, C. D. (2008), "Humans: Still Vital After All These Years," *Human Factors*, Vol. 50, 511–520.

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000), "A Model for Types and Levels of Human Interaction with Automation," *IEEE Transactions on Systems Man, and Cybernetics, Part A: Systems and Humans*, Vol. 30, pp. 276–297.

Pariès, J. (2006), "Complexity, Emergence, Resilience...," in *Resilience Engineering: Concepts and Precepts*, E. Hollnagel, D. D. Woods, and N. Leveson, Eds., Ashgate, Aldershot, England, pp. 43–53.

Pedersen, O. M. (1985), *Human Risk Contributions in Process Industry: Guides for Their Pre-Identification in Well-Structured Activities and for Post Incident Analysis*, Risø National Laboratory, Roskilde, Denmark.

Perrow, C. (1983), "The Organizational Context of Human Factors Engineering," *Administrative Science Quarterly*, Vol. 27, pp. 521–541.

Perrow, C. (1999), *Normal Accidents: Living with High-Risk Technologies*, Princeton University Press, Princeton, NJ.

Pesme, H., LeBot, P., and Meyer, P. (2007), "A Practical Approach of the MERMOS Method. Little Stories to Explore Human Reliability Assessment," IEEE/HPRCT Conference, Monterey, CA.

Peters, G. A., and Peters, B. J. (2006), *Human Error: Causes and Control*, Taylor & Francis, Boca Raton, FL.

Peterson, C. R., and Beach, L. R. (1967), "Man as an Intuitive Statistician," *Psychological Bulletin*, Vol. 68, pp. 29–46.

Phillips, L. D., Embrey, D. E., Humphreys, P., and Selby, D. L. (1990), "A Sociotechnical Approach to Assessing Human Reliability," in *Influence Diagrams, Belief Nets and Decision Making: Their Influence on Safety and Reliability*, R. M. Oliver and J. A. Smith, Eds., Wiley, New York.

Podofillini, L., Dang, V., Zio, E., Baraldi, P., and Librizzi, M. (2010), "Using Expert Models in Human Reliability

Analysis—A Dependence Assessment Method Based on Fuzzy Logic", *Risk Analysis*, Vol. 30, pp. 1277–1297.

Potter, S. S., Roth, E. M., Woods, D. D., and Elm, W. C. (1998), "A Framework for Integrating Cognitive Task Analysis into the System Development Process," in *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 395–399.

Rasmussen, J. (1982), "Human Errors: A Taxonomy for Describing Human Malfunction in Industrial Installations," *Journal of Occupational Accidents*, Vol. 4, pp. 311–333.

Rasmussen, J. (1986), *Information Processing and Human–Machine Interaction: An Approach to Cognitive Engineering*, Elsevier, New York.

Rasmussen, J., Pejterson, A. M., and Goodstein, L. P. (1994), *Cognitive Systems Engineering*, Wiley, New York.

Reason, J. (1990), *Human Error*, Cambridge University Press, New York.

Reason, J. (1997), *Managing the Risks of Organizational Accidents*, Ashgate, Aldershot, Hampshire, England.

Reason, J., and Hobbs, A. (2003), *Managing Maintenance Error: A Practical Guide*, Ashgate, Aldershot, Hampshire, England.

Reason, J. T., Parker, D., and Lawton, R. (1998), "Organizational Controls and Safety: The Varieties of Rule-Related Behaviour," *Journal of Occupational and Organizational Psychology*, Vol. 71, pp. 289–304.

Reynolds-Mozrall, J., Drury, C. G., Sharit, J., and Cerny, F. (2000), "The Effects of Whole-Body Restriction on Task Performance," *Ergonomics*, Vol. 43, pp. 1805–1823.

Roberts, K. H. (1990), "Some Characteristics of One Type of High Reliability Organization," *Organization Science*, Vol. 1, pp. 160–176.

Rochlin, G., La Porte, T. D., and Roberts, K. H. (1987), "The Self-Designing High Reliability Organization: Aircraft Carrier Flight Operations at Sea," *Naval War College Review*, Vol. 40, pp. 76–90.

Rouse, W. B., and Rouse, S. (1983), "Analysis and Classification of Human Error," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-13, pp. 539–549.

Rumelhart, D. E., and McClelland, J. L., Eds. (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, *Foundations*, MIT Press, Cambridge, MA.

Saaty, T. L. (1980), *The Analytic Hierarchy Process*, McGraw-Hill, New York.

Senders, J. W., and Moray, N. P. (1991), *Human Error: Cause, Prediction, and Reduction*, Lawrence Erlbaum Associates, Mahwah, NJ.

Sharit, J. (1997), "Allocation of Functions," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York, pp. 301–339.

Sharit, J. (1998), "Applying Human and System Reliability Analysis to the Design and Analysis of Written Procedures in High-Risk Industries," *Human Factors and Ergonomics in Manufacturing*, Vol. 8, pp. 265–281.

Sharit, J. (2003), "Perspectives on Computer Aiding in Cognitive Work Domains: Toward Predictions of Effectiveness and Use," *Ergonomics*, Vol. 46, pp. 126–140.

Shepherd, A. (2001), *Hierarchical Task Analysis*, Taylor & Francis, London.

Sheridan, T. B. (2008), "Risk, Human Error, and System Resilience: Fundamental Ideas," *Human Factors*, Vol. 50, pp. 418–426.

Simon, H. A. (1966), *Models of Man: Social and Rational*, Wiley, New York.

Spurgin, A. J. (2010), *Human Reliability Assessment: Theory and Practice*, CRC Press, Boca Raton, FL.

Spurgin, A. J., Moieni, P., Gaddy, C. D., Parry, G., Oris, D. D., Spurgin, J. P., Joksimovich, V., Gaver, D. P., and Hannaman, G. W. (1990), "Operator Reliability Experiments Using Power Plant Simulators (NP-6937)," Electric Power Research Institute, Palo Alto, CA.

Stout, R. M., Cannon-Bowers, J. A., Salas, E., and Milanovich, D. M. (1999), "Planning, Shared Mental Models, and Coordinated Performance: An Empirical Link Is Established," *Human Factors*, Vol. 41, pp. 61–71.

Strauch, B. (2002), *Investigating Human Error: Incidents, Accidents, and Complex Systems*, Ashgate, Aldershot, Hampshire, England.

Swain, A. D., and Guttmann, H. E. (1983), *Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications*, NUREG/CR-1278, U.S. Nuclear Regulatory Commission, Washington, DC.

Thomas, E. J., and Helmreich, R. L. (2002), "Will Airline Safety Models Work in Medicine?" in *Medical Error: What Do We Know? What Do We Do?* M. M. Rosenthal and K. M. Sutcliffe, Eds., Jossey-Bass, San Francisco, pp. 217–234.

Trost, W. A., and Nertney, R. J. (1985), *MORT User's Manual*, Department of Energy, 76/45-29, EG&G Idaho, Idaho Falls, ID.

Tversky, A., and Kahneman, D. (1974), "Judgment under Uncertainty: Heuristics and Biases," *Science*, Vol. 185, pp. 1124–1131.

U.S. Department of Defense (DoD) (1993), "System Safety Program Requirements," MIL-STD-882C, DoD, Washington, D.C.

Vicente, K. J. (1999), *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*, Lawrence Erlbaum Associates, Mahwah, NJ.

Vicente, K. J. (2004), *The Human Factor: Revolutionizing the Way People Live with Technology*, Routledge, New York.

*Wall Street Journal* (2010), "Near Collisions Raise Alarms on Air Safety," July 8, Associated Press.

Walsh, B. (2010), "The Meaning of the Mess," *TIME Magazine*, May 17, pp. 29–35.

Weick, K. E., Sutcliffe, K. M., and Obstfeld, D. (1999), "Organizing for High Reliability: Processes of Collective Mindfulness," *Research in Organizational Behavior*, Vol. 21, pp. 13–81.

Weiner, E. L. (1985), "Beyond the Sterile Cockpit," *Human Factors*, Vol. 27, pp. 75–90.

Westrum, R. (2006), "A Typology of Resilience Situations," in *Resilience Engineering: Concepts and Precepts*, E. Hollnagel, D. D. Woods, and N. Leveson, Eds., Ashgate, Aldershot, England, pp. 55–65.

Wickens, C. D. (1984), "Processing Resources in Attention," in *Varieties of Attention*, R. Parasuraman and R. Davies, Eds., Academic, New York, pp. 63–101.

Wickens, C. D., Liu, Y., Becker, S. E. G., and Lee, J. D. (2004), *An Introduction to Human Factors Engineering*, 2nd ed., Prentice-Hall, Upper Saddle River, NJ.

Wilde, G. J. S. (1982), "The Theory of Risk Homeostasis: Implications for Safety and Health," *Risk Analysis*, Vol. 2, pp. 209–225.

Williams, J. C. (1988), "A Data-Based Method for Assessing and Reducing Human Error to Improve Operational

Performance," in *Proceedings of IEEE Fourth Conference on Human Factors in Power Plants*, E. W. Hagen, Ed., IEEE, Piscataway, NJ, pp. 436–450.

Woods, D. D. (1984), "Some Results on Operator Performance in Emergency Events," *Institute of Chemical Engineers Symposium Series*, Vol. 90, pp. 21–31.

Woods, D. D. (2006), "Essential Characteristics of Resilience," in *Resilience Engineering: Concepts and Precepts*, E. Hollnagel, D. D. Woods, and N. Leveson, Eds., Ashgate, Aldershot, England, pp. 21–34.

Woods, D. D., and Hollnagel, E. (2005), *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering*, Taylor & Francis, Boca Raton, FL.

Woods, D. D., and Watts, J. C. (1997), "How Not to Navigate through Too Many Displays," in *Handbook of Human–Computer Interaction*, 2nd ed., M. Helander, T. K. Landauer, and P. Prabhu, Eds., Elsevier Science, New York, pp. 617–650.

Woods, D. D., Sarter, N. B., and Billings, C. E. (1997), "Automation Surprises," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York, pp. 1926–1943.

Wreathall, J. (2006), "Properties of Resilient Organizations: An Initial View," in *Resilience Engineering: Concepts and Precepts*, E. Hollnagel, D. D. Woods, and N. Leveson, Eds., Ashgate, Aldershot, England, pp. 275–285.

# CHAPTER 27

# MANAGING LOW-BACK DISORDER RISK IN THE WORKPLACE

**William S. Marras**
**Biodynamics Laboratory**
**The Ohio State University**
**Columbus, Ohio**

## 1 INTRODUCTION

Low-back disorders (LBDs) resulting in low-back pain (LBP) are common experiences in life. Although LBDs appear to occur more frequently as one ages, it does not need to be an inevitable result of aging. There is also abundant information about the work related-ness of LBD. Both the physical as well as organizational/psychosocial aspects of work have independently been associated with higher rates of LBDs. At a superficial level these findings may appear to represent a paradox relative to LBD causality, and there has been significant debate about the contribution of work factors compared to individual factors in defining risk. However, for most of us these factors coexist and are for the most part inexplicably linked.

When one steps away from the opinionated motivations behind many of the causal claims, it is clear that there is both a natural degenerative impact of aging upon the spine that is capable of leading to pain for some people. However, this degenerative process can be greatly accelerated through work exposure, thus leading to greater incidences of LBDs at the workplace.

This line of thinking suggests that one can never totally eliminate the risk of LBD in the workplace since a natural or base rate of LBDs would be expected to occur due to individual factors such as heredity and aging. However, through the proper design of work it is possible to minimize the additional (and often substantial) risk that could be offered through workplace risk factors. Therefore, this chapter will focus primarily upon what we now know about the causal factors leading to LBDs and LBP as well as how the workplace can be assessed and designed to minimize its impact on contributing to this additional workplace risk. Hence, this chapter will concentrate primarily upon the preventive aspects of workplace design from an ergonomics standpoint.

The science of ergonomics is concerned primarily with *prevention*. Many large and small companies have permanent ergonomic programs (processes) in place and have successfully controlled the risk as well as the costs associated with musculoskeletal disorders [Government Accountability Office (GAO), 1997]. Ergonomic approaches attempt to alter the work environment with the objective of controlling risk exposure and optimizing efficiency and productivity. Two types of risk control (interventions) categories are used in the workplace. The first control category involves engineering controls that physically change the orientation of the

work environment relative to the worker. Engineering controls alter the workplace and create a "smart" work environment where the risk has been minimized so that the work–person interface is optimal for productivity and minimal for risk. The second category of control involves administrative controls that are employed when it is not possible to provide engineering controls. It should be understood that administrative controls do not eliminate the risk. They attempt to control risk by managing the time of exposure to the risk in the workplace and, thus, require active management. Administrative controls often consist of rotation of workers to ensure that workers have adequate time to recover from exposure to risks through appropriate scheduling of non–risk exposure tasks.

While ergonomics typically addresses all aspects of musculoskeletal disorders as well as performance issues, this chapter will be limited to issues and principles associated with the prevention of LBDs due to repetitive physical work (not including vibration).

## 2   MAGNITUDE OF LOW-BACK PAIN PROBLEM AT WORK

Since most people work, workplace risk factors and individual risk factors are difficult to separate [National Research Council (NRC), 2001]. Nonetheless, the magnitude of LBDs in the workplace can be appreciated via surveys of working populations. Within the United States back disorders are associated with more days away from work than any other part of the body [National Institute for Occupational Safety and Health (NIOSH), 2000; Jacobs, 2008]. A study of 17,000 working-age men and women in Sweden (Vingard et al., 2002) indicated that 5% of workers sought care for a new LBP episode over a three-year period. In addition, they reported that many of these LBP cases became chronic. Assessment of information gathered in the National Health Injury Survey (NHIS) found that back

pain accounts for about one-quarter of the workers' compensation claims in the United States (Guo et al., 1995). About two-thirds of these LBP cases were related to occupational activities. Prevalence of lost-work days due to back pain was found to be 4.6% (Guo et al., 1999).

Recent efforts through the Bone and Joint Decade effort (Jacobs, 2008) have evaluated the burden of low-back problems on U.S. workers. This assessment reports that about 32% of the population reports pain that limits their ability to do work and 11% of workers report pain that limits their ability to do any work. Within these categories, 62 and 63% of workers, respectively, report low-back dysfunction as the limiting factor responsible for their work limitations. When work limitation due to back pain was considered as a function of gender, we see that more females report slightly more back pain than males. As shown in Figure 1, back pain that limits work or prevents one from working occurs more frequently as a function of age up until 65–74 years of age and then deceases slightly over the age of 75.

Certain types of occupations have also reported significantly greater rates of LBP. Reported risk was greatest for construction laborers (prevalence 22.6%) followed by nursing aides (19.8%) (Guo et al., 1995). However, a recent literature review (Hignett, 2008) has concluded that the annual LBP prevalence in nurses is as high as 40–50%. Figure 2 shows a summary of the distribution of lost-time back cases in private industry as a function of the type of work and the source of the injury based upon a NIOSH analysis of work-related LBDs (NIOSH, 2000). This figure suggests that the service industry followed by manufacturing jobs accounts for nearly half of all prevalence for occupationally related LBDs. The figure also indicates that handling of containers and worker motions or position assumed during work are very often associated with LBDs in industry. Therefore, these data strongly suggest that occupational factors can be related to risk of LBDs.
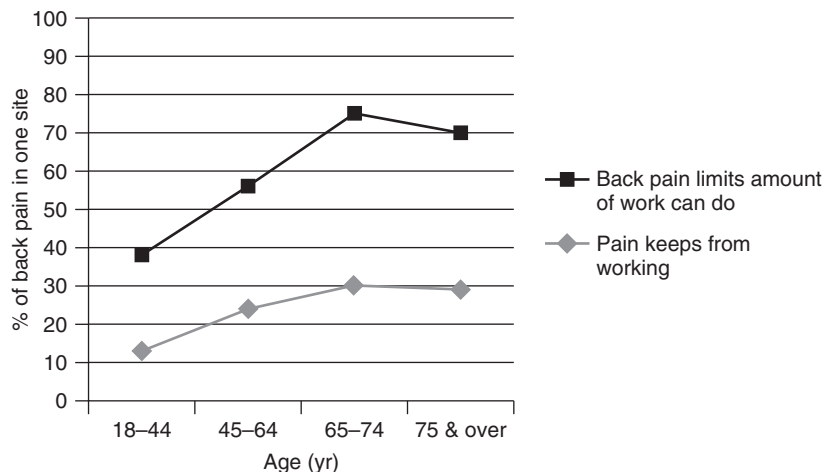


**Figure 1**   Age distribution and its relationship to work limitations.
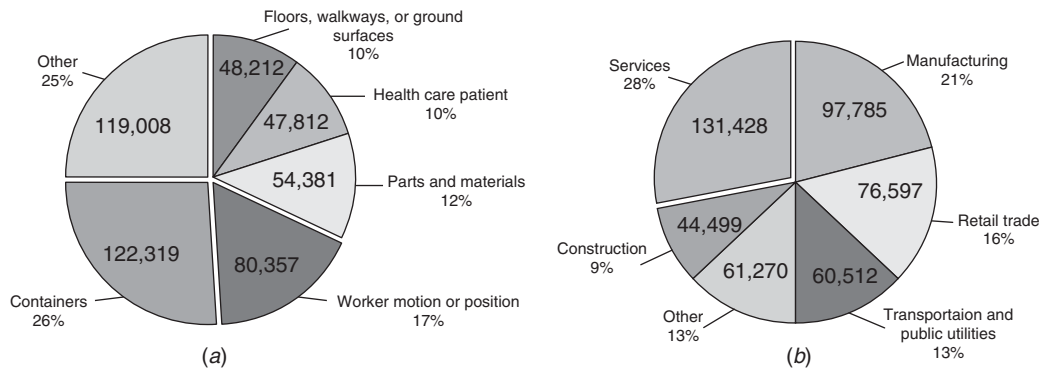
**Figure 2** (*a*) Number and distribution of back cases with days away from work in private industry by industry division during 1997 (NIOSH, 2000). (*b*) Number and distribution of back cases with days away from work in private industry by source of the disorder during 1997 (Jacobs, 2008).

## 3 EPIDEMIOLOGY OF WORK RISK FACTORS

Numerous literature reviews have endeavored to identify specific risk factors that may increase the risk of LBDs in the workplace. One of the first attempts at consolidating this information was performed by the NIOSH. In this critical review of the epidemiological evidence associated with musculoskeletal disorders (NIOSH, 1997) five categories of risk factors were evaluated. This evaluation suggested that strong evidence existed for an association between LBDs and lifting/forceful movements and LBDs and whole-body vibration. In addition, the evaluation concluded that there was significant evidence establishing associations between heavy physical work and awkward postures and back problems. Additionally, insufficient evidence was available to make any conclusions between static work postures and LBD risk.

Independent methodologically rigorous literature reviews by Hoogendoorn and colleagues (1999) were able to support these conclusions. Specifically, they concluded that manual materials handling, bending and twisting, and whole-body vibration were all significant risk factors for back pain.

Numerous investigations have attempted to assess the potential dose–response relationship among work risk factors and LBP. In particular, studies have been interested in the existence of an occupational "cumulative load" relationship with LBD. Two studies (Kumar, 1990; Norman et al., 1998) suggested the existence of such a cumulative load –LBD relationship in the workplace, although Videman et al. (1990) suggested that this relationship might not be a linear relationship. Videman et al. found that the relationship between history of physical loading due to occupation (cumulative load) and history of LBP was "J shaped" with sedentary jobs being associated with moderate levels of risk, heavy work being associated with the greatest degree of risk, and moderate exposure to loading being associated with the lowest level of risk (Figure 3). Seidler and colleagues (2001) have suggested a multifactor relationship with risk in that the combination of occupational lifting, trunk flexion, and duration of the activities significantly increased risk.

Recent studies have been able to identify risk with high levels of sensitivity and specificity when continuous dynamic biomechanical measures are employed (Marras et al., 2010b, 2010c). These efforts indicated that collective exposure to dynamic sagittal bending moments above 49 N-m, lateral trunk velocities greater than 84.1 deg/s, and exposure to the moment occurring after the midway point of the lift (more than 47.6% of the lift duration) yielded a sensitivity of 85% and a specificity of 87.5% in its ability to identify jobs resulting in reduced spine function.

Studies have also implicated psychosocial factors in the workplace as work risk factors for LBDs (Bigos et al., 1991; Bongers et al., 1993; Hoogendoorn et al., 2000; Karasek et al., 1998; van Poppel et al., 1998). Studies have indicated that monotonous work, high perceived work load, time pressure, low job satisfaction, and lack of social support were all related to LBD risk. Yet, the specific relationship with LBD appears to be unclear. Davis and Heaney (2000) found that the impact of psychosocial factors was diminished, although still significant, once biomechanical factors were accounted for in the study designs.
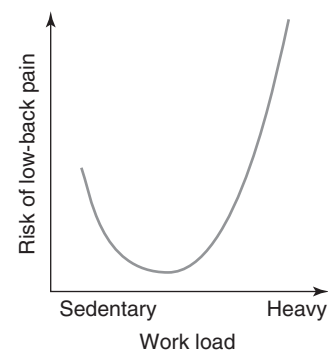


**Figure 3** Relationship between risk of LBP and work intensity exposure.

Several secondary prevention investigations of LBD have begun to explore the interaction between LBDs, physical factors, and psychosocial factors. Frank and colleagues (1996) as well as Waddell (1992, 1999) have concluded that much of LBP treatment is multidimensional. Primary prevention epidemiological studies have indicated that multiple categories of risk, such as physical stressors and psychosocial factors, play a role in LBD risk (Krause et al., 1998). Tubach and colleagues (2002) have reported that low social support at the workplace and bending at work were strongly associated with extended work absence due to LBP.

Perhaps the most comprehensive review of the epidemiological literature was performed by the National Research Council/Institute of Medicine (NRC, 2001). This assessment concluded that there is a clear relationship between LBDs and physical load imposed by manual material handling, frequent bending and twisting, physically heavy work, and whole-body vibration. Using the concept of attributable risk (attributable fraction), this analysis was able to determine the portion of LBP that would have been avoided if workers were not exposed to specific risk factors. As indicated in Table 1, the vast majority of high-quality epidemiological studies have associated LBDs with these risk factors and as much as two-thirds of risk can be attributed to materials handling activities. It was concluded that preventive measures may reduce the exposure to risk factors and reduce the occurrence of back problems.

## 4 OCCUPATIONAL BIOMECHANICS LOGIC

While epidemiological findings help us understand what exposure *factors* could be associated with work-related LBDs, the literature is problematic in that it cannot prescribe an optimal level of exposure in order to minimize risk. The previous section concluded that moderate levels of exposure are least risky for LBDs; however, we do not know what, precisely, constitutes moderate levels of exposure. The National Research Council/Institute of Medicine's (NRC, 2001) review of epidemiological evidence and LBDs states that "epidemiologic evidence itself is not specific enough to provide detailed, quantitative guidelines for design of the workplace, job, or task." This lack of specificity results from a dearth of the

**Table 1 Summary of Epidemiological Evidence with Risk Estimates (Attributable Fraction) of Associations with Work-Related Factors Associated with LBDs**

| Work-Related Risk Factor | Attributable Fraction (%) Range |
|---|---|
| Manual material handling | 11–66 |
| Frequent bending and twisting | 19–57 |
| Heavy physical load | 31–58 |
| Static work posture | 14–32 |
| Repetitive movements | 41 |
| Whole-body vibration | 18–80 |

continuous exposure measures. Most epidemiological studies have documented workplace exposures in a binary fashion where they document if a specific threshold of exposure has been exceeded. For example, many studies document whether workers lift more than 25 lb or not. Without continuous measures, it is impossible to ascertain the specific "levels" of exposure that would be associated with an increased risk of LBDs (NRC, 2001). In addition, from a biomechanical standpoint, we know that risk is a much more complex issue. We need to understand the load origin in terms of distance from the body and height off the floor as well as the load destination location if we are to understand the forces imposed on the body through the lifting task. Defining risk is most likely multidimensional. Hence, in order to more fully understand "how much exposure is too much exposure" to risk factors, it is necessary to understand *how* work-related factors interact and lead to LBDs. Thus, causal pathways are addressed through biomechanical and ergonomic analyses. Collectively, the biomechanical literature as a whole provides specificity of exposure and a promising approach to controlling LBD risk in the workplace.

Biomechanical logic provides a logic structure to help us understand the mechanisms that might effect the development of a LBD. At the center of this logic is the notion that risk can be defined by comparing the load imposed upon a structure with the tolerance of that same structure. As shown in Figure 4*a*, McGill (1997) suggests that, during work, the structures and tissues of the spine undergo a loading pattern with each repeated job cycle. When the magnitude of the load imposed upon a structure or tissue exceeds the structural tolerance limit, tissue damage occurs. The tissue damage might be capable of setting off the sequence of events that could lead to LBD. With this logic, if the magnitude of the imposed load is below the structural tolerance, the task can be considered free of risk to the tissue. The magnitude of the distance between the structure loading and the tolerance can be thought of as a safety margin. On the other hand, if the load exceeds the tolerance, significant risk is present.

Biomechanics reasoning can also be employed to describe the processes believed to be at play during cumulative trauma exposure. When exposed to repetitive exertions, one would expect the tolerance to be subject to degradation over time (Figure 4*b*). Yet, as the work is performed repeatedly, we would expect that the loading pattern would remain relatively constant, whereas with overuse we would expect the tolerance limit to drop over time. This process would make it more probable that the tissue load exceeds the tissue tolerance and trigger a potential disorder.

## 5 BIOMECHANICS OF RISK

There are numerous pathways to the pain perception associated with LBDs. These pain pathways are the key to understanding how tissue loading results in LBP. In addition, if one appreciates how pain is related to the factors associated with tissue loading, then one can use
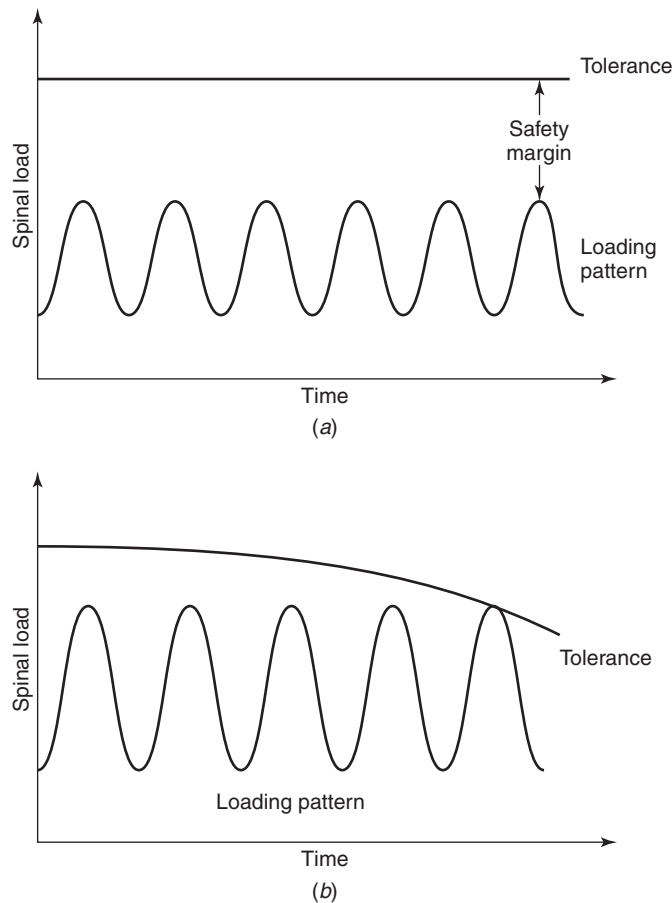
**Figure 4** Biomechanical load–tolerance relationships. (*a*) When the tolerance exceeds the load, the situation is considered safe with the distance between the two benchmarks considered a safety margin. (*b*) Cumulative trauma occurs when the tolerance decreases over time.

this knowledge to minimize the exacerbation of pain in workplace design. Thus, this knowledge forms the basis of ergonomics thinking. One can quantitatively target the limits above which a pain pathway is initiated as a tolerance limit for ergonomic purposes. Although these pathways have not been explicitly defined, designing tasks relative to these general principles is appealing since they represent biologically plausible mechanisms that are consistent with the injury association derived from the epidemiological literature.

Three general pain pathways are believed to be present for the spine that may affect the design of the workplace. These pathways are related to (1) structural and tissue stimulation, (2) physiological limits, and (3) psychophysical acceptance. It is expected that each of these pathways have different tolerances to the mechanical loading of the tissue. Thus, in order to optimally design a workplace, one must orient the specific tasks so that the ultimate tolerances within each of these categories are not exceeded.

## 5.1 Relationship between Tissue Stimulation and Pain

There are several structures in the back that when stimulated are capable of initiating pain perception. Both cellular and neural mechanisms can initiate and exacerbate pain perception. Several investigations have described the neurophysiological and neuroanatomical origins of back pain (Bogduk, 1995; Cavanaugh, 1995; Cavanaugh et al., 1997; Kallakuri et al., 1998; Siddall & Cousins, 1997b). These pathways involve the application of force or pressure on a structure that can directly stimulate pain receptors and can trigger the release of pain-stimulating chemicals.

Pain pathways in the low back have been identified for pain originating from the facet joints, disc, longitudinal ligaments, and sciatica. Facet joint pain is believed to be associated with the distribution of small nerve fibers and endings in the lumbar facet joint, nerves containing substance P (a pain-enhancing biochemical), high-threshold mechanoreceptors in the facet joint capsule,

and sensitization and excitation of nerves in the facet joint and surrounding muscle when the nerves were exposed to inflammatory biochemicals (Dwyer et al., 1990; Ozaktay et al., 1995; Yamashita et al., 1996). The pathway for disc pain is believed to activate through an extensive distribution of small nerve fibers and free nerve endings in the superficial annulus of the disc as well as in small fibers and free nerve endings in the adjacent longitudinal ligaments (Bogduk, 1991, 1995; Cavanaugh et al., 1995; Kallakuri et al., 1998). Sciatic pain is thought to be associated with mechanical stimulation of some of the spine structures. Moderate pressure placed on the dorsal root ganglia can result in vigorous and long-lasting excitatory discharges that could easily explain sciatica. In addition, sciatica might be explained through excitation of the dorsal root fibers when the ganglia are exposed to the nucleus pulposus. Stimulation and nerve function loss in nerve roots exposed to phospholipase A$_2$ could also explain the pain associated with sciatica (Cavanaugh et al., 1997; Chen et al., 1997; Ozaktay et al., 1998).

Studies are demonstrating the importance of proinflammatory agents such as tumor necrosis factor alpha (TNF$\alpha$) and interleukin-1 (IL-1) (Dinarello, 2000) in the development of pain. Proinflammatory agents are believed to upregulate vulnerability to inflammation under certain conditions and set the stage for pain perception. Thus, it is thought that mechanical stimulation of tissues can initiate this sequence of events and thus become the initiator of pain. It may be possible to consider the role of these agents in a load–tolerance model where tolerance may be considered the point at which these agents are upregulated. A preliminary study (Yang et al., 2010) has demonstrated that loads on spine structures due to occupational tasks are capable of initiating such a chemical reaction.

This body of work is providing a framework for a logical link between the mechanical stimulation of spinal tissues and structures and the sensation of LBP that is the foundation of occupational biomechanics and ergonomics.

## 5.2 Functional Lumbar Spinal Unit Tolerance Limits

Individual structure tolerances within lumbar functional spinal units are often considered, collectively, as part of the structural support system. The vertebral body can withstand fairly large loads when compressed, and since the end plate is usually the first structure to yield, the end-plate tolerance is often considered as a key marker of spine damage leading to pain. A review by Jager (1987) indicated the compressive tolerance of the end plate reported in the literature can be large (over 8 kN), especially in upright postures, but highly variable (depending greatly on age) with some specimens indicating failure at 2 kN. Damage to human vertebral cancellous bone often results from shear loading and the ultimate strength is correlated with tissue stiffness when exposed to compressive loading (Fyhrie and Schaffler, 1994). Bone failure typically occurs along with disc herniation and annular delamination (Gunning et al., 2001). Thus, damage to the bone itself appears

to often be part of the cascading series of events associated with LBP (Brinkmann, 1985; Kirkaldy-Willis, 1998; Siddall and Cousins, 1997b).

There are several lines of thinking about how vertebral end plate microfractures can lead to low-back problems. One line of thinking contends that the health of the vertebral body end plate is essential for proper mechanical functioning of the spine. When damage occurs to the end plate, nutrient supply is restricted to the disc and this can lead to degeneration of the disc fibers and disruption of spinal function (Moore, 2000). The literature supports the notion that the disruption of nutrient flow is capable of initiating a cascading series of events leading to LBP (Brinkmann, 1985; Kirkaldy-Willis, 1998; Siddall and Cousins, 1997a, 1997b). The literature suggests that the end plate is often the first structure to be damaged when the spine is loaded, especially at low load rates (Brinkmann et al., 1988; Callaghan and McGill, 2001; Holmes et al., 1993; Moore, 2000). Vertebral end-plate tolerance levels have been documented in numerous investigations. End-plate failure typically occurs when the end plate is subjected to compressive loads of 5.5 kN (Holmes et al., 1993). In addition, end-plate tolerances decrease by 30–50% with exposure to repetitive loading (Brinkmann et al., 1988) and suggests the disc is affected by cumulative trauma. The literature suggests that spine integrity can also be weakened by anterior–posterior (forward–backward) shear loading. Shear limit tolerance levels beginning at 1290–1770 N for soft tissue and 2000–2800 N for hard tissue have been reported for the spinal structures (Begeman et al., 1994; Krypton et al., 1995).

Load-related damage might also be indicated by the presence of Schmorls nodes in the vertebral bodies. Some have suggested that Schmorls nodes could be remnants of healed end-plate fractures (Vernon-Roberts and Pirie, 1973, 1977) and might be linked to trauma (Kornberg, 1988; Vernon-Roberts and Pirie, 1973).

Position or posture of the spine is also closely related to end-plate tolerance to loading. Flexed spine postures greatly reduce end-plate tolerance (Adams and Hutton, 1982; Gunning et al., 2001). In addition, trunk posture has been documented as an important consideration for occupational risk assessment. Industrial surveillance studies by Punnett et al. (1991), Marras et al. (1993a, 1995), and Norman and colleagues (1998) have all suggested that LBD risk increases when trunk posture deviates from a neutral upright posture during the work cycle.

It also appears that individual factors also influence end-plate integrity. Most notably, age and gender appear to greatly influence the biomechanical tolerance of the end plate (Jager et al., 1991) in that age and gender are related to bone integrity. Brinkmann et al. 1988 have demonstrated that bone mineral content and end-plate cross-sectional area are responsible for much of the variance in tolerance (within 1 kN).

There is little doubt that the disc can be subject to damage with sufficient loading. Disc herniations occur frequently when the spine is subject to compression and positioned in an excessively flexed posture (Adams and Hutton, 1982). In addition, repeated flexion, even under

moderate compressive loading conditions, can produce repeated disc herniations (Callaghan and McGill, 2001). Under anterior–posterior shear conditions avulsion of the lateral annulus can occur (Yingling and McGill, 1999a, 1999b). The torsion tolerance limit of the disc can be exceeded at loads as low as 88 N-m in an intact disc and 54 N-m in the damaged disc (Adams and Hutton, 1981; Farfan et al., 1970). When the spine is loaded in multiple dimensions simultaneously, risk also increases. The literature indicates that when the spine assumes complex spinal postures such as hyperflexion with lateral bending and twisting, disc herniation is increasingly likely (Adams and Hutton, 1985; Gordon et al., 1991).

Disc tolerance can also be associated with diurnal cycles or time of day when the lifting exposure occurs. Snook and associates (1998) found that flexion early in the day was associated with an increased risk of a LBP report. In addition, Fathallah and colleagues (1995) found similar results reporting that risk of injury was greater early in the day when disc hydration was at a high level. Therefore, the temporal component of risk associated with work exposure must be considered when assessing risk.

This brief review of the spine's tolerance limits indicated that the tolerance limits of the functional lumbar spinal unit vary considerably. Adams et al. (1993) describe a process where repeated vertebral microfractures and scarring of the end plate can lead to an interruption of nutrient flow to the disc. This process can result in weakening of the annulus that can result in protrusion of the disc into the surrounding structures. In addition, the weakened disc can result in spinal instability. The end plate and most of the inner portions of the annulus are not capable of sensing pain. However, once disc protrusion and/or disc instability occurs, loads are transmitted to the outer portions of the annulus and surrounding tissues. These structures are indeed capable of sensing pain. In addition, inflammatory responses can occur and nociceptors of surrounding tissues can be further sensitized and stimulated, thus initiating a sequence of events resulting in pain. Quantitative ergonomics approaches attempt to design work tasks so that spine loads are well within the tolerance limits of the spine structures.

While a wide range of tolerance limits have been reported for the functional lumbar spinal unit, most authorities have adopted the NIOSH lower limit of 3400 N for compression as the protective limit for most male workers and 75% of female workers (Chaffin et al., 1999). This limit represents the point at which end-plate microfracture is believed to begin within a large, diverse, population of workers. Similarly, 6400 N of compressive load represents the limit at which 50% of workers would be at risk (NIOSH, 1981). Furthermore, current quantitative assessments are recognizing the complex interaction of spine position, frequency, and complex spine forces (compression, shear, and torsion) as more realistic assessments of risk. However, these complex relationships have yet to find their way into ergonomic assessments nor have they resulted in best practices or standards.

## 5.3 Ligament Tolerance Limits

Ligament tolerances are affected primarily by load rate (Noyes et al., 1994). Avulsion occurs at low load rates and tearing occurs mostly at high load rates. Therefore, load rate may explain the increased risk associated with bending kinematics (velocity) that have been identified as risk factors in surveillance studies (Fathallah et al., 1998) as well as injuries from slips or falls (McGill, 1997). Posture appears to also play a role in tolerance. While loaded the architecture of the interspinous ligaments can result in significant anterior shear forces imposed on the spine during forward flexion (Heylings, 1978). This finding is consistent with field observations of risk (Marras et al., 1993a, 1995, 1993b, 2000b; Norman et al., 1998; Punnett et al., 1991). Field observations have identified 60 N-m as the point at which tissue damage is initiated (Adams and Dolan, 1995). Similarly, surveillance studies (Marras et al., 1993a, 1995) have identified exposures to external load moments of at least 73.6 N-m as being associated with greater risk of occupationally related LBP reporting. Also reinforcing these findings was a study by Norman and colleagues (1998), who reported nearly 30% greater load moment exposure in those jobs associated with risk of LBP. In this study, mean moment exposure associated with the back pain cases was 182 N-m of *total* load moment (load lifted plus body segment weights).

Spine curvature or lumbar lordosis may also affect the loading and tolerance of the spinal structures. Findings from Canadian researchers have demonstrated that when lumbar spinal curvature is maintained during bending the extensor muscles support the shear forces of the torso. However, when the spine (and posterior ligaments) are flexed during bending significant shear can be imposed on the ligaments (McGill et al., 1994; McGill and Norman, 1987; Potvin et al., 1991). The shear tolerance of the spine can be easily exceeded (2000–2800 N) exceeded when the spine is in full flexion (Krypton et al., 1995).

As with most living tissues, temporal factors play a large role in recovery of the ligaments. Solomonow and colleagues have found that ligaments require long periods of time to regain structural integrity. During this recovery period it is likely that compensatory muscle activities are recruited (Gedalia et al., 1999; Solomonow et al., 1998, 1999, 2000, 2002; Stubbs et al., 1998; Wang et al., 2000), and these muscle activities can easily increase spine loading. Required recovery time has been estimated to be several times the loading period duration and thus may easily exceed the typical work–rest cycles common in industry.

## 5.4 Facet Joint Tolerance

The facet joints are capable of supporting a significant portion of the load transmitted along the spinal column. Therefore, it is important to understand the tolerance limits of this structure. Facet joint failure can occur in response to shear loading of the spine. McGill and colleagues have reported that much of the tissues that load the facets are capable of producing significant horizontal forces and thus place these structures at risk

during occupational tasks (McGill, 2002). Crypton et al. (1995) estimate a shear tolerance for the facet joints of 2000 N. These findings are consistent with industrial observations indicating that exposure to lateral motions and shear loading was associated with an increased risk of LBDs (Marras et al., 1993a, 1995; Norman et al., 1998). In addition, laboratory-based studies confirm that exposure to high lateral velocity can result in significant lateral shear forces in the lumbar spine (Marras and Granata, 1997a).

Torsion can also load the facet joints to a failure point (Adams and Hutton, 1981). Exposure to excessive twisting moments, especially when combined with high-velocity motions, have been associated with excessive tissue loading (Marras & Granata, 1995; McGill, 1991; Pope et al., 1986, 1987). Field-based studies have also identified these movements as being associated with high-risk (for LBP) tasks (Marras et al., 1993a, 1995; Norman et al., 1998). The load imposed upon the spinal tissues when exposed to torsional moments also depends greatly upon the posture of the spine, with greater loads occurring when deviated postures (from neutral) are adopted (Marras and Granata, 1995). The specific structure loading pattern depends upon both the posture assumed during the task as well as curvature of the spine since a great amount of load sharing occurs between the apophyseal joints and the disc (Adams and Dolan, 1995). Therefore, spine posture dictates both the nature of spine loading as well as the degree of risk to the facet joints or the disc.

## 5.5 Adaptation

An important consideration in the interpretation of the load–tolerance relationship of the spine is that of adaptation. Wolff's law suggests that tissues adapt and remodel in response to the imposed load. In the spine, adaptation in response to load has been reported for bone (Carter, 1985), ligaments (Woo et al., 1985), disc (Porter et al., 1989), and vertebrae (Brinkmann et al., 1989). Adaptation may explain the observation that the greatest risk has been associated with jobs involving both high loads and low levels of spinal load, whereas jobs associated with moderate spine loads appear to enjoy the lowest levels of risk (Chaffin and Park, 1973; Videman et al., 1990). Hence, there appears to be an optimal loading zone for the spine that minimizes risk of exceeding the tolerance limit.

## 5.6 Psychophysical Limits as a Tolerance Threshold

Tolerance limits used in biomechanical assessment of tissue are typically derived from cadaveric testing studies. While these mechanical limits for tissue strength are probably reasonable for the analysis of tasks resulting in acute-trauma-type injuries, the application of cadaver-based tolerances to repetitive tasks is less logical. Repetitive loading must consider the influence of repeated weakening of the structure as well as the impact of tissue repair. Since adaptation is a key distinction in living tissue, quantitative analyses of the load–tolerance relationship under repeated loading become problematic. Highly dynamic tasks may be difficult to characterize

through quantitative biomechanical analyses and their injury pathway may be poorly understood. Hence, there is a dearth of biomechanical tolerance limit data that describe how living tissues respond to such repeated loading conditions.

In circumstances where mechanical tolerances are not known, an alternative approach to establishing tolerance limits has been to use the psychophysical limit as a tolerance limit. Psychophysics has been used as a means of strength testing where subjects are asked to progressively adjust the amount of load they can push, pull, lift, or carry until they feel that the magnitude of the force exertion would be acceptable to them over an 8-h work shift. Work variables included in such evaluations typically include measures such as lift origin, height, load dimensions, frequency of exertion, push/pull heights, and carrying distance. These variables are systematically altered to yield a database of acceptable conditions or thresholds of acceptance that would be tolerable for a specified range of male and female workers. These data are typically presented in table form and indicate the percentage of subjects who would find a particular load condition acceptable for a given task. Snook and colleagues are best known for publishing extensive descriptions of these psychophysical tolerances (Ciriello et al., 1990; Snook, 1978, 1985a, 1985b, 1987; Snook and Ciriello, 1991). Table 2 shows an example of such data for a pulling task.

Very few investigations have reported whether the design of work tasks using these psychophysical tolerance limits results in a minimization of LBP reports at work. However, one study by Snook (1978) has reported that low-back-related injury claims were three times more prevalent in jobs exceeding the psychophysically determined strength tolerance of 75% of men compared with jobs demanding less strength.

## 5.7 Physiological Tolerance Limits

Energy expenditure limits can also be used as tolerance limits for those jobs where physiological load limits the workers' ability to perform the work. These limits are associated with the ability of the body to deliver oxygen to the muscles. When muscles go into oxygen debt, insufficient release of adenosine triphosphate (ATP) occurs within the muscle and prolonged muscle contractions cannot be sustained. Therefore, under these extremely high energy expenditure work conditions, aerobic capacity can be considered as a physiological tolerance limit for LBP.

The NIOSH has established physiological criteria for limiting heavy physical work based upon high levels of energy expenditure (Waters et al., 1994). These criteria established an energy expenditure rate of 9.5 kcal/min as a baseline for maximum aerobic lifting capacity. Seventy percent of this baseline limit is considered the aerobic tolerance limit for work that is defined primarily as "arm work." Fifty percent, 40%, and 33% of the baseline energy expenditure have been designated as the tolerance limits for lifting task durations of 1, 1–2, and 2–8 h, respectively. While limited epidemiological evidence is available to support these limits, Cady and associates (1979, 1985) have demonstrated the

**Table 2** Example of psychophysical table used to determine the acceptable load an individual is willing to accept. The table indicates the maximum amount of pull force acceptable for males and females under various conditions. From Stover H. Snook and Vincent M Ciriello, The design of manual handling tasks; revised tables of maximum acceptable weights and forced, Ergonomics, 1991, Vol. 34, No. 9, 1197–1213.

| Height[†] | Percent[⇓] | 2.1 m pull One pull every | | | | | | | 7.6 m pull One pull every | | | | | | | 15.2 m pull One pull every | | | | | | | 30.3 m pull One pull every | | | | | 45.7 m pull One pull every | | | | | 61.0 m pull One pull every | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 6 | 12 | 1 | 2 | 5 | 30 | 8 | 15 | 22 | 1 | 2 | 5 | 30 | 8 | 25 | 35 | 1 | 2 | 5 | 30 | 8 | 1 | 2 | 5 | 30 | 8 | 1 | 2 | 5 | 30 | 8 | 2 | 5 | 30 | 8 |
| | | s | | min | | | | h | s | | min | | | | h | s | | min | | | | h | min | | | | h | min | | | | h | min | | | h |
| | | | | | | | | | | | | | | **Initial Forces*** | | | | | | | | | | | | | | | | | | | | | | |
| 144 | 90 | 14 | 16 | 18 | 18 | 19 | 19 | 23 | 11 | 13 | 16 | 16 | 17 | 18 | 21 | 13 | 15 | 15 | 15 | 16 | 17 | 20 | 12 | 13 | 15 | 15 | 19 | 10 | 11 | 13 | 13 | 16 | 10 | 11 | 11 | 14 |
| | 75 | 17 | 19 | 22 | 22 | 23 | 24 | 28 | 14 | 15 | 20 | 20 | 21 | 21 | 26 | 16 | 18 | 19 | 19 | 20 | 20 | 24 | 14 | 16 | 19 | 19 | 23 | 12 | 14 | 16 | 16 | 20 | 12 | 14 | 14 | 17 |
| | 50 | 20 | 23 | 26 | 26 | 28 | 28 | 33 | 16 | 18 | 24 | 24 | 25 | 26 | 31 | 19 | 21 | 22 | 22 | 24 | 24 | 29 | 17 | 19 | 22 | 22 | 27 | 15 | 16 | 19 | 19 | 24 | 14 | 16 | 16 | 20 |
| | 25 | 24 | 27 | 31 | 31 | 32 | 33 | 39 | 19 | 21 | 28 | 28 | 29 | 30 | 36 | 22 | 25 | 26 | 26 | 28 | 28 | 33 | 20 | 22 | 26 | 26 | 32 | 17 | 19 | 22 | 22 | 28 | 16 | 19 | 19 | 24 |
| | 10 | 26 | 30 | 34 | 34 | 36 | 37 | 44 | 21 | 24 | 31 | 31 | 33 | 33 | 40 | 24 | 28 | 29 | 29 | 31 | 31 | 38 | 22 | 25 | 29 | 29 | 37 | 20 | 22 | 25 | 25 | 31 | 18 | 21 | 21 | 27 |
| 95 | 90 | 19 | 22 | 25 | 25 | 27 | 27 | 32 | 15 | 18 | 23 | 23 | 24 | 24 | 29 | 18 | 20 | 21 | 21 | 23 | 23 | 28 | 16 | 18 | 21 | 21 | 26 | 14 | 16 | 18 | 18 | 23 | 13 | 16 | 16 | 19 |
| | 75 | 23 | 27 | 31 | 31 | 32 | 33 | 39 | 19 | 21 | 28 | 28 | 29 | 30 | 36 | 22 | 25 | 26 | 26 | 28 | 28 | 33 | 20 | 22 | 26 | 26 | 32 | 17 | 19 | 22 | 22 | 28 | 16 | 19 | 19 | 24 |
| | 50 | 28 | 32 | 36 | 36 | 39 | 39 | 47 | 23 | 26 | 33 | 33 | 35 | 35 | 42 | 26 | 29 | 31 | 31 | 33 | 33 | 40 | 24 | 27 | 31 | 31 | 38 | 20 | 23 | 27 | 27 | 33 | 20 | 23 | 23 | 28 |
| | 25 | 33 | 37 | 42 | 42 | 45 | 45 | 54 | 26 | 3 | 39 | 39 | 41 | 41 | 49 | 30 | 34 | 36 | 36 | 38 | 39 | 46 | 27 | 31 | 36 | 36 | 45 | 24 | 27 | 31 | 31 | 38 | 23 | 26 | 26 | 33 |
| | 10 | 37 | 42 | 48 | 48 | 51 | 51 | 61 | 30 | 33 | 43 | 43 | 46 | 47 | 56 | 33 | 38 | 41 | 41 | 43 | 44 | 52 | 31 | 35 | 40 | 40 | 50 | 27 | 30 | 35 | 35 | 43 | 26 | 30 | 30 | 37 |
| 64 | 90 | 22 | 25 | 28 | 28 | 30 | 30 | 36 | 18 | 20 | 26 | 26 | 27 | 28 | 33 | 20 | 23 | 24 | 24 | 26 | 26 | 31 | 18 | 21 | 24 | 24 | 30 | 16 | 18 | 21 | 21 | 26 | 15 | 18 | 18 | 22 |
| | 75 | 27 | 30 | 34 | 34 | 37 | 37 | 44 | 21 | 24 | 31 | 31 | 33 | 34 | 40 | 24 | 28 | 29 | 29 | 31 | 32 | 38 | 22 | 25 | 29 | 29 | 36 | 19 | 22 | 25 | 25 | 31 | 19 | 21 | 21 | 27 |
| | 50 | 32 | 36 | 41 | 41 | 44 | 44 | 53 | 25 | 29 | 37 | 37 | 40 | 40 | 48 | 29 | 33 | 35 | 35 | 37 | 38 | 45 | 27 | 30 | 35 | 35 | 43 | 23 | 26 | 30 | 30 | 37 | 22 | 26 | 26 | 32 |
| | 25 | 37 | 42 | 48 | 48 | 51 | 51 | 61 | 30 | 34 | 44 | 44 | 46 | 47 | 56 | 34 | 39 | 41 | 41 | 43 | 44 | 52 | 31 | 35 | 41 | 41 | 50 | 27 | 30 | 35 | 35 | 43 | 26 | 30 | 30 | 37 |
| | 10 | 42 | 48 | 54 | 54 | 57 | 58 | 69 | 33 | 38 | 49 | 49 | 52 | 53 | 63 | 38 | 43 | 46 | 46 | 49 | 49 | 59 | 35 | 39 | 46 | 46 | 57 | 30 | 34 | 39 | 39 | 49 | 29 | 34 | 34 | 42 |
| | | | | | | | | | | | | | | **Sustained Forces**** | | | | | | | | | | | | | | | | | | | | | | |
| 144 | 90 | 8 | 10 | 12 | 13 | 15 | 15 | 18 | 6 | 8 | 10 | 11 | 12 | 12 | 15 | 7 | 8 | 9 | 9 | 10 | 11 | 13 | 7 | 8 | 9 | 11 | 13 | 6 | 7 | 8 | 9 | 10 | 6 | 6 | 7 | 9 |
| | 75 | 10 | 13 | 16 | 17 | 19 | 20 | 23 | 8 | 10 | 13 | 14 | 16 | 16 | 19 | 9 | 10 | 12 | 12 | 14 | 14 | 17 | 9 | 10 | 12 | 14 | 16 | 7 | 9 | 10 | 11 | 14 | 7 | 8 | 10 | 11 |
| | 50 | 13 | 16 | 20 | 21 | 23 | 24 | 28 | 10 | 13 | 16 | 17 | 19 | 20 | 23 | 11 | 13 | 14 | 15 | 17 | 17 | 20 | 11 | 13 | 15 | 17 | 20 | 9 | 11 | 12 | 14 | 17 | 9 | 10 | 12 | 14 |
| | 25 | 15 | 20 | 24 | 25 | 28 | 29 | 34 | 12 | 15 | 20 | 20 | 23 | 24 | 28 | 13 | 15 | 17 | 18 | 20 | 21 | 24 | 13 | 15 | 18 | 20 | 24 | 11 | 13 | 15 | 17 | 20 | 11 | 12 | 14 | 17 |
| | 10 | 17 | 22 | 27 | 28 | 32 | 33 | 39 | 14 | 17 | 22 | 23 | 26 | 27 | 32 | 14 | 17 | 19 | 20 | 23 | 24 | 28 | 15 | 17 | 20 | 23 | 27 | 12 | 14 | 17 | 19 | 23 | 12 | 14 | 16 | 19 |
| 95 | 90 | 10 | 13 | 16 | 17 | 19 | 20 | 24 | 8 | 10 | 13 | 14 | 16 | 16 | 19 | 9 | 10 | 12 | 12 | 14 | 14 | 17 | 9 | 10 | 12 | 14 | 17 | 7 | 9 | 10 | 12 | 14 | 7 | 9 | 10 | 12 |
| | 75 | 13 | 17 | 21 | 22 | 25 | 25 | 30 | 11 | 13 | 17 | 18 | 20 | 21 | 25 | 11 | 14 | 15 | 15 | 18 | 18 | 22 | 12 | 13 | 16 | 18 | 21 | 10 | 11 | 13 | 15 | 18 | 9 | 11 | 13 | 15 |
| | 50 | 16 | 21 | 26 | 27 | 31 | 32 | 37 | 13 | 17 | 21 | 22 | 25 | 26 | 31 | 14 | 17 | 19 | 19 | 22 | 23 | 27 | 14 | 17 | 19 | 22 | 26 | 12 | 14 | 16 | 19 | 22 | 12 | 14 | 16 | 18 |
| | 25 | 19 | 26 | 31 | 33 | 37 | 38 | 45 | 16 | 20 | 26 | 27 | 30 | 31 | 37 | 17 | 20 | 22 | 23 | 26 | 27 | 32 | 17 | 20 | 23 | 27 | 32 | 14 | 17 | 19 | 22 | 26 | 14 | 16 | 19 | 22 |
| | 10 | 22 | 29 | 36 | 37 | 42 | 43 | 51 | 18 | 23 | 29 | 31 | 34 | 36 | 42 | 19 | 23 | 26 | 27 | 30 | 31 | 37 | 19 | 23 | 27 | 31 | 36 | 16 | 19 | 22 | 25 | 30 | 16 | 19 | 21 | 25 |
| 64 | 90 | 11 | 14 | 17 | 18 | 20 | 21 | 25 | 9 | 11 | 14 | 15 | 17 | 17 | 20 | 9 | 11 | 12 | 13 | 15 | 15 | 18 | 9 | 11 | 13 | 15 | 18 | 8 | 9 | 11 | 12 | 15 | 8 | 9 | 10 | 12 |
| | 75 | 14 | 19 | 23 | 23 | 26 | 27 | 32 | 11 | 14 | 19 | 19 | 22 | 22 | 26 | 12 | 14 | 16 | 17 | 19 | 19 | 23 | 12 | 14 | 17 | 19 | 23 | 10 | 12 | 14 | 16 | 19 | 10 | 12 | 13 | 16 |
| | 50 | 17 | 23 | 28 | 29 | 32 | 34 | 40 | 14 | 18 | 23 | 24 | 27 | 28 | 33 | 15 | 18 | 20 | 21 | 23 | 24 | 28 | 15 | 18 | 21 | 24 | 27 | 13 | 15 | 17 | 20 | 23 | 12 | 14 | 16 | 20 |
| | 25 | 20 | 27 | 33 | 35 | 39 | 40 | 48 | 17 | 21 | 27 | 28 | 32 | 33 | 39 | 18 | 21 | 24 | 25 | 28 | 29 | 34 | 18 | 21 | 25 | 28 | 33 | 15 | 18 | 21 | 24 | 28 | 15 | 17 | 20 | 23 |
| | 10 | 23 | 31 | 38 | 40 | 45 | 46 | 54 | 19 | 24 | 31 | 32 | 37 | 38 | 45 | 20 | 24 | 27 | 28 | 32 | 33 | 39 | 21 | 24 | 28 | 32 | 38 | 17 | 20 | 24 | 27 | 32 | 17 | 20 | 23 | 27 |

[†]Vertical distance from floor to hands (cm).
[⇓]Percentage of industrial population.
*The force required to get an object in motion.
**The force required to keep an object in motion.
Italicized values exceed 8 h physiological criteria.

importance of aerobic capacity limits associated with back problems for firefighters.

## 5.8  Psychosocial Pathways

A body of literature has attempted to describe how psychosocial factors might relate to the risk of suffering a LBD. Psychosocial factors have been associated with risk of LBP in several reviews (Bongers et al., 1993; Burton et al., 1995) and some researchers have dismissed the role of biomechanical factors as a causal factor (Bigos et al., 1991). However, few studies have appropriately considered biomechanical exposure along with psychosocial exposure in these assessments. A study by Davis and Heaney (2000) demonstrated that no studies have been able to effectively assess both risk dimensions concurrently.

More sophisticated biomechanical assessments (Davis et al., 2002; Marras et al., 2000c) have shown that psychosocial stress has the capacity to influence biomechanical loading. These studies demonstrate that individual factors such as personality can interact with perception of psychosocial stress to increase trunk muscle coactivation and subsequently increase spine loading. Therefore, these studies provide evidence that psychosocial stress is capable of influencing LBP risk through a biomechanical pathway.

## 5.9  Spine-Loading Assessment

A critical part of evaluating the load–tolerance relationship and the subsequent risk associated with work is an accurate and quantitative assessment of the loading experienced by back and spine tissues. The tolerance literature suggests that it is important to understand the specific nature of the tissue loading, including the dimensions of tissue loading such as compression force, shear force in multiple dimensions, load rates, positions of the spine structures during loading, and frequency of loading. Hence, *accurate and specific* measures associated with spine loading are essential if one is to use this information to assess the potential risk associated with occupational tasks.

Currently, it is not practical to directly monitor the loads imposed upon living spine structures and tissues while workers are performing a work task. Alternatively, indirect assessments such as biomechanical models are typically used to estimate tissue loads. The goal of biomechanical models is to understand how exposure to external (to the body) loads results in internal (to the body) forces that may exceed specific tolerance limits. External loads are imposed on the musculoskeletal system through the external environment (e.g., gravity or inertia) and must be countered or overcome by the worker in order to perform work. Internal forces are supplied by the musculoskeletal structures within the body (e.g., muscles, ligaments) that must supply counterforces to support the external load. However, since the internal forces are typically at a severe biomechanical disadvantage (relative to the external moment), these internal forces can be very large and result in large force applications on spine tissues. Since these internal forces are so large it is extremely important to accurately assess the nature of these loads in order to appreciate

the risk of a musculoskeletal disorder. Several biomechanical modeling approaches have been employed for these purposes and these different approaches often result in significant different trade-offs between their ability to realistically and accurately assess spine loading associated with a task and ease of model use.

Early models used to assess spine loading during occupational tasks needed to make assumptions about which trunk muscles supported the external load during a lifting task (Chaffin and Baker, 1970; Chaffin et al., 1977). These initial models assumed that a single "equivalent" muscle vector within the trunk could characterize the trunk's internal supporting force (and thus define spine loading) required to counteract an external load lifted by a worker. These crude models assumed that a lift could be portrayed as a static equilibrium lifting situation and that no muscle coactivation occurs among the trunk musculature during lifting. The models employed anthropometric regression relationships to estimate body segment lengths representative of the general population. Two output variables were predicted that could be used in a load–tolerance assessment of work exposure. One commonly used model output involved spine compression that is typically compared to the NIOSH-established compression limits of 3400 and 6400 N. A second model output was population static strength capacity associated with six joints of the body. Specifically lumbosacral (L5/S1) joint strength was used to assess overexertion risk to the back. This model evolved into a personal computer–based model and was used for general assessments of materials handling tasks involving slow movements (that were assumed to be quasi-static) where excessive compression loads were suspected of contributing to risk. An example of the model program output is shown in Figure 5. Early field-based risk assessments of the workplace have used this method to assess spine loads on the job (Herrin et al., 1986).

As computational power became more available, workplace biomechanical models were expanded to account for the contribution of multiple internal muscle reactions in response to the lifting of an external load. The assessment of multiple muscles resulted in models that were much more accurate and realistic. In addition, the spine tolerance literature was beginning to recognize the significance of three-dimensional spine loads as compared to purely compression loads in characterizing potential risk. The multiple-muscle biomechanical models were capable of predicting spine compression forces as well as spine shear forces.

The first multiple-muscle system model was developed by Schultz and Andersson (1981). The model demonstrated how loads manipulated outside the body could impose large spinal loads upon the system of muscle within the torso due to the coactivation of trunk muscles necessary to counteract this external load. This model was able to predict asymmetric loading of the spine. Hence, the model represented an advancement in realism compared to previous models. However, the approach resulted in indeterminate solutions in that there were more muscle forces represented in the model than functional constraints available to uniquely solve
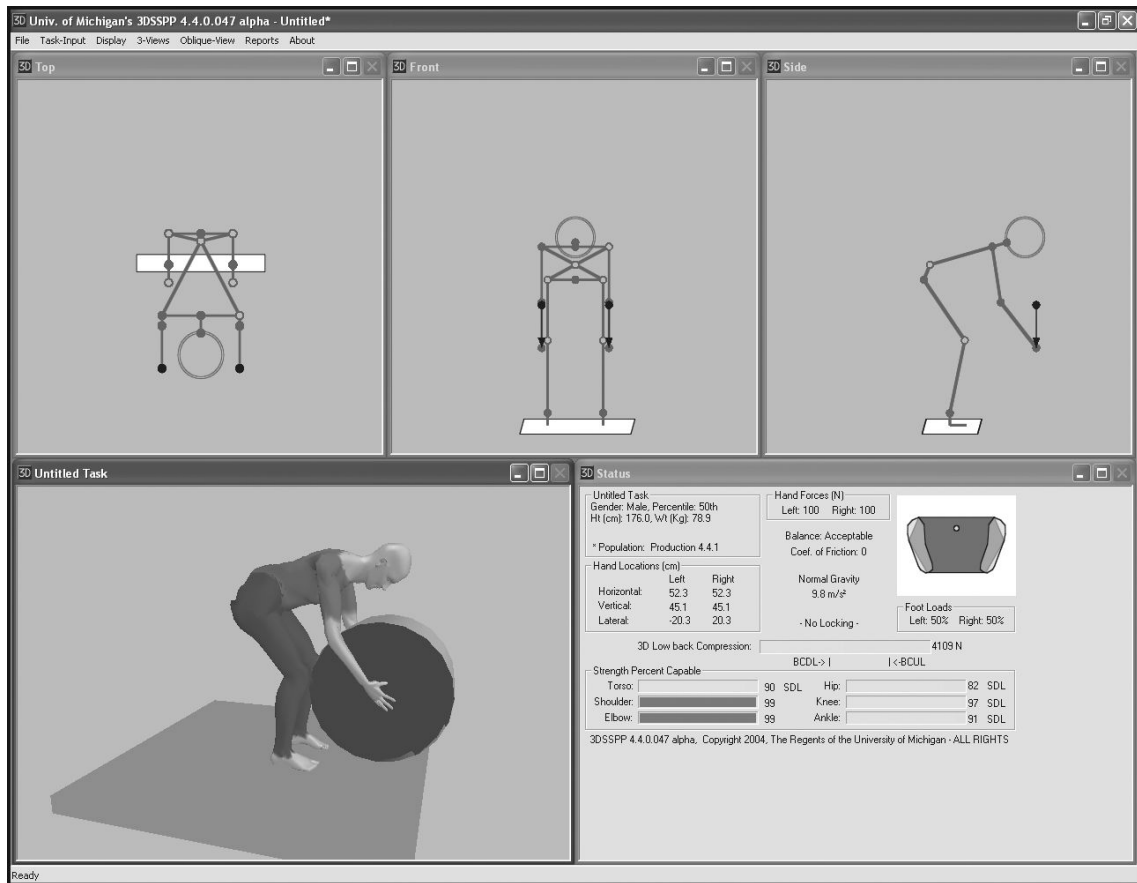
**Figure 5** Example of static strength prediction program used to assess spine load and strength requirements for a given task. (Courtesy of D. Chaffin.)

for the forces, so unique solutions were not apparent. Modeling efforts attempted to overcome this difficulty by assuming that certain muscles are inactive during the task (Bean et al., 1988; Hughes and Chaffin, 1995; Schultz et al., 1982). These efforts resulted in models that worked well for steady-state static representations of a lift but not for dynamic lifting situations (Marras et al., 1984).

Later efforts attempted to account for the influence of muscle coactivation upon spine loading under dynamic, complex lifting situations by directly monitoring muscle activity using electromyography (EMG) as an input to multiple-muscle models. EMG measures eliminated the problem of indeterminacy since specific muscle activities were uniquely defined through the neural activation of each muscle. Because of the use of direct muscle activity, these models were termed biologically assisted models. They were able not only to accurately assess compression and shear spine loads for a specific occupationally related movements (Granata and Marras, 1993, 1995a, 1999; Marras and Davis, 1998; Marras et al., 1998, 2001b; Marras and Granata, 1995, 1997b,

1997c; Marras and Sommerich, 1991a, 1991b; McGill, 1991, 1992a, 1992b) but also to predict differences among individuals so that variations in loading among a population could be evaluated (Granata et al., 1999; Marras et al., 2000c, 2000a, 2002; Marras and Granata, 1997a; Mirka and Marras, 1993) (Figure 6). These models were reported to have excellent external as well as internal validity (Granata et al., 1999; Marras et al., 1999c). The significance of accounting for trunk muscle coactivation when assessing realistic dynamic lifting was demonstrated by Granata and Marras (1995b). They found that not accounting for coactivation models could miscalculate spinal loading by up to 70%.

The disadvantage of biologically assisted models is that they require EMG recordings from the worker, which is often not practical in a workplace environment. Hence, many biologically assisted modeling assessments of the spine during work have been performed under laboratory conditions and have attempted to assess specific aspects of the work that may be common to many work conditions. For example, studies have employed EMG-assisted models to assess
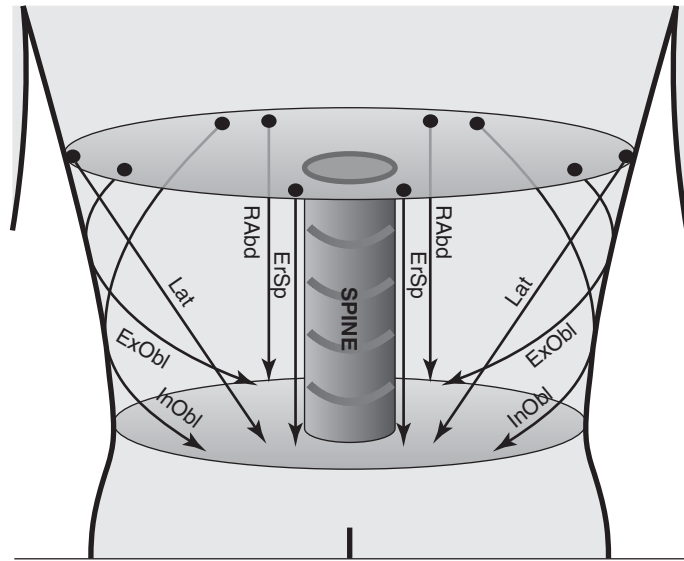
**Figure 6** Biologically (EMG) assisted model used to evaluate spine loading during simulated work activities.
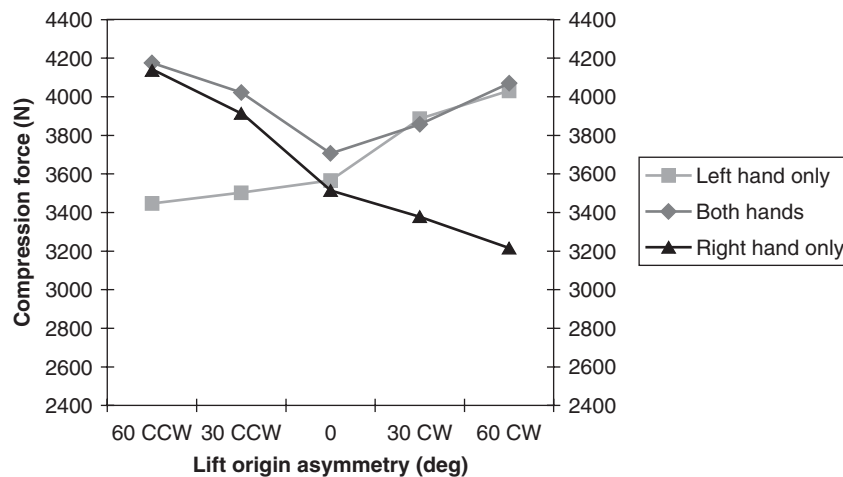


**Figure 7** Mean peak compression force as a function of lift asymmetry (clockwise (CW) vs. counterclockwise (CCW)) and hand(s) used to lift load. Results derived from EMG-assisted model simulation of tasks.

three-dimensional spine loading during materials handling activities (Davis and Marras, 2000; Davis et al., 1998a, 1998b; Marras and Davis, 1998; Marras et al., 2001b; Marras and Granata, 1997a). In addition, numerous studies have yielded information about various dimensions of lifting using biologically assisted models. Figure 7 illustrates the difference in spine compression as subjects lift with one hand verus two hands as a function of lift asymmetry (Marras and Davis, 1998). This assessment indicates that compressive loading of the spine is not simply a matter of load weight lifted. Considerable trade-offs occur as a function of asymmetry and the number of hands involved with the lift. Trade-offs among workplace factors were evaluated in a study

that assessed order-selecting activities in a laboratory setting (Marras et al., 1999d). Results from this study are shown in Table 3. This table highlights the interaction between load weight, location of the lift (region on the pallet), and presence of handles on spine compression (benchmark). The analysis indicates that all three factors influence the loading on the spine. Another study indicated the trade-offs between spine compressive and shear loads as a function of the number of hands used by the worker during the lift, whether both feet were in contact with the ground, lift origin, and height of a bin from which objects were lifted (Ferguson et al., 2002) (Table 4). Other studies have evaluated spine-loading trade-offs associated with team lifting

**Table 3  Percentage of Lifts during Order Selection Tasks in Various Spine Compression Benchmark Zones as Function of Interaction between Load Weight, Location of Lift (Region on Pallet), and Presence of Handles**

| Region on Pallet | Benchmarks ($n$) | Box Weight | | | | | |
| | | 18.2 kg | | 22.7 kg | | 27.3 kg | |
| | | Handles | No Handles | Handles | No Handles | Handles | No Handles |
|---|---|---|---|---|---|---|---|
| | <3400 | 100.0 | 100.0 | 100.0 | 99.2 | 99.2 | 100.0 |
| Front top | 3400–6400 | 0.0 | 0.0 | 0.0 | 0.8 | 0.8 | 0.0 |
| | >6400 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | <3400 | 98.2 | 89.1 | 84.5 | 76.4 | 83.6 | 67.3 |
| Back top | 3400–6400 | 1.8 | 10.9 | 15.5 | 23.6 | 16.4 | 32.7 |
| | >6400 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | <3400 | 98.7 | 91.3 | 94.7 | 82.7 | 92.6 | 76.0 |
| Front middle | 3400–6400 | 1.3 | 8.7 | 5.3 | 17.3 | 7.4 | 23.3 |
| | >6400 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 |
| | <3400 | 88.7 | 82.0 | 80.7 | 75.3 | 76.7 | 64.7 |
| Back middle | 3400–6400 | 11.3 | 18.0 | 19.3 | 24.7 | 23.3 | 34.6 |
| | >6400 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 |
| | <3400 | 45.3 | 30.0 | 29.3 | 14.0 | 16.0 | 3.3 |
| Front bottom | 3400–6400 | 52.0 | 62.0 | 62.7 | 65.3 | 72.0 | 66.0 |
| | >6400 | 2.7 | 8.0 | 8.0 | 20.7 | 12.0 | 30.7 |
| | <3400 | 35.3 | 24.0 | 30.0 | 10.7 | 9.3 | 2.0 |
| Back bottom | 3400–6400 | 60.7 | 67.3 | 56.7 | 65.3 | 71.3 | 62.0 |
| | >6400 | 4.0 | 8.7 | 13.3 | 24.0 | 19.3 | 36.0 |

*Source*: From W. S. Marras et al. (1999), *Ergonomics*, Vol. 42., No. 7, pp. 980–996.
*Note*: Spine loads estimated by an EMG-assisted model.

**Table 4  Spine Forces (Means and Standard Deviations for Lateral Shear, Anterior–Posterior Shear, and Compression) as Function of Number of Hands Used, Number of Feet Supporting Body during Lift, Region of Pallet and Height of Bin When Lifting Items from Industrial Bin**

| Independent Variables | Condition | Lateral Shear Force | Anterior–posterior Shear Force | Compression Force |
|---|---|---|---|---|
| Hand | one hand | 472.2 (350.5)* | 1093.3 (854.7) | 6033.6 (2981.2) |
| | two hand | 233.8 (216.9)* | 1136.9 (964.1) | 5742.3 (1712.3) |
| Feet | one foot | 401.7 (335.1)* | 1109.4 (856.1) | 6138.6 (2957.5)* |
| | two feet | 304.3 (285.1)* | 1120.8 (963.3) | 5637.3 (2717.9)* |
| Region of bin | upper front | 260.2 (271.7)[a] | 616.6 (311.1)[a] | 3765.7 (1452.8)[a] |
| | upper back | 317.0 (290.8)[a] | 738.0 (500.0)[a] | 5418.1 (2364.2)[b] |
| | lower front | 414.4 (335.0)[b] | 1498.3 (1037.8)[b] | 6839.8 (2765.4)[c] |
| | lower back | 420.4 (329.0)[b] | 1607.5 (1058.4)[b] | 7528.2 (2978.4)[d] |
| Bin height | 94 cm | 361.9 (328) | 1089.9 (800.8) | 5795.8 (2660.4) |
| | 61 cm | 344.1 (301) | 1140.3 (1009.1) | 5980.2 (3027.4) |

*Source*: From W.S. Marras, K.G. Davis, B.C. Kirking, and P.K. Bertsche (1999), ''A Comprehensive Analysis of Low-Back Disorder Risk and Spinal Loading during the Transferring and Repositioning of Patients Using Different Techniques,'' *Ergonomics*, Vol. 42, No. 7, pp. 904–926.
*Significant difference at $x = 0.05$
Region has four experimental conditions; superscript letters indicate which regions were significantly different from one another at $x = 0.05$.

(Marras et al., 1999a), patient lifting (Table 5) (Wang, 1999), the assessment of lifting belts (Granata et al., 1997; Jorgensen and Marras, 2000; Marras et al., 2000d; McGill et al., 1990, 1994), and the use of lifting assistance devices (Marras et al., 1996). Efforts have also endeavored to translate these in-depth laboratory studies for use in the field through the use of regression models of workplace characteristics (Fathallah et al., 1999; McGill et al., 1996). In addition, biologically assisted models have been used to assess the role of psychosocial factors, personality, and mental processing on spine loading (Davis et al., 2002; Marras et al., 2000c).

**Table 5 Spine Loads Estimated during Patient Transfer as Function of Number of Lifters and Transfer Technique**

| Transfer task | One-PersonTransfers | Two-PersonTransfers |
|---|---|---|
| **Lateral shear forces (N)** | | |
| Lower to wheelchair without an arm from bed | 1176.8 (891.0)[d] | 754.0 (144.9)[ab] |
| Lower to bed from wheelchair without an arm | 1256.2 (778.8)[d] | 908.6 (589.4)[c] |
| Lower to wheelchair from bed | 1066.8 (490.0)[d] | 639.1 (351.6)[a] |
| Lower to bed from wheelchair | 1017.0 (370.9)[d] | 942.5 (508.3)[c] |
| Lower to commode chair from hospital chair | 1146.8 (587.5)[d] | 833.4 (507.3)[b] |
| Lower to hospital chair from commode chair | 1104.1 (526.6)[d] | 834.1 (425.6)[b] |
| **Anterior-posterior shear forces (N)** | | |
| Lower to wheelchair without an arm from bed | 1031.8 (681.7)[a] | 986.8 (496.8)[a] |
| Lower to bed from wheelchair without an arm | 1089.7 (615.6)[a] | 1032.9 (472.1)[a] |
| Lower to wheelchair from bed | 1180.8 (716.7)[a] | 1020.7 (503.4)[a] |
| Lower to bed from wheelchair | 1108.7 (544.5)[a] | 1049.4 (511.4)[a] |
| Lower to commode chair from hospital chair | 1137.1 (587.5)[a] | 1018.4 (544.9)[a] |
| Lower to hospital chair from commode chair | 1122.0 (536.0)[a] | 982.6 (484.6)[a] |
| **Compression forces (N)** | | |
| Lower to wheelchair without an arm from bed | 5895.4 (1998.1)[d] | 4483.2 (1661.7)[b] |
| Lower to bed from wheelchair without an arm | 6457.2 (1930.6)[e] | 4663.3 (1719.2)[b] |
| Lower to wheelchair from bed | 5424.0 (2133.8)[c] | 4245.2 (1378.7)[a] |
| Lower to bed from wheelchair | 5744.0 (1728.5)[cd] | 4630.7 (1656.2)[b] |
| Lower to commode chair from hospital chair | 6062.3 (1669.7)[d] | 4645.7 (1450.8)[b] |
| Lower to hospital chair from commode chair | 6464.7 (1698.0)[E] | 4630.6 (1621.4)[b] |

Source: From W. S. Marras, K. G. Davis, B. C. Kirking, and P. K. Bertsche (1999), ''A Comprehensive Analysis of Low-Back Disorder Risk and Spinal Loading during the Transferring and Repositioning of Patients Using Different Techniques,'' *Ergonomics*, Vol. 42, No. 7, pp. 904–926.
Note: Superscript letters indicate significant difference at $p = 0.05$.

In an effort to eliminate the need for biological measures (EMG) to assess muscle coactivity and subsequent spine loading, several studies have attempted to use stability as criteria to govern detailed biologically assisted biomechanical models of the torso (Cholewicki and McGill, 1996; Cholewicki et al., 2000b; Cholewicki and VanVliet, 2002; Granata and Marras, 2000; Granata and Orishimo, 2001; Granata and Wilson, 2001; Panjabi, 1992a, 1992b; Solomonow et al., 1999). This is thought to be important because a potential injury pathway for LBDs suggests that the unnatural rotation of a single spine segment may create loads on passive tissue or other muscle tissue that result in irritation or injury (McGill, 2002). However, nearly all of the work performed in this area to date has been directed toward static response of the trunk or sudden loading responses (Cholewicki et al., 2000a; Cholewicki et al., 2000b; Cholewicki and VanVliet, 2002; Granata and Orishimo, 2001; Granata et al., 2001; Granata and Wilson, 2001). Thus, these assessments may have limited value for the assessment of the most common workplace risk factors for LBP.

## 6 ASSESSMENT METHODS AND IDENTIFICATION OF LOW-BACK DISORDER RISK AT WORK

The logic associated with various risk assessment approaches has been described in previous sections. These approaches have been used to develop a rich body of literature that describes spine loading and subsequent risk in response to various work-related factors that are common to workplaces (e.g. one-hand vs. two-hand lifting). These studies can be used as a guide for the proper design of many work situations. However, there is still a need to assess unique work situations that may not have been assessed in these in-depth laboratory studies. High-fidelity spine-loading assessment techniques (e.g., EMG-assisted models) may not be practical for the assessment of some work situations since they require extensive instrumentation and typically require the task to be simulated in a laboratory environment. Therefore, tools with less precision and accuracy may be necessary to estimate risk to the spine due to the work. This section reviews the

methods and tools available for such assessments along with a review of the literature that supports their usage.

## 6.1 3DSSPP

The three-dimensional static strength prediction program (3DSSPP) has been available for quite some time. The logic associated with this approach was described previously. The computer program considers the load–tolerance relationship from both the spine compression and joint strength aspects. Spine compression is estimated with a linked segment–single equivalent muscle model and compared to the NIOSH-established compression tolerance limit of 3400 N.

Strength tolerance is assessed by estimating the joint load imposed by a task on six joints and comparing these loads to a population-based static strength database. This strength relationship has been defined as a lifting strength rating (LSR) and has been used to assess low-back injuries in industrial environments (Chaffin and Park, 1973). The LSR is defined as the weight of the maximum load lifted on the job divided by the lifting strength. The assessment concluded that "the incidence rate of low back pain (was) correlated (monotonically) with higher lifting strength requirements as determined by assessment of both the location and magnitude of the load lifted." This was one of the first studies to emphasize the importance of load moment exposure (importance of load location relative to the body in addition to load weight) when assessing risk. The study also found that exposure to moderate lifting frequencies appeared to be protective, whereas high or low lift rates were associated with jobs linked to greater reports of back injury.

One study used both the LSR and estimates of back-compression forces to observe job risk over a three-year period in five large industrial plants where 2934 material handling tasks were evaluated (Herrin et al., 1986). The findings indicated a positive association between the lifting strength ratio and back pain incidence rates. This study also found that musculoskeletal injuries were twice as likely when spine compression forces exceeded 6800 N. However, this relationship did not hold for low-back-specific incident reports. This study indicated that injury risk prediction was best associated with the most stressful tasks (as opposed to indices that represent risk aggregation).

## 6.2 Job Demand Index

Ayoub developed the concept of a sob severity index (JSI), which is somewhat similar to the LSR (Ayoub et al., 1978). The JSI is defined as the ratio of the job demands relative to the lifting capacities of the worker. Job demands include the variables of object weight (lifted), the frequency of lifting, exposure time, and lifting task origins and destinations. A comprehensive task analysis is necessary to assess the job demands in this context. Worker capacity includes the strength as well as the body size of the worker where strength is determined via psychophysical testing (as discussed earlier). Liles

and associates (1984) performed a prospective study using the JSI and identified a threshold of a job demand relative to worker strength above which the risk of low-back injury increased. These authors suggest that this method could identify the more high risk (costly) jobs.

## 6.3 NIOSH Lifting Guide and Revised Lifting Equation

The NIOSH has developed two lift assessment tools to help those in industry assess the risk associated with materials handling. The objective of these tools was to "prevent or reduce the occurrence of lifting-related low back pain among workers" (Waters et al., 1993). These assessments considered biomechanical, physiological, and psychophysical limits as criteria for assessing task risk.

The original tool was a guide to help define safe lifting limits based upon biomechanical, physiological, and psychophysical tolerance limits ((NIOSH, 1981). This method requires the evaluator to assess workplace characteristics. Based upon these work characteristics, the guide estimates that the magnitude of the load that must be lifted for spine compression reaches 3400 N [the action limit (AL)] and 6400N [the maximum permissible limit (MPL)]. From a biomechanical standpoint, the AL was defined as the spine compression limit at which damage just begins to occur in the spine in a large portion of the population. Based upon this logic, "safe" work tasks should be designed so that the load lifted by the worker is below the calculated AL limit. The AL is calculated through a functional equation that considers four discounting functions multiplied by a constant. The constant (90 lb, or 40 kg) is assumed to be the magnitude of the weight that, when lifted under ideal lifting conditions, would result in a spine compression of 3400 N. The four workplace-based discounting factors are (1) horizontal distance of the load from the spine, (2) the vertical height of the load off the floor, (3) the vertical travel distance of the load, and (4) the frequency of lifting. These factors are governed by functional relationships that reduce the magnitude of the allowable load (constant) proportionally according to their contribution to increases in spine compression. The MPL is determined by simply multiplying the AL by a factor of 3. If the load lifted by the worker exceeds the MPL, it is assumed that more than 50% of the workers would be at risk of damaging the disc. Under these conditions engineering controls would be required. If the load lifted is between the AL and the MPL values, then the task is assumed to place less than half the workers at risk. In this case, either engineering or administrative controls were permitted. If the load lifted is less than the AL, the task is considered safe. This guide was designed primarily for sagittally symmetric lifts that were slow (no appreciable acceleration) and smooth. Only one independent assessment of the guide's effectiveness could be found in the literature (Marras et al., 1999b). When predictions of risk were compared with historical data of industrial back injury reporting,

this evaluation indicated an odds ratio of 3.5 with good specificity and low sensitivity.

The most recent version of a NIOSH lifting guideline is known as the "revised NIOSH lifting equation" (Waters et al., 1993). The revised equation was developed with the intent of also including asymmetric lifting tasks as well as tasks with different types of coupling (handles) features in the assessment. The functional structure of the revised equation is similar in form to the 1981 guide in that it includes a load constant that mediates several work characteristic "multipliers." However, several differences are apparent between these two guides. First, the revised equation yields a recommended weight limit (RWL) (instead of an AL or MPL). If the magnitude of the load lifted by the worker is below the RWL, the load is considered safe. Second, the functional equation's load constant is reduced to 23 kg (51 lb) (from the 40 kg, or 90 lb, in the 1981 guide). Third, the functional relationship between the equation multipliers and the workplace factors is changed. Functionally, these relationships are slightly more liberal for the four factors in order to account for the lower value of the load constant. Fourth, two additional multipliers are included to account for task asymmetry and coupling. Once the RWL is calculated for a given workplace configuration, it is compared (as a denominator) to the load lifted by the worker to yield a lifting index (LI). If the LI is less than unity, the job is considered safe. If the LI is greater than 1, then risk is associated with the task. LI values above 3.0 are thought to place many of the workers at an increased risk of LBP (Waters et al., 1994). The equations that govern both the 1981 and 1993 versions of this guide are described in Chapter 12 of this handbook.

Two effectiveness studies have evaluated the revised equation. The first evaluation compared the ability of the revised equation to identify high- and low-risk jobs based upon a historical LBP reporting in industry (Marras et al., 1999b). This evaluation reported an overall odds ratio of 3.1. In-depth analyses indicated higher sensitivity than the 1981 guide but lower specificity. A second study assessed odds ratios as a function of the LI magnitude (Waters et al., 1999). LIs between 1 and 3 yielded odds ratios ranging from 1.54 to 2.45, suggesting increasing risk with increasing LIs. Conversely, when the LIs were over 3, the odds ratio was lower (odds ratio of 1.63), indicating a nonmonotonic relationship between the LI and risk.

## 6.4   Video-Based Biomechanical Models

Quantitative video-based assessments have been used to better understand the association of LBP risk with workplace factors. A study by Norman and colleagues (1998) employed a quasi-dynamic two-dimensional (2D) biomechanical model to evaluate cumulative biomechanical loading of the spine in 234 automotive assembly workers. The study identified four independent factors for LBP reporting. These factors consisted of integrated load moment (over a work shift), hand forces, peak shear force on the spine, and peak trunk velocity. The analysis found that workers exposed to the upper 25% of loading to all risk factors had about six times

more risk of reporting back pain than those exposed to the lowest 25% of loading.

### 6.4.1   Lumbar Motion Monitor Risk Assessment

The previously reviewed LBP risk assessment tools have not attempted to understand the role of motion in defining risk. We have known since the days of Sir Isaac Newton that force is a function of mass times acceleration. Hence, motion can have a very large influence on spine loading. Yet, most of the available assessment tools represent work tasks as static or quasi-static in their assessments.

The contribution of trunk dynamics combined with traditional workplace biomechanical factors contribution to LBP risk has been assessed by Marras and colleagues (1993a, 1995). These studies evaluated over 400 industrial jobs (along with documented LBD risk history) by observing 114 workplace and worker-related variables. Of the variables documented, load moment (load magnitude times distance of load from spine) exposure was identified as the single most powerful predictor of LBD reporting. The studies also identified 16 trunk kinematic variables that resulted in statistically significant odds ratios associated with risk of LBD reporting in the workplace. None of the individual kinematic variables were as strong a predictor as load moment. However, when load moment was considered in combination with three trunk kinematic variables (describing three dimensions of trunk motion) and an exposure frequency measure, a strong multiple logistic regression model that quantifies the risk of LBD risk (resulting from work design) was identified (odds ratio, O.R. = 10.7). This analysis indicated that risk was multidimensional in nature and that exposure to the combination of the five variables described LBP reporting well. This information was used to develop a functional risk model (Figure 8) that accounts for trade-offs between risk variables. As an example, a job that exposes a worker to a low magnitude of load moment can represent a high-risk situation if the other four variables in the model are of sufficient magnitude. Thus, the model is able to assess the interactions or collective influence of the risk variables. This model has been validated via a prospective workplace intervention study (Marras et al., 2000a). The risk model has been designed to work with a lumbar motion monitor (LMM) (Figure 9) and a computer program to document trunk motion exposure on the job.

When these conclusions are combined with the findings of epidemiological studies exploring the influence of nonneutral postures in the workplace (Punnett et al., 1991), a potential injury pathway is suggested. These studies indicate that as trunk posture becomes more extreme or the trunk motion becomes more rapid (during the performance of work) LBP reporting increases. From a biomechanical perspective, these results suggest that the occupational risk of LBD is associated with mechanical loading of the spine and indicates that when tasks involve greater three-dimensional loading the association with risk becomes much stronger.

Fathallah and associates (1998) evaluated data from 126 jobs to assess the complex trunk motions of groups
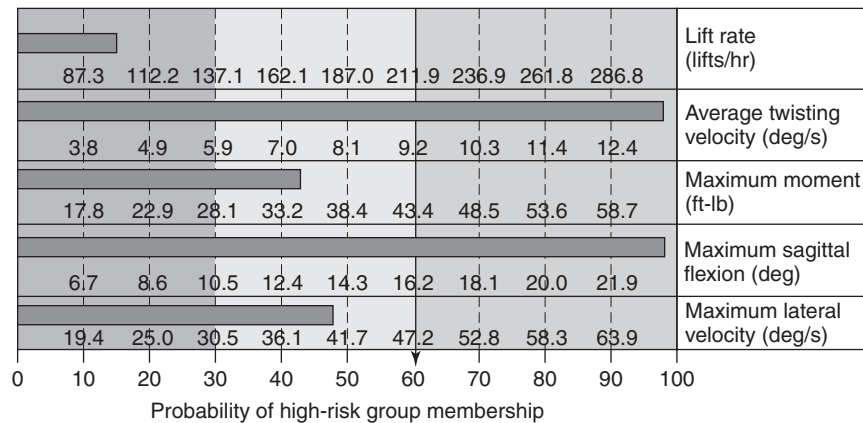
**Figure 8** Lumbar motion monitor (LMM) risk model. The probability of high-risk (LBP) group membership is quantitatively indicated for a particular task for each of five risk factors indicating how much exposure is too much exposure for a particular risk factor. The vertical arrow indicates the overall probability of high-risk group membership due to the combination of risk factors.

associated with varying degrees of LBP reporting. They found that groups with greater LBP reporting rates exhibited complex trunk motion patterns that consisted of high values of combined trunk velocities, especially



**Figure 9** LMM used to track trunk kinematics during occupational activities.

at extreme sagittal flexion. In contrast, the low-risk groups did not exhibit these patterns. This suggests that elevated levels of complex simultaneous velocity patterns along with key workplace factors (load moment and frequency) are correlated with increased LBP risk.

### 6.4.2 Dynamic Moment Exposure in the Workplace

Since earlier studies (Marras and Granata, 1995; Marras et al., 1993a) have shown that exposure to load moment is one of the best indicators of LBP reporting, a recent effort has investigated exposure to dynamic moment exposure and its relationship to decrements in low-back function. An ultrasound-based measurement device (Figure 10) was used to monitor dynamic load moment exposure in distribution center workers over extended periods throughout the workday (Marras et al., 2010a). This effort was able to precisely document the range of dynamic moment exposure associated with different types of work (Marras et al., 2010b). Assessment of these exposures relative to low-back function decrement risk indicated that lateral velocity along with dynamic moment exposure and timing of the peak load exposure allowed the identification of job characteristics leading to low-back dysfunction with excellent sensitivity and specificity (Marras et al., 2010c). This study demonstrates that with proper quantification of realistic (dynamic) task exposures and an appreciation for risk factor interactions, one can indeed identify the characteristics of jobs that lead to LBDs.

### 6.4.3 Workplace Assessment Summary

While there are many superficial reviews of the literature that have not been able to identify relationships between LBP and work factors, none of these studies have assessed quantitative studies of workplace exposure. Only with quantitative measures of the workplace can one assess "how much exposure is too much exposure." The studies described in this chapter are

**Figure 10** Moment monitor used to precisely measure exposure to dynamic load moments on the job.

insightful in that, even though some of these studies have not evaluated spinal loading directly, the exposure measures included can be considered indirect indicators of spinal load. Collectively these studies suggest that as the risk factors increase in magnitude the risk increases monotonically. While load location and strength limits both appear to be indicators of the risk to the spine, other exposure metrics (load location, kinematics, and three-dimensional analyses) are important from a biomechanical standpoint because they influence the ability of the trunk's internal structures to support the external load. Therefore, as these measures change, they can change the nature of the loading on the back's tissues.

These studies indicate that when biomechanically meaningful assessments are collected in the workplace associations between physical factors and risk of LBD reporting are apparent. Several common features of biomechanical risk can be identified from these studies. First, increasingly accurate LBP risk can be identified in the workplace when the specific load magnitude and location relative to the body (load moment) are quantified. Second, studies have demonstrated that increased reporting of LBP can be characterized well when the trunk's three-dimensional kinematic demands

due to work are described. Finally, these assessments have shown that LBP risk is multidimensional. There appears to be a synergy among risk factors that is often associated with increased reporting of LBP. Many studies have also suggested that some of these relationships are nonmonotonic. In summary, these efforts have suggested that the better the exposure characteristics are quantified in terms of biomechanical demand, the better one can assess the association with risk.

## 7 PROCESS OF IMPLEMENTING ERGONOMIC CHANGE

The literature demonstrated that there are interactions between biomechanical loading of the spine and psychosocial factors (Davis et al., 2002; Marras et al., 2000c). Consequently, one must not only address the physical aspects of the workplace but also consider the organizational environment. Ergonomic changes to the work environment must consider biomechanical loading as well as the psychosocial environment. The ergonomic process represents an excellent mechanism for accomplishing these dual goals. Ergonomic processes have been proven effective in introducing and accepting physical change in the workplace (GAO, 1997). Interventions can reduce workers' compensation costs if they are implemented correctly.

The ergonomics process is designed to address occupational health issues in a timely manner and establish an environment that makes the workers accepting of engineering interventions. Ergonomics processes were originally designed to control musculoskeletal disorders in high-risk meat-packing facilities [Occupational Safety and Health Administration (OSHA), 1993]. The objective of this approach is to develop a system or process to identify musculoskeletal problems before they become disabling and correct the features of the work. This is considered a process instead of a program because it is intended to become an ongoing surveillance and correction component of the business operation instead of a one-time effort.

The ergonomics process is intended to encourage communication between management and labor and working as a team to accomplish a common goal of worker health. In order to address the psychosocial environment in the workplace, a key component of the process is worker empowerment. Workers are expected to take an active role in the process and assume control and ownership of work design suggestions and changes. Thus, this process is based upon a participatory approach. Benefits of this approach include positive worker motivation, increased job satisfaction, and greater acceptance of change. The ultimate objective is to create a work environment where the success of the operation is the common goal as opposed to focusing on the interests of any given individual.

There are several common features of a successful ergonomics processes: management leadership and commitment, employee participation, job analysis resulting in injury prevention and control, training, medical management, program evaluation, and effort

documentation. The process is initiated with the creation of an ergonomics committee. Ideally, the committee should be constructed so that it has a balanced representation between management and labor. Committee members should include those involved with the layout of the work as well as those empowered to control scheduling. It is also important to include labor representatives on the committee as well as those employees who have broad experience with many of the jobs in the facility. Finally, it is often wise to include those employees who can communicate well with the majority of the work force as committee members. The ergonomics committee should be at the center of all ergonomic-related activities within the facility.

The ergonomics process can be thought of as a system where the different components of the system interact to produce the desired effect. The interactions among the system components are shown in Figure 11. As shown here, all activities interact in some way with the ergonomics committee to "drive" the process. The ergonomics process begins with management involvement. Ergonomic processes should be top down and therefore must be driven from the top. Management must establish and initiate the process and visibly demonstrate commitment to the process. Furthermore, commitment must be matched with resources made available to the committee. Resources should include not only financial resources so that physical interventions can be implemented but also access to information such as injury records and production schedules.

As shown in Figure 11, there are three fundamental functions of the ergonomics committee. First, the committee must monitor the workplace in order to identify where clusters of work-related LBDs are occurring. Surveillance techniques include monitoring of injury reports for LBP reporting as well as active surveillance of workers for symptoms of LBP. In order to make the effort proactive rather than reactive, it is important to solicit the cooperation of the workforce in this effort. Medical personnel can be recruited to help facilitate this effort by assisting the committee in the interpretation of LBP trends. The second function of the committee involves the control and prevention of work-related LBDs. The variety of techniques discussed earlier can be employed to help isolate and understand the underlying nature of the problems associated with the design of work. In this framework, the question of interest is often "how much exposure to risk factors is too much exposure?" Quantitative methods can be used to help determine which changes are necessary and to estimate their probable impact. As shown in Figure 11, it is important to involve an ergonomic expert in assisting the committee in performing these assessments. The third function of the committee involves the training and education of the workforce. Several levels of training are necessary within the process. All workers should receive awareness training to introduce them to the ergonomics process, familiarize them with LBP risk factors, and explain to them how low-back problems develop. All workers should receive training as to the types of symptoms that need to be reported to the committee in order to optimize prevention efforts. More detailed ergonomics training should also be made available to engineers and supervisors. The goal of the training should be to provide sufficient detail so that management understands the functioning of the process. If this is accomplished, they should not become an impediment to the process success.

Medical management specialists and ergonomics experts serve as resources to the committee within the ergonomics process framework. The goal is not to turn the ergonomics committee into ergonomics experts, but to encourage them to actively work with trained
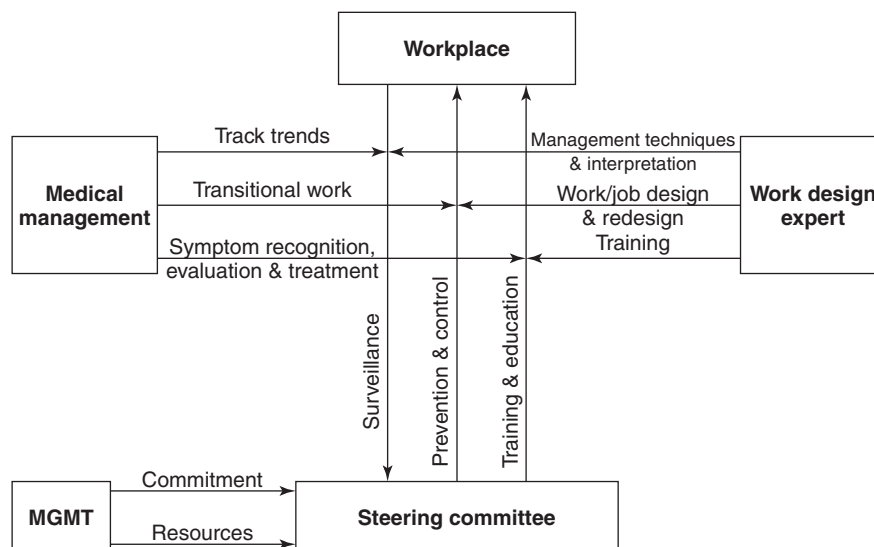


**Figure 11** Interaction of elements within an ergonomics process.

ergonomics experts to accomplish the objectives of the process.

It is essential that the process be evaluated periodically in order to justify its continuation. Metrics such as the achievement of program goals, reductions of LBD reports, hazard reduction, and employee feedback should be considered as indicators of process success. It is also important to recognize that the evaluation provides an opportunity to fine tune the process. All ergonomics processes need to be custom fit to the organization and thus fine tuning is essential. Finally, documentation is a critical part of a successful process. Records documenting the changes made to the workplace as well as their impact on LBDs can serve as justification for process expenditures. These records are also important for process history so that knowledge can be transferred to new team members.

The ergonomics process can have a significant impact on LBD risk, but only if the process is performed consistently and maintained adequately within the organization. Keys to process maintenance consist of strong direction, realistic and achievable goals, establishment of a system to address employee concerns, demonstration of early intervention success, and publicity for the intervention.

## 8  CONCLUSIONS

This chapter has demonstrated that LBDs are common at the work site and strongly linked with occupational tasks when risk factors such as manual materials handling, large moment exposure, bending and twisting, and whole-body vibration are present. The concept of the load–tolerance relationship represents a biomechanically plausible avenue to support the epidemiological findings regarding risk. Advanced biomechanical laboratory-based models have been developed that have been used to quantitatively assess and understand the risk associated with many work situations. There are also a host of workplace assessment tools available to assess risk directly at the work site; however, these tools must be used appropriately, recognizing their limitations. Workplace assessment tools have been shown to be most sensitive for minimizing risk to the low back if they consider multiple risk factors collectively, including load moment exposure and torso kinematic responses to work situations in three-dimensional space. The more precisely these job requirements can be quantified and assessed, the better the association with risk. Finally, the implementation of interventions to minimize risk to the low back must consider psychosocial issues within the workplace in order to foster worker acceptance. An ergonomics process can be useful for these purposes.

## REFERENCES

Adams, M., and Dolan, P. (1995), "Recent Advances in Lumbar Spinal Mechanics and Their Clinical Significance," *Clinical Biochemistry* (*Bristol*, *Avon*), Vol. 10, No. 1, pp. 3–19.

Adams, M., McNally, D., Wagstaff, J., and Goodship, A. E. (1993), "Abdominal Stress Concentrations in Lumbar Intervertebral Discs Followng Damage to the Vertebral Bodies: A Cause of Disc Failure," *European Spine Journal*, Vol. 1, pp. 214–221.

Adams, M. A., and Hutton, W. C. (1981), "The Relevance of Torsion to the Mechanical Derangement of the Lumbar Spine," *Spine*, Vol. 6, No. 3, pp. 241–248.

Adams, M. A., and Hutton, W. C. (1982), "Prolapsed Intervertebral Disc. A Hyperflexion Injury 1981 Volvo Award in Basic Science," *Spine*, Vol. 7, No. 3, pp. 184–191.

Adams, M. A., and Hutton, W. C. (1985), "Gradual Disc Prolapse," *Spine*, Vol. 10, No. 6, pp. 524–531.

Ayoub, M. M., et al. (1978), *Determination and Modeling of Lifting Capacity*, Final Report, NIOSH Grant No. 5R01OH00545-02, Texas Tech University, Lubbock, TX.

Bean, J. C., Chaffin, D. B., and Schultz, A. B. (1988), "Biomechanical Model Calculation of Muscle Contraction Forces: A Double Linear Programming Method," *Journal of Biomechanics*, Vol. 21, No. 1, pp. 59–66.

Begeman, P. C., Visarius, H., Nolte, L. P., and Parasad, P. (1994), "Viscoelastic Shear Responses of the Cadaver and Hybrid III Lumbar Response," paper presented at the 38th Stapp Car Crash Conference, Ft. Lauderdale, FL.

Bigos, S. J., et al. (1991), "A Prospective Study of Work Perceptions and Psychosocial Factors Affecting the Report of Back Injury," *Spine*, Vol. 16, No. 1, pp. 1–6. Published erratum, *Spine*, 1991, Vol. 16, No. 6, p. 688.

Bogduk, N. (1991), "The Lumbar Disc and Low Back Pain," *Neurosurgery Clinics of North America*, Vol. 2, No. 4, pp. 791–806.

Bogduk, N. (1995), "The Anatomical Basis for Spinal Pain Syndromes," *Journal of Manipulative and Physiological Therapy*, Vol. 18, No. 9, pp. 603–605.

Bongers, P. M., de Winter, C. R., Kompier, M. A., and Hildebrandt, V. H. (1993), "Psychosocial Factors At Work and Musculoskeletal Disease," *Scandinavian Journal of Work, Environment & Health*, Vol. 19, No. 5, pp. 297–312.

Brinkmann, P. (1985), "Pathology of the Vertebral Column," *Ergonomics*, Vol. 28, pp. 77–80.

Brinkmann, P., Biggermann, M., and Hilweg, D. (1988), "Fatigue Fracture of Human Lumbar Vertebrae," *Clinical Biochemistry* (*Bristol, Avon*), Vol. 3, Suppl. 1, pp. S1–S23.

Brinkmann, P., Biggermann, M., and Hilweg, D. (1989), "Prediction of the Compressive Strength of Human Lumbar Vertebrae," *Clinical Biochemistry* (*Bristol, Avon*), Vol. 4, Suppl. 2, pp. S1–S27.

Burton, A. K., Tillotson, K. M., Main, C. J., and Hollis, S. (1995), "Psychosocial Predictors of Outcome in Acute and Subchronic Low Back Trouble," *Spine*, Vol. 20, No. 6, pp. 722–728.

Cady, L. D., Bischoff, D. P., O'Connell, E. R., Thomas, P. C., and Allan, J. H. (1979), "Strength and Fitness and Subsequent Back Injuries in Firefighters," *Journal of Occupational Medicine*, Vol. 21, No. 4, pp. 269–272.

Cady, L. D., Jr., Thomas, P. C., and Karwasky, R. J. (1985), "Program for Increasing Health and Physical Fitness of Fire Fighters," *Journal of Occupational Medicine*, Vol. 27, No. 2, pp. 110–114.

Callaghan, J. P., and McGill, S. M. (2001), "Intervertebral Disc Herniation: Studies on a Porcine Model Exposed to Highly Repetitive Flexion/Extension Motion with Compressive Force," *Clinical Biochemistry* (*Bristol*, *Avon*), Vol. 16, No. 1, pp. 28–37.

Carter, D. R. (1985), "Biomechanics of bone," in *Biomechanics of Trauma*, H. M. Nahum and J. Melvin, Eds., Appleton Century Crofts, Norwalk, CT.

Cavanaugh, J. M. (1995), "Neural Mechanisms of Lumbar Pain," *Spine*, 20, No. 16, pp. 1804–1809.

Cavanaugh, J. M., Kallakuri, S., and Ozaktay, A. C. (1995), "Innervation of the Rabbit Lumbar Intervertebral Disc and Posterior Longitudinal Ligament," *Spine*, Vol. 20, No. 19, pp. 2080–2085.

Cavanaugh, J. M., Ozaktay, A. C., Yamashita, T., Avramov, A., Getchell, T. V., and King, A. I. (1997), "Mechanisms of Low Back Pain: A Neurophysiologic and Neuroanatomic Study," *Clinical Orthopaedics*, Vol. 335, pp. 166–180.

Chaffin, D. B., and Baker, W. H. (1970), "A Biomechanical Model for Analysis of Symmetric Sagittal Plane Lifting," *AIIE Transactions*, Vol. 2, No. 1, pp. 16–27.

Chaffin, D. B., and Park, K. S. (1973), "A Longitudinal Study of Low-Back Pain as Associated with Occupational Weight Lifting Factors," *American Industrial Hygiene Association Journal*, Vol. 34, No. 12, pp. 513–525.

Chaffin, D. B., Herrin, G. D., Keyserling, W. M., and Garg, A. (1977), "A Method for Evaluating the Biomechanical Stresses Resulting from Manual Materials Handling Jobs," *American Industrial Hygiene Association Journal*, Vol. 38, No. 12, pp. 662–675.

Chaffin, D. B., Andersson, G. B. J., and Martin, B. J. (1999), *Occupational Biomechanics*, 3rd ed., Wiley, New York.

Chen, C., Cavanaugh, J. M., Ozaktay, A. C., Kallakuri, S., and King, A. I. (1997), "Effects of Phospholipase A2 on Lumbar Nerve Root Structure and Function," *Spine*, Vol. 22, No. 10, pp. 1057–1064.

Cholewicki, J., and McGill, S. (1996), "Mechanical Stability of the In Vivo Lumbar Spine: Implications of Injury and Chronic Low Back Pain," *Clinical Biochemistry* (*Bristol*, *Avon*), Vol. 11, No. 1, pp. 1–15.

Cholewicki, J., and VanVliet, I. J. (2002), "Relative Contribution of Trunk Muscles to the Stability of the Lumbar Spine during Isometric Exertions," *Clinical Biochemistry* (*Bristol*, *Avon*), Vol. 17, No. 2, pp. 99–105.

Cholewicki, J., Polzhofer, G. K., and Radebold, A. (2000a), "Postural Control of Trunk during Unstable Sitting," *Journal of Biomechanics*, Vol. 33, No. 12, pp. 1733–1737.

Cholewicki, J., Simons, A. P., and Radebold, A. (2000b), "Effects of External Trunk Loads on Lumbar Spine Stability," *Journal of Biomechanics*, Vol. 33, No. 11, pp. 1377–1385.

Ciriello, V. M., Snook, S. H., Blick, A. C., and Wilkinson, P. L. (1990), "The Effects of Task Duration on Psychophysically-Determined Maximum Acceptable Weights and Forces," *Ergonomics*, Vol. 33, No. 2, pp. 187–200.

Cripton, P., Berlemen, U., Visarino, H., Begman, P. C., Nolte, L. P., and Prasad, P. (1995, May 4–5), "Response of the Lumbar Spine Due to Shear Loading," paper presented at the Injury Prevention through Biomechanics, Wayne State Unvisersity, Detroit, MI.

Davis, K., Marras, W. S., Heaney, C. A., Waters, T. R., and Gupta, P. (2002), "The Impact of Mental Processing and Pacing on Spine Loading," *Spine*, Vol. 27, No. 23, pp. 2645–2653.

Davis, K. G., and Heaney, C. A. (2000), "The Relationship between Psychosocial Work Characteristics and Low Back Pain: Underlying Methodological Issues," *Clinical Biochemistry* (*Bristol*, *Avon*), Vol. 15, No. 6, pp. 389–406.

Davis, K. G., and Marras, W. S. (2000), "Assessment of the Relationship between Box Weight and Trunk Kinematics: Does a Reduction in Box Weight Necessarily Correspond to a Decrease in Spinal Loading? [In Process Citation]," *Human Factors*, Vol. 42, No. 2, pp. 195–208.

Davis, K. G., Marras, W. S., and Waters, T. R. (1998a), "The Evaluation of Spinal Loads during Lowering and Lifting," *Clinical Biochemistry* (*Bristol*, *Avon*), Vol. 13, No. 3, pp. 141–152.

Davis, K. G., Marras, W. S., and Waters, T. R. (1998b), "Reduction of Spinal Loading through the Use of Handles," *Ergonomics*, Vol. 41, No. 8, pp. 1155–1168.

Dinarello, C. A. (2000), "Proinflammatory Cytokines," *Chest*, Vol. 118, No. 2, pp. 503–508.

Dwyer, A., Aprill, C., and Bogduk, N. (1990), "Cervical Zygapophyseal Joint Pain Patterns. I: A Study in Normal Volunteers," *Spine*, Vol. 15, No. 6, pp. 453–457.

Farfan, H. F., Cossette, J. W., Robertson, G. H., Wells, R. V., and Kraus, H. (1970), "The Effects of Torsion on the Lumbar Intervertebral Joints: The Role of Torsion in the Production of Disc Degeneration," *Journal of Bone and Joint Surgery* (*American*), Vol. 52, No. 3, pp. 468–497.

Fathallah, F. A., Marras, W. S., and Wright, P. L. (1995), "Diurnal Variation in Trunk Kinematics during a Typical Work Shift," *Journal of Spinal Disorders*, Vol. 8, No. 1, pp. 20–25.

Fathallah, F. A., Marras, W. S., and Parnianpour, M. (1998), "The Role of Complex, Simultaneous Trunk Motions in the Risk of Occupation-Related Low Back Disorders," *Spine*, Vol. 23, No. 9, pp. 1035–1042.

Fathallah, F. A., Marras, W. S., and Parnianpour, M. (1999), "Regression Models for Predicting Peak and Continuous Three-Dimensional Spinal Loads during Symmetric and Asymmetric Lifting Tasks," *Human Factors*, Vol. 41, No. 3, 373–388.

Ferguson, S. A., Gaudes-MacLaren, L. L., Marras, W. S., Waters, T. R., and Davis, K. G. (2002), "Spinal Loading When Lifting from Industrial Storage Bins," *Ergonomics*, Vol. 45, No. 6, pp. 399–414.

Frank, J. W., et al. (1996), "Disability Resulting from Occupational Low Back Pain. Part I: What Do We Know about Primary Prevention? A Review of the Scientific Evidence on Prevention before Disability Begins," *Spine*, Vol. 21, No. 24, pp. 2908–2917.

Fyhrie, D. P., and Schaffler, M. B. (1994), "Failure Mechanisms in Human Vertebral Cancellous Bone," *Bone*, Vol. 15, No. 1, 105–109.

Gedalia, U., Solomonow, M., Zhou, B. H., Baratta, R. V., Lu, Y., and Harris, M. (1999), "Biomechanics of Increased Exposure to Lumbar Injury Caused by Cyclic Loading. Part 2. Recovery of Reflexive Muscular Stability with Rest," *Spine*, Vol. 24, No. 23, pp. 2461–2467.

Gordon, S. J., Yang, K. H., Mayer, P. J., Mace, A. H., Jr., Kish, V. L., and Radin, E. L. (1991), "Mechanism of Disc Rupture. A Preliminary Report," *Spine*, Vol. 16, No. 4, pp. 450–456.

Government Accountability Office (GAO) (1997), *Worker Protection: Private Sector Ergonomics Programs Yield Positive Results*, No. GAO/HHS 97-163, GAO, Washington, DC.

Granata, K. P., and Marras, W. S. (1993), "An EMG-Assisted Model of Loads on the Lumbar Spine during Asymmetric Trunk Extensions," *Journal of Biomechanics*, Vol. 26, No. 12, pp. 1429–1438.

Granata, K. P., and Marras, W. S. (1995a), "An EMG-Assisted Model of Trunk Loading During Free-Dynamic Lifting," *Journal of Biomechanics*, Vol. 28, No. 11, pp. 1309–1317.

Granata, K. P., and Marras, W. S. (1995b), "The Influence of Trunk Muscle Coactivity on Dynamic Spinal Loads," *Spine*, Vol. 20, No. 8, pp. 913–919.

Granata, K. P., and Marras, W. S. (1999), "Relation between Spinal Load Factors and the High-Risk Probability of Occupational Low-Back Disorder," *Ergonomics*, Vol. 42, No. 9, pp. 1187–1199.

Granata, K. P., and Marras, W. S. (2000), "Cost-Benefit of Muscle Cocontraction in Protecting Against Spinal Instability," *Spine*, Vol. 25, No. 11, pp. 1398–1404.

Granata, K. P., and Orishimo, K. F. (2001), "Response of Trunk Muscle Coactivation to Changes in Spinal Stability," *Journal of Biomechanics*, Vol. 34, No. 9, pp. 1117–1123.

Granata, K. P., and Wilson, S. E. (2001), "Trunk Posture and Spinal Stability," *Clinical Biochemistry* (*Bristol*, *Avon*), Vol. 16, No. 8, pp. 650–659.

Granata, K. P., Marras, W. S., and Davis, K. G. (1997), "Biomechanical Assessment of Lifting Dynamics, Muscle Activity and Spinal Loads While Using Three Different Styles of Lifting Belt," *Clinical Biochemistry* (*Bristol*, *Avon*), Vol. 12, No. 2, pp. 107–115.

Granata, K. P., Marras, W. S., and Davis, K. G. (1999), "Variation in Spinal Load and Trunk Dynamics during Repeated Lifting Exertions," *Clinical Biochemistry* (*Bristol*, *Avon*), Vol. 14, No. 6, pp. 367–375.

Granata, K. P., Orishimo, K. F., and Sanford, A. H. (2001), "Trunk Muscle Coactivation in Preparation for Sudden Load," *Journal of Electromyography and Kinesiology*, Vol. 11, No. 4, pp. 247–254.

Gunning, J. L., Callaghan, J. P., and McGill, S. M. (2001), "Spinal Posture and Prior Loading History Modulate Compressive Strength and Type of Failure in the Spine: A Biomechanical Study Using a Porcine Cervical Spine Model," *Clinical Biochemistry* (*Bristol*, *Avon*), Vol. 16, No. 6, pp. 471–480.

Guo, H. R., et al. (1995), "Back Pain among Workers in the United States: National Estimates and Workers at High Risk," *American Journal of Industrial Medicine*, Vol. 28, No. 5, pp. 591–602.

Guo, H. R., Tanaka, S., Halperin, W. E., and Cameron, L. L. (1999), "Back Pain Prevalence in US Industry and Estimates of Lost Workdays," *American Journal of Public Health*, Vol. 89, No. 7, pp. 1029–1035.

Herrin, G. D., Jaraiedi, M., and Anderson, C. K. (1986), "Prediction of Overexertion Injuries Using Biomechanical and Psychophysical Models," *American Industrial Hygiene Association Journal*, Vol. 47, No. 6, pp. 322–330.

Heylings, D. J. (1978), "Supraspinous and Interspinous Ligaments of the Human Lumbar Spine," *Jornal of Anatomy*, Vol. 125, No. 1, pp. 127–131.

Hignett, S. (2008). "Work-Related Back Pain in Nurses," *Journal of Advanced Nursing*, Vol. 23, No. 6, pp. 1238–1246.

Holmes, A. D., Hukins, D. W., and Freemont, A. J. (1993), "End-Plate Displacement during Compression of Lumbar Vertebra-Disc-Vertebra Segments and the Mechanism of Failure," *Spine*, Vol. 18, No. 1, pp. 128–135.

Hoogendoorn, W. E., van Poppel, M. N., Bongers, P. M., Koes, B. W., and Bouter, L. M. (1999), "Physical Load during Work and Leisure Time as Risk Factors for Back Pain [see comments]," *Scandinavian Journal of Work, Environment & Health*, Vol. 25, No. 5, pp. 387–403.

Hoogendoorn, W. E., van Poppel, M. N., Bongers, P. M., Koes, B. W., and Bouter, L. M. (2000), "Systematic Review of Psychosocial Factors at Work and Private Life as Risk Factors for Back Pain," *Spine*, Vol. 25, No. 16, pp. 2114–2125.

Hughes, R. E., and Chaffin, D. B. (1995), "The Effect of Strict Muscle Stress Limits on Abdominal Muscle Force Predictions for Combined Torsion and Extension Loadings," *Journal of Biomechanics*, Vol. 28, No. 5, pp. 527–533.

Jacobs, J., Ed. (2008), *The Burden of Musculoskeletal Diseases in the United States*, American Academy of Orthopaedic Surgeons, Rosemont, IL.

Jager, M. (1987), *Biomechanisches Modell des Menschen zu Analyse und Beureilung der Belastung der Wirbelsaule bie der Handhabung von Lasten*, Fortschritt-Bericht 17/33, VDI Verlag, Dusseldorf, Germany.

Jager, M., Luttmann, A., and Laurig, W. (1991), "Lumbar Load during One-Hand Bricklaying," *International Journal of Industrial Ergonomics*, Vol. 8, pp. 261–277.

Jorgensen, M. J., & Marras, W. S. (2000), "The Effect of Lumbar Back Support Tension on Trunk Muscle Activity," *Clinical Biochemistry* (*Bristol*, *Avon*), Vol. 15, No. 4, pp. 292–294.

Kallakuri, S., Cavanaugh, J. M., and Blagoev, D. C. (1998), "An Immunohistochemical Study of Innervation of Lumbar Spinal Dura and Longitudinal Ligaments," *Spine*, Vol. 23, No. 4, pp. 403–411.

Karasek, R., Brisson, C., Kawakami, N., Houtman, I., Bongers, P., and Amick, B. (1998), "The Job Content Questionnaire (JCQ): An Instrument for Internationally Comparative Assessments of Psychosocial Job Characteristics," *Journal of Occupational Medicine and Psychology*, Vol. 3, No. 4, pp. 322–355.

Kirkaldy-Willis, W. H. (1998), "The Three Phases of the Spectrum of Degenerative Diesars," in *Managing Low Back Pain*, 2nd ed., W. H. Kirkaldy-Willis, Ed., Churchill-Livingston, New York.

Kornberg, M. (1988), "MRI Diagnosis of Traumatic Schmorl's Node. A Case Report," *Spine*, Vol. 13, No. 8, pp. 934–935.

Krause, N., Ragland, D. R., Fisher, J. M., and Syme, S. L. (1998), "Psychosocial Job Factors, Physical Workload, and Incidence of Work- Related Spinal Injury: A 5-Year Prospective Study of Urban Transit Operators," *Spine*, Vol. 23, No. 23, pp. 2507–2516.

Krypton, P., Berlemen, U., Visarino, H., Begman, P. C., Nolte, L. P., and Prasad, P. (1995, May 4–5), "Response of the Lumbar Spine Due to Shear Loading," paper presented at the Injury Prevention through Biomechanics, Wayne State Unvisersity, Detroit, MI.

Kumar, S. (1990), "Cumulative Load as a Risk Factor for Back Pain," *Spine*, Vol. 15, No. 12, pp. 1311–1316.

Liles, D. H., Deivanayagam, S., Ayoub, M. M., and Mahajan, P. (1984), "A Job Severity Index for the Evaluation and Control of Lifting Injury," *Human Factors*, Vol. 26, No. 6, pp. 683–693.

Marras, W. S., and Davis, K. G. (1998), "Spine Loading during Asymmetric Lifting Using One versus Two Hands," *Ergonomics*, Vol. 41, No. 6, pp. 817–834.

Marras, W. S., and Granata, K. P. (1995), "A Biomechanical Assessment and Model of Axial Twisting in the Thoracolumbar Spine," *Spine*, Vol. 20, No. 13, pp. 1440–1451.

Marras, W. S., and Granata, K. P. (1997a), "Changes in Trunk Dynamics and Spine Loading during Repeated Trunk Exertions," *Spine*, Vol. 22, No. 21, pp. 2564–2570.

Marras, W. S., and Granata, K. P. (1997b), "The Development of an EMG-Assisted Model to Assess Spine Loading during Whole-Body Free-Dynamic Lifting," *Journal of Electromyography and Kinesiology*, Vol. 7, No. 4, pp. 259–268.

Marras, W. S., and Granata, K. P. (1997c), "Spine Loading during Trunk Lateral Bending Motions," *Journal of Biomechanics*, Vol. 30, No. 7, pp. 697–703.

Marras, W. S., King, A. I., and Joynt, R. L. (1984), "Measurement of Loads on the Lumbar Spine under Isometric and Isokinetic Conditions," *Spine*, Vol. 9, No. 2, pp. 176–187.

Marras, W. S., and Sommerich, C. M. (1991a), "A Three-Dimensional Motion Model of Loads on the Lumbar Spine: I. Model Structure," *Human Factors*, Vol. 33, No. 2, pp. 123–137.

Marras, W. S., and Sommerich, C. M. (1991b), "A Three-Dimensional Motion Model Of Loads on the Lumbar Spine: II. Model Validation," *Human Factors*, Vol. 33, No. 2, pp. 139–149.

Marras, W. S., et al. (1993a), "The Role of Dynamic Three-Dimensional Trunk Motion in Occupationally- Related Low Back Disorders. The Effects of Workplace Factors, Trunk Position, and Trunk Motion Characteristics on Risk of Injury," *Spine*, Vol. 18, No. 5, pp. 617–628.

Marras, W. S., et al. (1993b), "The Role of Dynamic Three-Dimensional Trunk Motion in Occupationally-Related Low Back Disorders. The Effects of Workplace Factors, Trunk Position, and Trunk Motion Characteristics on Risk of Injury," *Spine*, Vol. 18, No. 5, pp. 617–628.

Marras, W. S., et al. (1995), "Biomechanical Risk Factors for Occupationally Related Low Back Disorders," *Ergonomics*, Vol. 38, No. 2, pp. 377–410.

Marras, W. S., Granata, K. P., Davis, K. G., Allread, W. G., and Jorgensen, M. J. (1996), *Low Back Disorder Risk and Spine Loading during the Use of Lifting Assistance Devicess (LADs)*, final report for Zimmerman Inc., The Ohio State University, Columbus, OH.

Marras, W. S., Davis, K. G., and Granata, K. P. (1998), "Trunk Muscle Activities during Asymmetric Twisting Motions," *Journal of Electromyography and Kinesiology*, Vol. 8, No. 4, pp. 247–256.

Marras, W. S., Davis, K. G., Kirking, B. C., and Granata, K. P. (1999a), "Spine Loading and Trunk Kinematics during Team Lifting," *Ergonomics*, Vol. 42, No. 10, pp. 1258–1273.

Marras, W. S., Fine, L. J., Ferguson, S. A., and Waters, T. R. (1999b), "The Effectiveness of Commonly Used Lifting Assessment Methods to Identify Industrial Jobs Associated with Elevated Risk of Low-Back Disorders," *Ergonomics*, Vol. 42, No. 1, pp. 229–245.

Marras, W. S., Granata, K. P., and Davis, K. G. (1999c), "Variability in Spine Loading Model Performance," *Clinical Biochemistry (Bristol, Avon)*, Vol. 14, No. 8, pp. 505–514.

Marras, W. S., Granata, K. P., Davis, K. G., Allread, W. G., and Jorgensen, M. J. (1999d), "Effects of Box Features on Spine Loading during Warehouse Order Selecting," *Ergonomics*, Vol. 42, No. 7, pp. 980–996.

Marras, W. S., Allread, W. G., Burr, D. L., and Fathallah, F. A. (2000a), "Prospective Validation of a Low-Back Disorder Risk Model and Assessment of Ergonomic Interventions Associated with Manual Materials Handling Tasks," *Ergonomics*, Vol. 43, No. 11, pp. 1866–1886.

Marras, W. S., Allread, W. G., Burr, D. L., and Fathallah, F. A. (2000b), "Prospective Validation of a Low-Back Disorder Risk Model and Assessment of Ergonomic Interventions Associated with Manual Materials Handling Tasks [In Process Citation]," *Ergonomics*, Vol. 43, No. 11, pp. 1866–1886.

Marras, W. S., Davis, K. G., Heaney, C. A., Maronitis, A. B., and Allread, W. G. (2000c), "The Influence of Psychosocial Stress, Gender, and Personality on Mechanical Loading of the Lumbar Spine," *Spine*, Vol. 25, No. 23, pp. 3045–3054.

Marras, W. S., Jorgensen, M. J., and Davis, K. G. (2000d), "Effect of Foot Movement and an Elastic Lumbar Back Support on Spinal Loading during Free-Dynamic Symmetric and Asymmetric Lifting Exertions," *Ergonomics*, Vol. 43, No. 5, pp. 653–668.

Marras, W. S., Davis, K. G., Ferguson, S. A., Lucas, B. R., and Gupta, P. (2001a), "Spine Loading Characteristics of Patients with Low Back Pain Compared with Asymptomatic Individuals," *Spine*, Vol. 26, No. 23, pp. 2566–2574.

Marras, W. S., Davis, K. G., and Splittstoesser, R. E. (2001b), "Spine Loading during Whole Body Free Dynamic Lifting," final project report, Grant No. R01 OH03289, The Ohio State University, Columbus, OH.

Marras, W. S., Davis, K. G., and Jorgensen, M. J. (2002), "Spine Loading as a Function of Gender," *Spine*, Vol. 27, No. 22, pp. 2514–2520.

Marras, W. S., Lavender, S. A., Ferguson, S. A., Splittstoesser, R. E., Yang, G., and Schabo, P. (2010a), "Instrumentation for Measuring Dynamic Spinal Load Moment Exposures in the Workplace," *Journal of Electromyography and Kinesiology*, Vol. 20, No. 1, pp. 1–9.

Marras, W. S., Lavender, S. A., Ferguson, S. A., Splittstoesser, R. E., and Yang, G. (2010b), "Quantitative Biomechanical Workplace Exposure Measures: Distribution Centers," *Journal of Electromyography and Kinesiology*, Vol. 20, No. 5, pp. 813–822.

Marras, W. S., Lavender, S. A., Ferguson, S. A., Splittstoesser, R. E., and Yang, G. (2010c), "Quantitative Dynamic Measures of Physical Exposure Predict Low Back Functional Impairment," *Spine (Phila Pa 1976)*, Vol. 35, No. 8, pp. 914–923.

McGill, S. (2002), *Low Back Disorder*, Human Kinetics, Champaign, IL.

McGill, S., Seguin, J., and Bennett, G. (1994), "Passive Stiffness of the Lumbar Torso in Flexion, Extension, Lateral Bending, and Axial Rotation. Effect of Belt Wearing and Breath Holding," *Spine*, Vol. 19, No. 6, pp. 696–704.

McGill, S. M. (1991), "Electromyographic Activity of the Abdominal and Low Back Musculature during the Generation of Isometric and Dynamic Axial Trunk Torque: Implications for Lumbar Mechanics," *Journal of Orthopaedic Research*, Vol. 9, No. 1, pp. 91–103.

McGill, S. M. (1992a), "The Influence of Lordosis on Axial Trunk Torque and Trunk Muscle Myoelectric Activity," *Spine*, Vol. 17, No. 10, pp. 1187–1193.

McGill, S. M. (1992b), "A Myoelectrically Based Dynamic Three-Dimensional Model to Predict Loads on Lumbar Spine Tissues during Lateral Bending," *Journal of Biomechanics*, Vol. 25, No. 4, pp. 395–414.

McGill, S. M. (1997), "The Biomechanics of Low Back Injury: Implications on Current Practice in Industry and the Clinic," *Journal of Biomechanics*, Vol. 30, No. 5, pp. 465–475.

McGill, S. M., and Norman, R. W. (1987), "Effects of an Anatomically Detailed Erector Spinae Model On L4/L5 Disc Compression and Shear," *Journal of Biomechanics*, Vol. 20, No. 6, pp. 591–600.

McGill, S. M., Norman, R. W., and Sharratt, M. T. (1990), "The Effect of an Abdominal Belt on Trunk Muscle Activity and Intra-Abdominal Pressure during Squat Lifts," *Ergonomics*, Vol. 33, No. 2, pp. 147–160.

McGill, S. M., Norman, R. W., and Cholewicki, J. (1996), "A Simple Polynomial That Predicts Low-Back Compression during Complex 3-D Tasks," *Ergonomics*, Vol. 39, No. 9, pp. 1107–1118.

Mirka, G. A., and Marras, W. S. (1993), "A Stochastic Model of Trunk Muscle Coactivation during Trunk Bending," *Spine*, Vol. 18, No. 11, pp. 1396–1409.

Moore, R. J. (2000), The Vertebral End-Plate: What Do We Know? *European Spine Journal*, Vol. 9, pp. 92–96.

National Institute for Occupational Safety and Health (NIOSH) (1981), *Work Practices Guide for Manual Lifting*, Publication No. 81-122, Department of Health and Human Services (DHHS), NIOSH, Cincinnati, OH.

National Institute for Occupational Safety and Health (NIOSH) (1997), *Musculoskeletal Disorders and Workplace Factors*, Publication No. 97-141, Department of Health and Human Services (DHHS), NIOSH, Cincinnati, OH.

National Institute for Occupational Safety and Health (NIOSH) (2000), *Worker Health Chartbook, 2000*, NIOSH, Cincinnati, OH.

National Research Council (NRC) (2001), *Musculoskeletal Disorders and the Workplace: Low Back and Upper Extremity*, National Academy Press, Washington, DC.

Norman, R., Wells, R., Neumann, P., Frank, J., Shannon, H., and Kerr, M. (1998), "A Comparison of Peak vs Cumulative Physical Work Exposure Risk Factors for the Reporting of Low Back Pain in the Automotive Industry", *Clinical Biochemistry* (*Bristol, Avon*), Vol. 13, No. 8, pp. 561–573.

Noyes, F. R., De Lucas, J. L., and Torvik, P. J. (1994), Biomechanics of Ligament Failure: An Analysis Of Strain-Rate Sensitivty And Mechanisms Of Failure In Primates. *Journal of Bone Joint Surgery*, Vol. 56A, pp. 236–253.

Occupational Safety and Health Administration (OSHA) (1993), *Ergonomics Program MAnagement Guidelines for Meatpacking Plants*, Publication No. OSHA 3123, U.S. Department of Labor, Washington DC.

Ozaktay, A. C., Cavanaugh, J. M., Blagoev, D. C., and King, A. I. (1995), "Phospholipase A2-Induced Electrophysiologic and Histologic Changes in Rabbit Dorsal Lumbar Spine Tissues," *Spine*, Vol. 20, No. 24, pp. 2659–2668.

Ozaktay, A. C., Kallakuri, S., and Cavanaugh, J. M. (1998), "Phospholipase A2 Sensitivity of the Dorsal Root and Dorsal Root Ganglion," *Spine*, Vol. 23, No. 12, pp. 1297–1306.

Panjabi, M. M. (1992a), "The Stabilizing System of the Spine. Part I. Function, Dysfunction, Adaptation, and Enhancement," *Journal of Spinal Disorders*, Vol. 5, No. 4, pp. 383–389; discussion 397.

Panjabi, M. M. (1992b), "The Stabilizing System of the Spine. Part II. Neutral Zone and Instability Hypothesis," *Journal of Spinal Disorders*, Vol. 5, No. 4, pp. 390–396; discussion 397.

Pope, M. H., Andersson, G. B., Broman, H., Svensson, M., and Zetterberg, C. (1986), "Electromyographic Studies of the Lumbar Trunk Musculature during the Development of Axial Torques," *Journal of Orthopaedic Research*, Vol. 4, No. 3, pp. 288–297.

Pope, M. H., Svensson, M., Andersson, G. B., Broman, H., and Zetterberg, C. (1987), "The Role of Prerotation of the Trunk in Axial Twisting Efforts," *Spine*, Vol. 12, No. 10, pp. 1041–1045.

Porter, R. W., Adams, M. A., and Hutton, W. C. (1989), "Physical Activity and the Strength of the Lumbar Spine," *Spine*, Vol. 14, No. 2, pp. 201–203.

Potvin, J. R., McGill, S. M., and Norman, R. W. (1991), "Trunk Muscle and Lumbar Ligament Contributions to Dynamic Lifts with Varying Degrees of Trunk Flexion [see comments]," *Spine*, Vol. 16, No. 9, pp. 1099–1107.

Punnett, L., Fine, L. J., Keyserling, W. M., Herrin, G. D., and Chaffin, D. B. (1991), "Back Disorders and Nonneutral Trunk Postures of Automobile Assembly Workers," *Scandinavian Journal of Work, Environment & Health*, Vol. 17, No. 5, pp. 337–346.

Schultz, A., Andersson, G. B., Ortengren, R., Bjork, R., and Nordin, M. (1982), "Analysis and Quantitative Myoelectric Measurements of Loads on the Lumbar Spine When Holding Weights in Standing Postures," *Spine*, Vol. 7, No. 4, pp. 390–397.

Schultz, A. B., and Andersson, G. B. (1981), "Analysis of Loads on the Lumbar Spine," *Spine*, Vol. 6, No. 1, pp. 76–82.

Seidler, A., et al. (2001), "The Role of Cumulative Physical Work Load in Lumbar Spine Disease: Risk Factors for Lumbar Osteochondrosis and Spondylosis Associated with Chronic Complaints," *Occupational and Environmental Medicine*, Vol. 58, No. 11, pp. 735–746.

Siddall, P. J., and Cousins, M. J. (1997a), "Neurobiology of Pain," *International of Anesthesiology Clinics*, Vol. 35, No. 2, pp. 1–26.

Siddall, P. J., and Cousins, M. J. (1997b), "Spinal Pain Mechanisms," *Spine*, Vol. 22, No. 1, pp. 98–104.

Snook, S. H. (1978), "The Design of Manual Handling Tasks," *Ergonomics*, Vol. 21, No. 12, pp. 963–985.

Snook, S. H. (1985a), "Psychophysical Acceptability as a Constraint in Manual Working Capacity," *Ergonomics*, Vol. 28, No. 1, pp. 331–335.

Snook, S. H. (1985b), "Psychophysical Considerations in Permissible Loads," *Ergonomics*, Vol. 28, No. 1, pp. 327–330.

Snook, S. H. (1987), "Approaches to Preplacement Testing and Selection of Workers," *Ergonomics*, Vol. 30, No. 2, pp. 241–247.

Snook, S. H., and Ciriello, V. M. (1991), "The Design of Manual Handling Tasks: Revised Tables of Maximum Acceptable Weights and Forces," *Ergonomics*, Vol. 34, No. 9, pp. 1197–1213.

Snook, S. H., Webster, B. S., McGorry, R. W., Fogleman, M. T., and McCann, K. B. (1998), "The Reduction of Chronic Nonspecific Low Back Pain through the Control of Early Morning Lumbar Flexion. A Randomized Controlled Trial," *Spine*, Vol. 23, No. 23, pp. 2601–2607.

Solomonow, M., Zhou, B. H., Harris, M., Lu, Y., and Baratta, R. V. (1998), "The Ligamento-Muscular Stabilizing System of the Spine," *Spine*, Vol. 23, No. 23, pp. 2552–2562.

Solomonow, M., Zhou, B. H., Baratta, R. V., Lu, Y., and Harris, M. (1999), "Biomechanics of Increased Exposure to Lumbar Injury Caused by Cyclic Loading: Part 1. Loss of Reflexive Muscular Stabilization," *Spine*, Vol. 24, No. 23, pp. 2426–2434.

Solomonow, M., He Zhou, B., Baratta, R. V., Lu, Y., Zhu, M., and Harris, M. (2000), "Biexponential Recovery Model of Lumbar Viscoelastic Laxity and Reflexive Muscular Activity after Prolonged Cyclic Loading," *Clinical Biochemistry* (*Bristol*, *Avon*), Vol. 15, No. 3, pp. 167–175.

Solomonow, M., Zhou, B., Baratta, R. V., Zhu, M., and Lu, Y. (2002), "Neuromuscular Disorders Associated with Static Lumbar Flexion: A Feline Model," *Journal of Electromyography and Kinesiology*, Vol. 12, No. 2, pp. 81–90.

Stubbs, M., Harris, M., Solomonow, M., Zhou, B., Lu, Y., and Baratta, R. V. (1998), "Ligamento-Muscular Protective Reflex in the Lumbar Spine of the Feline," *Journal of Electromyography and Kinesiology*, Vol. 8, No. 4, pp. 197–204.

Tubach, F., Leclerc, A., Landre, M. F., and Pietri-Taleb, F. (2002), "Risk Factors for Sick Leave Due to Low Back Pain: A Prospective Study," *Journal of Occupational and Environmental Medicine*, Vol. 44, No. 5, pp. 451–458.

van Poppel, M. N., Koes, B. W., Deville, W., Smid, T., and Bouter, L. M. (1998), "Risk Factors for Back Pain Incidence in Industry: A Prospective Study," *Pain*, Vol. 77, No. 1, pp. 81–86.

Vernon-Roberts, B., and Pirie, C. J. (1973), "Healing Trabecular Microfractures in the Bodies of Lumbar Vertebrae," *Annals of Rheumatic Diseases*, Vol. 32, No. 5, pp. 406–412.

Vernon-Roberts, B., and Pirie, C. J. (1977), "Degenerative Changes in the Intervertebral Discs of the Lumbar Spine and Their Sequelae," *Rheumatology Rehabilitation*, Vol. 16, No. 1, pp. 13–21.

Videman, T., Nurminen, M., and Troup, J. D. (1990), "1990 Volvo Award in Clinical Sciences. Lumbar Spinal Pathology in Cadaveric Material in Relation to History of Back Pain, Occupation, and Physical Loading," *Spine*, Vol. 15, No. 8, pp. 728–740.

Vingard, E., et al. (2002), "Seeking Care for Low Back Pain in the General Population: A Two-Year Follow-up Study: Results from the MUSIC-Norrtalje Study," *Spine*, Vol. 27, No. 19, pp. 2159–2165.

Waddell, G. (1992), "Biopsychosocial Analysis of Low Back Pain," *Bailliere's Clinical Rheumatology*, Vol. 6, No. 3, pp. 523–557.

Waddell, G. (1999), *The Back Pain Revolution*, Churchill Livingstone, Edinburgh.

Wang, E. (1999), "Age-Dependent Atrophy and Microgravity Travel: What Do They Have in Common?" *FASEB Journal*, Vol. 13, Suppl., pp. S167–174.

Wang, J. L., Parnianpour, M., Shirazi-Adl, A., and Engin, A. E. (2000), "Viscoelastic Finite-Element Analysis of a Lumbar Motion Segment in Combined Compression and Sagittal Flexion. Effect of Loading Rate," *Spine*, Vol. 25, No. 3, pp. 310–318.

Waters, T. R., Putz-Anderson, V., Garg, A., and Fine, L. J. (1993), "Revised NIOSH Equation for the Design and Evaluation of Manual Lifting Tasks," *Ergonomics*, Vol. 36, No. 7, pp. 749–776.

Waters, T. R., Putz-Anderson, V., and Garg, A. (1994), *Applications Manual for the Revised NIOSH Lifting Equation*, National Institute for Occupation Safety and Health.

Waters, T. R., et al. (1999), "Evaluation of the Revised NIOSH Lifting Equation. A Cross-Sectional Epidemiologic Study," *Spine*, Vol. 24, No. 4, pp. 386–394; discussion 395.

Woo, S. L.-Y., Gomez, M. A., and Akeson, W. H. (1985), "Mechanical Behaviors of Soft Tissues: Measurements, Modifications, Injuries, and Treatment," in *Biomechanics of Trauma*, H. M. Nahum and J. Melvin, Eds., Appleton Century Crofts, Norwalk, CT, pp. 109–133.

Yamashita, T., Minaki, Y., Ozaktay, A. C., Cavanaugh, J. M., and King, A. I. (1996), "A Morphological Study of the Fibrous Capsule of the Human Lumbar Facet Joint," *Spine*, Vol. 21, No. 5, pp. 538–543.

Yang, G., Marras, W., and Best, T. (2010), "The Biochemical Response to Biomechanical Tissue Loading on the Low Back during Physical Work Exposure," *Clinical Biomechanics*, in review.

Yingling, V. R., and McGill, S. M. (1999a), "Anterior Shear of Spinal Motion Segments. Kinematics, Kinetics, and Resultant Injuries Observed in a Porcine Model," *Spine*, Vol. 24, No. 18, pp. 1882–1889.

Yingling, V. R., and McGill, S. M. (1999b), "Mechanical Properties and Failure Mechanics of the Spine under Posterior Shear Load: Observations from a Porcine Model," *Journal of Spinal Disorders*, Vol. 12, No. 6, 501–508.

# CHAPTER 28

# WORK-RELATED UPPER EXTREMITY MUSCULOSKELETAL DISORDERS

**David Rodrick**
**University of Michigan—Dearborn**
**Dearborn, Michigan**

**Waldemar Karwowski**
**University of Central Florida**
**Orlando, Florida**

**William S. Marras**
**The Ohio State University**
**Columbus, Ohio**

## 1 INTRODUCTION

### 1.1 Extent of Problem

The annual health and economic burden of work-related musculoskeletal disorders (WMSDs) in the United States was estimated at between $13 and $54 billion [National Institute for Occupational Safety and Health (NIOSH), 2001]. To adequately address the severity and burden of WMSDs, the NIOSH formed the National Occupational Research Agenda (NORA), which designated musculoskeletal disorders as one of its 21 priority research areas. Musculoskeletal disorders include nonfatal and/or nontraumatic injuries to the

upper extremity, neck, back, trunk, and lower extremity. The U.S. Bureau of Labor Statistics (BLS) reported a total of 335,390 cases, an incidence rate of 35.4, and median days away from work as 9 days in 2007 (Bureau of Labor Statistics, 2008, reissued in 2009). WMSD cases accounted for about one-third of 1.15 million cases of nonfatal injuries and illnesses involving days away from work in private industries. The BLS reported highest incidence rates in the following five occupations: (1) nursing aides, orderlies, and attendants; (2) emergency medical technicians and paramedics; (3) laborers and freight stock and material movers; (4) reservation and transportation ticket agents; and (5) light or delivery service truck drivers. According to the report, in 2007, days away from work declined by 21,770 cases while the incidence rate decreased by 8% from 2006. The survey utilized four characteristics—nature (e.g., sprain), part of the body (e.g., back), source (e.g., health care patient), and event of exposure (e.g., overexertion in lifting)—to characterize the cases of WMSDs.

## 1.2   Definitions

Sommerich et al. (2006) reviewed the definition of the umbrella term WMSD. According to Hagberg et al. (1995), the term *work-related musculoskeletal disorders* defines those disorders and diseases of the musculoskeletal system that have a proven or hypothetical work-related causal component. Musculoskeletal disorders are pathological entities in which the functions of the musculoskeletal system are disturbed or abnormal, whereas diseases are defined pathological entities with observable impairments in body configuration and function. Although work-related upper extremity disorders (WUEDs) are a heterogeneous group of disorders, and the current state of knowledge does not allow for a general description of the course of these disorders, it is possible nevertheless to identify a group of generic risk factors, including biomechanical factors, such as static and dynamic loading on the body and posture, cognitive demands, and organizational and psychosocial factors, for which there is evidence of work relatedness and a higher risk of developing WUEDs.

The generic risk factors, which typically interact and accumulate to form cascading cycles, are assumed to be directly responsible for pathophysiological phenomena which depend on location, intensity, temporal variation, duration, and repetitiveness of the generic risk factors (Hagberg et al., 1995). It was also proposed that both insufficient and excessive loading on the musculoskeletal system have deleterious effects and that the pathophysiological process is dependent on a person's characteristics with respect to body responses, coping mechanisms, and adaptation to risk factors. The generic risk factors, workplace design features, and pathophysiological phenomena are parts of the generic model for WUED prevention proposed by Armstrong et al. (1993) and demonstrated in later in this chapter in Figure 4.

## 1.3   Cumulative-Trauma Disorders of Upper Extremity

Since WMSDs are used as an umbrella term rather a diagnostic term, different regions of the world label

these disorders differently. For instance, WMSD is labeled as repetitive-strain injury (RSI) in Canada and Europe, both RSI and occupational overuse syndrome in Australia, and cumulative-trauma disorder (CTD) in the United States. Putz-Anderson (1993) defined CTD by combining the literal meaning of each word. *Cumulative* indicates that these disorders develop gradually over periods of time as a result of repeated stresses. The cumulative concept is based on the assumption that each repetition of an activity produces some trauma or wear and tear on the tissues and joints of the particular body part. The term *trauma* indicates bodily injury from mechanical stresses, whereas *disorders* refers to physical ailments. The definition above also stipulates a simple cause-and-effect model for CTD development. According to such a model, since the human body needs sufficient intervals of rest time between episodes of repeated strains to repair itself, if the recovery time is insufficient, combined with a high repetition of forceful and awkward postures, the worker is at higher risk of developing a CTD. In the context of the generic model for prevention proposed by Armstrong et al. (1993), the definition above is oriented primarily toward biomechanical risk factors for WUEDs and therefore is incomplete.

Ranney et al. (1995) pointed out that the evidence that chronic musculoskeletal disorders of the upper extremities are work related is growing rapidly. Several comprehensive reviews have examined multiple sources of evidence and data [Bernard, 1997; Viikari-Juntura and Silverstein, 1999; National Research Council (NRC), 2001; Buckle and Devereux, 2002]. Palmer and Smedley (2007) noted that most investigations suffer from such limitations as small sample size, confounding, incomplete blinding, and crude exposure assessment. Despite those limitations, they have found evidence of neck pain associated with work-related exposure. Another recent study by Harrington et al. (2009) found association between increase in job demand and upper extremity symptoms.

Previously, Armstrong et al. (1993) concluded that it was not possible to define the dose–response relationships and exposure limits for the WUED problems. To establish the work relatedness of these disorders, both the quantification of exposures involved in work and a determination of health outcomes, including details of the specific disorders (Luopajarvi et al., 1979; Moore et al., 1991; Stock, 1991; Hagberg, 1992), are needed. Also, more detailed medical diagnoses are required for choosing appropriate exposure measures as well as for structuring treatment, screening, and prevention programs (Ranney et al., 1995). However, due to recent development in valid and reliable construction of exposure assessment and quantification, a few studies were able to establish dose–response relationships. For example, recently, Engholm and Holmstrom (2005) found a location-specific dose–response relationship between work-related physical factors, that is, awkward posture and musculoskeletal disorders among construction workers. Another study by Sauni et al. (2009) on finish metal workers, by means of a questionnaire on exposure and symptoms, found a dose–response

relationship between the cumulative lifetime vibration dose of hand–arm vibration and finger blanching, sensorineural symptoms, symptoms of carpal tunnel syndrome, and musculoskeletal symptoms of upper limbs and neck.

## 2 CONCEPTS AND CHARACTERISTICS

### 2.1 Epidemiology

The World Health Organization (WHO, 1985) defines an *occupational disease* as a disease for which there is a direct cause–effect relationship between hazard and disease (e.g., silica-silicosis). Work-related diseases (WRDs) are defined as multifactorial when the work environment and the performance of work contribute significantly to the causation of disease (WHO, 1985). Work-related diseases can be partially caused by adverse work conditions. However, personal characteristics, environmental factors, and sociocultural factors are also recognized as risk factors for these diseases.

As reviewed by Armstrong et al. (1993) and summarized by Hagberg et al. (1995), Bernard (1997), and NRC (2001), evidence of the work relatedness of musculoskeletal disorders is established by the pattern supplied through numerous epidemiological studies conducted over the last 35 years of research in the field. The incidence and prevalence of musculoskeletal disorders in the reference populations of the studies were low but not zero, indicating that there are non-work-related causes of these disorders as well. Such variables as cultural differences and psychosocial and economic factors, which may influence one's perception and tolerance of pain and consequently affect the willingness to report musculoskeletal problems, may have a significant impact on the progressions from disorder to work disability (WHO, 1985; Leino, 1989). Descriptions of common musculoskeletal disorders and related job activities were summarized by Kroemer et al. (1994) and are given in Table 1.

### 2.2 Evidence of Work Relatedness of Upper Extremity Disorders

The WRDs of the upper extremity include, among others, carpal tunnel syndrome, tendonitis, ganglionitis, tenosynovitis, bursitis, and epicondylitis (Putz-Anderson, 1993). As reported in a recent BLS (Bureau of Labor Statistics, 2008) account, workers employed in nursing, emergency medicine, laborers and freight stock and material movers, reservation and transportation ticket agents, and light or delivery service truck drivers are at a high risk of developing WUEDs, having incidence rates between 117 and 252. Although the occurrence of WUEDs at work has been well documented (Hagberg et al., 1995), because of the high complexity of the problem, there is a lack of clear understanding of the cause–effect relationship characteristics for these disorders, which prevents accurate classification, implementation of effective control measures, and/or subsequent rehabilitation and return-to-work strategies. The problem may be confounded

by poor management–labor relationships and lack of willingness for open communication about the potential problems and how to solve them for the fear of legal litigation, including claims of unfair labor practices (Sommerich et al., 2006).

### 2.2.1 Definition of Risk Factors

Risk factors are defined as variables that are believed to be related to the probability of a person's developing a disease or disorder (Kleinbaum et al., 1982). Hagberg et al. (1995) classified the generic risk factors for development of WMSDs by considering their explanatory value, biological plausibility, and the relation to the work environment: These generic risk factors are (1) fit, reach, and see; (2) musculoskeletal load; (3) static load; (4) postures; (5) cold, vibration, and mechanical stresses; (6) task in invariability; (7) cognitive demands; and (8) organizational and psychosocial work characteristics. These WMSD risk factors are present at varying levels for different jobs and tasks. They are assumed to interact and to have an accumulative effect, forming the cascading cycles described by Armstrong et al. (1993), the extent and severity of which depends on their intensity, duration, and so on, meaning that the mere presence of a risk factor does not necessarily suggest that an exposed worker is at excessive risk of injury.

### 2.2.2 Biomechanical Factors

According to Armstrong et al. (1986), the following are categories of biomechanical risk factors for development of WUEDs: (1) forceful exertions and motions; (2) repetitive exertions and motions; (3) extreme postures of the shoulder (elbow above midtorso or reaching down and behind), forearm (inward or outward rotation with a bent wrist), wrist (palmar flexion or full extensions), and hand (pinching); (4) mechanical stress concentrations over the base of the palm, on the palmar surface of the fingers, and on the sides of the fingers; (5) duration of exertions, postures, and motions; (6) effects of hand–arm vibration; (7) exposure to a cold environment; (8) insufficient rest or break time; and (9) the use of gloves. Furthermore, wrist angular flexion–extension acceleration was also determined to be a potential risk factor for hand–wrist CTDs under conditions of dynamic industrial tasks (Marras and Schoenmarklin, 1993; Schoenmarklin et al., 1994). More recently, comprehensive reviews of epidemiological and experimental studies of work-related MSDs have concluded that there is sufficient evidence of associations between physical exposures in the workplace and MSDs (Bernard, 1997; Viikari-Juntura and Silverstein, 1999; NRC, 2001; Buckle and Devereux, 2002; Larsson et al., 2007). Conclusions from the reviews by Bernard (1997) and NRC (2001) are summarized in Tables 2 and 3, respectively. Tables 4 and 5 present a summary of work-related postural risk factors for wrist and shoulder disorders.

Inadequate design of tools with respect to weight and size can impose extreme wrist positions and high forces on a worker's musculoskeletal system (Armstrong et al., 1993). For example, holding a heavier object requires increased power grip and high tension in the finger flexor

**Table 1 Common WMSDs**

| Disorder[a] | Description | Typical Job Activities |
|---|---|---|
| Carpal tunnel syndrome (writer's cramp, neuritis, median neuritis) (N) | The result of compression of the median nerve in the carpal tunnel of the wrist. This tunnel is an opening under the carpal ligament on the palmar side of the carpal bones. Through this tunnel pass the median nerve and the finger flexor tendons. Thickening of the tendon sheaths increases the volume of tissue in the tunnel, thereby increasing pressure on the median nerve. The tunnel volume is also reduced if the wrist is flexed or extended or ulnarly or radially pivoted. | Buffing, grinding, polishing, sanding, assembly work, typing, keying, cashiering, playing musical instruments, surgery, packing, housekeeping, cooking, butchering, hand washing, scrubbing, hammering |
| Cubital tunnel syndrome (N) | Compression of the ulnar nerve below the notch of the elbow. Tingling, numbness, or pain radiating into ring or little fingers. | Resting forearm near elbow on a hard surface and/or sharp edge, also when reaching over obstruction |
| DeQuervain's syndrome (or disease) (T) | A special case of tenosynovitis which occurs in the abductor and extensor tendons of the thumb, where they share a common sheath. This condition often results from combined forceful gripping and hand twisting, as in wringing clothes. | Buffing, grinding, polishing, sanding, pushing, pressing, sawing, cutting, surgery, butchering, use of pliers, "turning" control such as on a motorcycle, inserting screws in holes, forceful hand wringing |
| Epicondylitis ("tennis elbow") | Tendons attaching to the epicondyle (the lateral protrusion at the distal end of the humerus bone) become irritated. This condition is often the result of impacts of jerky throwing motions, repeated supination and pronation of the forearm, and forceful wrist extension movements. The condition is well known among tennis players, pitchers, bowlers, and people hammering. A similar irritation of the tendon attachments on the inside of the elbow is called medical epicondylitis, also known as "golfer's elbow." | Turning screws, small parts assembly, hammering, meat cutting, playing musical instruments, playing tennis, pitching, bowling |
| Ganglion (T) | A tendon sheath swelling that is filled with synovial fluid or a cystic tumor at the tendon sheath or a joint membrane. The affected area swells up and causes a bump under the skin, often on the dorsal or radial side of the wrist. (Since it was in the past occasionally smashed by striking with a Bible or heavy book, it was also called a "Bible bump.") | Buffing, grinding, polishing, sanding, pushing, pressing, sawing, cutting, surgery, butchering, use of pliers, "turning" control such as on a motorcycle, inserting screws in holes, forceful hand wringing |
| Neck tension syndrome (M) | An irritation of the levator scapulae and the trapezius muscle of the neck, commonly occurring after repeated or sustained work. | Belt conveyor assembly, typing, keying, small parts assembly, packing, load carrying in hand or on shoulder, overhead work |
| Pronator (teres) syndrome | Result of compression of the median nerve in the distal third of the forearm, often where it passes through the two heads of the pronator teres muscle in the forearm; common with strenuous flexion of elbow and wrist. | Soldering, buffing, grinding, polishing, sanding |
| Shoulder tendonitis (rotator cuff syndrome or tendonitis, supraspinatus tendonitis, subacromial bursitis, subdeltoid bursitis, partial tear of the rotator cuff) (T) | The rotator cuff consists of four muscles and their tendons that fuse over the shoulder joint. They medially and laterally rotate the arm and help to abduct it. The rotator cuff tendons must pass through a small bony passage between the humerus and the acromion with a bursa as cushion. Irritation and swelling of the tendon or of the bursa are often caused by continuous muscle and tendon effort to keep the arm elevated. | Punch press operations, overhead assembly, overhead welding, overhead painting, overhead auto repair, belt conveyor assembly work, packing, storing, construction work, postal "letter carrying," reaching, lifting, carrying load on shoulder |

**Table 1** (*Continued*)

| Disorder[a] | Description | Typical Job Activities |
|---|---|---|
| Tendonitis (T) | An inflammation of a tendon. Often associated with repeated tension, motion, bending, being in contact with a hard surface, vibration. The tendon becomes thickened, bumpy, and irregular in its surface. Tendon fibers may be frayed or torn apart. In tendons without sheaths, such as the biceps tendon, the injured area may calcify. | Punch press operation, assembly work, wiring, packaging, core making, use of pliers |
| Tenosynovitis (tendosynovitis, tendovaginitis) (T) | Inflammation of the synovial sheaths. The sheath swells. Consequently, movement of the tendon with the sheath is impeded and painful. The tendon surfaces can become irritated, rough, and bumpy. If the inflamed sheath presses progressively onto the tendon, the condition is called stenosing tendosynovitis. ''DeQuervain's syndrome'' (see there) is a special case occurring at the thumb; the ''trigger finger'' (see there) condition occurs in flexors of the fingers. | Buffing, grinding, polishing, sanding, pushing, pressing, sawing, cutting, surgery, butchering, use of pliers, ''turning'' control such as on a motorcycle, inserting screws in holes, forceful hand wringing |
| Thoracic outlet syndrome (neurovascular compression syndrome, cervicobrachial disorder, brachial plexus neuritis, costoclavicular syndrome, hyperabduction syndrome) (V, N) | A disorder resulting from compression of the nerves and blood vessels of the brachial plexus between clavicle and first and second ribs. If this neurovascular bundle is compressed by the pectoralis minor muscle, blood flow to and from the arm is reduced. This ischemic condition makes the arm numb and limits muscular activities. | Buffing, grinding, polishing, sanding, overhead assembly, overhead welding, overhead painting, overhead auto repair, typing, keying, cashiering, wiring, playing musical instruments, surgery, truck driving, stacking, material handling, postal ''letter carrying,'' carrying heavy loads with extended arms |
| Trigger finger (or thumb) (T) | A special case of tenosynovitis (see there) where the tendon forms a nodule and becomes nearly locked, so that its forced movement is not smooth but in a snapping or jerking manner. This is a special case of stenosing tenosynovitis crepitans, a condition usually found with the digit flexors at the AI ligament. | Operating trigger finger, using hand tools that have sharp edges pressing into the tissue or whose handles are too far apart for the user's hand so that the end segments of the fingers are flexed while the middle segments are straight |
| Ulnar artery aneurysm (V, N) | Weakening of a section of the wall of ulnar artery as it passes through the Guyon tunnel in the wrist; often from pounding or pushing with the heel of the hand. The resulting ''bubble'' presses on the ulnar nerve in the Guyon tunnel. | Assembly work |
| Ulnar nerve entrapment (Guyon tunnel syndrome) (N) | Results from the entrapment of the ulnar nerve as it passes through the Guyon tunnel in the wrist. It can occur from prolonged flexion and extension of the wrist and repeated pressure on the hypothenar eminence of the palm. | Playing musical instruments, carpentering, brick laying, use of pliers, soldering, hammering |
| White finger (''dead finger,'' Raynaud's syndrome, vibration syndrome) (V) | Stems from insufficient blood supply bringing about noticeable blanching. Finger turns cold or numb and tingles, and sensation and control of finger movement may be lost. The condition results from closure of the digit's arteries caused by vasospasms triggered by vibrations. A common cause in continued forceful gripping of vibrating tools particularly in a cold environment. | Chain sawing, jackhammering, use of vibrating tool, sanding, paint scraping, using vibrating tool too small for the hand, often in a cold environment |

*Source:* Adapted from Kroemer et al. (1994).
[a]Type of disorder: N, nerve; M, muscle; V, vessel; T, tendon.

**Table 2  Evidence for Causal Relationship between Physical Work Factors and MSDs**

| Body Part | Risk Factor | Strong Evidence | Evidence | Insufficient Evidence | Evidence of No Effect |
|---|---|:---:|:---:|:---:|:---:|
| Neck and | Repetition | | x | | |
| neck/shoulder | Force | | x | | |
| | Posture | x | | | |
| | Vibration | | | x | |
| Shoulder | Repetition | | x | | |
| | Force | | | x | |
| | Posture | | x | | |
| | Vibration | | | x | |
| Elbow | Repetition | | | x | |
| | Force | | x | | |
| | Posture | | | x | |
| | Combination | x | | | |
| Hand/wrist | Repetition | | x | | |
| Carpal tunnel | Force | | x | | |
| syndrome | Posture | | | x | |
| | Vibration | | x | | |
| | Combination | x | | | |
| Tendonitis | Repetition | | x | | |
| | Force | | x | | |
| | Posture | | x | | |
| | Combination | x | | | |
| Hand–arm vibration syndrome | Vibration | x | | | |

*Source:* Bernard (1997).

**Table 3  Summary of Epidemiological Studies Reviewed by NRC (2001) for Evidence of Association between Work-Related Physical Exposures and WUED**

| Focus of Study | Number of Studies Providing Risk Estimates | Number of Studies That Found Significant Positive Associations between Physical Risk Factor Exposure and WUED |
|---|:---:|:---:|
| Risk factor | | |
| Manual materials handling | 28 | 24 |
| Repetition | 8 | 4 |
| Force | 3 | 2 |
| Repetition and force | 2 | 2 |
| Repetition and cold | 1 | 1 |
| Vibration | 32 | 26 |
| Disorder | | |
| Carpal tunnel syndrome | 18[a] | 12 |
| Hand–arm vibration syndrome | 13 | 12 |

[a]Nine studies provided 18 risk estimators.

tendons, which causes increased pressure in the carpal tunnel. Furthermore, a task that induces hand and arm vibration causes an involuntary increase in power grip through a reflex of the strength receptors. Vibration can also cause protein leakage from the blood vessels in the nerve trunks and result in edema and increased pressure in the nerve trunks and therefore can also result in edema

and increased pressure in the nerve (Lundberg et al., 1990).

Historically, several millions of workers in occupations such as vehicle operation are intermittently exposed every year to hand–arm vibration that significantly stresses the musculoskeletal system (Haber, 1971). Hand–arm vibration syndrome (HAVS) that is

**Table 4 Postural Risk Factors Reported in the Literature for the Wrist**

| Risk Factor | Results: Outcome and Details |
| --- | --- |
| Wrist flexion | Carpal tunnel syndrome (CTS), exposure of 20–40 h per week; increased median nerve stresses (pressure); increased finger flexor muscle activation for grasping; median nerve compression by flexor tendons. |
| Wrist extension | Median nerve compression by flexor tendons; CTS, exposure of 20–40 h per week; increased intracarpal tunnel pressure for extreme extension of 90°. |
| Wrist ulnar deviation | Exposure response effect found: If deviation greater than 2W, increased pain and pathological findings. |
| Deviated wrist positions | Workers with CTS used these postures more often. |
| Hand manipulations | More than 1500–2000 manipulations per hour led to tenosynovitis. |
| Wrist motion | 1276 flexion extension motions led to fatigue; higher wrist accelerations, and velocities in high-risk wrist WMSD jobs. |

*Source:* Adapted from Kuorinka and Forcier (1995).

characterized by intermittent numbness and blanching of the fingers with reduced sensitivity to heat, cold, and pain can affect up to 90% of workers in occupations such as chipping, grinding, and chain-sawing (Wasserman et al., 1974; Taylor and Pelmear, 1976). HAVS is caused primarily by vibration of a part or parts of the body of which the main sources are hand-held power tools such as chain saws and jackhammers (Sommerich et al., 2006).

A very recent study by da Costa and Vieira (2010) that reviewed only case–control or cohort studies utilized five criteria for association: strength of association, consistency, between studies, temporality, dose–response relationship, and coherence (following Bernard, 1997). In their study, da Costa and Vieira (2010) considered "strong evidence risk factors" for those that would satisfy at least four of the five, "reasonable evidence risk factors" for those that would satisfy at least one of the five, and "insufficient evidence risk factors" for those that would satisfy none of the criteria for causality but rather presented clear bias or confounding factors. The review study categorized these three levels of risk factors to specific body segments; see Table 6.

### 2.2.3 Work Organization Factors

According to Sommerich et al. (2006), work organization and environmental factors may also play a significant role in development of WUEDs. Work organization is defined as "the objective nature of the work process. It deals with the way in which work is structured, supervised, and processed" (Hagberg et al., 1995).

**Table 5 Postural Risk Factors Reported in the Literature for the Shoulder**

| Risk Factor | Results: Outcome and Details |
| --- | --- |
| More than 60° abduction or flexion for more than 1 h per day | Acute shoulder and neck pain |
| Less than 15° median upper arm flexion and 10° abduction for continuous work with low loads | Increased sick leave resulting from musculoskeletal problems |
| Abduction greater than 30° | Rapid fatigue at greater abduction angles |
| Abduction greater than 45° | Rapid fatigue at 90° |
| Shoulder forward flexion of 30°, abduction greater than 30° | Hyperabduction syndrome with compression of blood vessels |
| Hands no greater than 35° above shoulder level | Impairment of blood flow in the supraspinatus muscle |
| Upper arm flexion or abduction of 90° | Onset of local muscle fatigue |
| Hands at or above shoulder height | Electromyographic signs of local muscle fatigue in less than 1 min |
| Repetitive shoulder flexion | Tendonitis and other shoulder disorders |
| Repetitive shoulder abduction or flexion | Acute fatigue |
| Postures invoking static shoulder loads | Neck–shoulder symptoms negatively related to movement rate |
| Arm elevation | Tendonitis and other shoulder disorders |
| Shoulder elevation | Pain |
| Shoulder elevation and upper arm abduction | Neck–shoulder symptoms |
| Abduction and forward flexion invoking static shoulder loads | Neck–shoulder symptoms, shoulder pain and sick leave resulting from musculoskeletal problems |
| Overhead reaching and lifting | Pain |

*Source:* Adapted from Kuorinka and Forcier (1995).

**Table 6  Levels of Evidence of Different Risk Factors of WMSDs by Body Parts**

| Body Part | Strong Evidence | Reasonable Evidence | Insufficient Evidence |
|---|---|---|---|
| Neck | None | Psychosocial factors<br>Smoking<br>Gender<br>Posture<br>Comorbidity | Heavy physical work<br>Lifting<br>Sedentary work<br>Aging<br>High body mass index |
| Nonspecific upper limb | None | Comorbidity | Psychosocial factors<br>Aging<br>Smoking<br>Heavy physical work<br>High body mass index |
| Shoulder | None | Heavy physical work<br>Psychosocial factors | Repetitive work<br>Aging<br>High body mass index<br>Sedentary work |
| Elbow/forearm | None | Awkward posture<br>Comorbidity<br>Repetitive work<br>Aging | High body mass index<br>Heavy physical work<br>Gender<br>Monotonous work<br>Associated WUED |
| Wrist/hand | None | Prolonged computer work<br>Heavy physical work<br>High body mass index<br>Aging<br>Gender<br>Awkward posture<br>Repetitive work | Smoking<br>Comorbidity<br>Psychosocial factors |

The mechanisms by which work organizational factors can modify the risk for WUEDs include modifying the extent of exposure to other risk factors (physical and environmental) and modifying a person's stress response, thereby increasing the risk associated with a given level of exposure. Specific work organization factors that have been shown to fall into at least one of these categories include (but are not limited) to the following: (1) wage incentives, (2) machine-paced work, (3) workplace conflicts of many types, (4) absence of worker decision latitude, (5) time pressures and work overload, and (6) unaccustomed work during training periods or after returning from long-term leave. As discussed by Hagberg et al. (1995), the organizational context in which work is carried out has major influences on a worker's physical and psychological stress and health. The work organization defines the level of work output required (work standards), the work process (how the work is carried out), the work cycle (work–rest regimens), the social structure, and the nature of supervision.

### 2.2.4  Psychosocial Work Factors

Psychosocial work factors are defined as "the subjective aspects of work organization and how they are perceived by workers and managers" (Hagberg et al., 1995). Factors commonly investigated include job dissatisfaction and perceptions of workload, supervisor, and co-worker support, job control, monotony of work, job clarity, and interactions with clients. Recent reviews have concluded that high perceived job stress and non-work-related stress (worry, tension, distress) are consistently linked to WUEDs (NRC, 2001; Bongers et al., 2002). Bernard (1997) and NRC (2001) also found high job demands/workload to be linked to WUED in the majority of studies they reviewed that considered it. The NRC (2001) review concluded that the evidence was insufficient for linking WUED and low decision latitude, social support, or limited rest–break opportunities. MacDonald et al. (2001) presented evidence of high degrees of correlation between some physical and psychosocial work factors. This makes studying these factors more difficult, requiring collection of information on both types of factors and use of sophisticated data analysis techniques in order to draw correct conclusions about any associations between risk factors and WUED. However, it also means that the risk factors may be linked through work organization elements and that, by thoughtfully addressing those elements, exposures to both physical and psychosocial risk factors may be reduced.

### 2.2.5  Individual Factors

Individual characteristics of a worker, including anthropometry, health, sex, and age, may alter the way in

which work is performed and may affect a worker's capacity for or tolerance of exposure to physical or other risk factors. In particular, Lundberg (2002) and Treaster and Burr (2004) determined that women were at greater risk than men for WUED development. Treaster and Burr (2004) found this to hold even after accounting for confounders such as age and work factor exposure. They suggested a number of reasons why this might occur, including differences in exposures due to mismatches between female workers and their workstations, tools, and strength requirements of tasks, and so on (anthropometry); psychosocial or psychological factors, which may be the result of differences in job status (e.g., many women's jobs may tend to have less autonomy); the perceived need to work harder to prove one's self in a male-dominated workplace or profession; additional psychological pressure or workload due to responsibilities outside work (care of home, children, or aging parents); or biological differences, possibly related to the effects of sex hormones on soft tissues (Hart et al., 1998).

### 2.2.6 Basic Classification of Disorders

Since most manual work requires the active use of the arms and hands, the structures of the upper extremity are particularly vulnerable to soft tissue injury. WUEDs are typically associated with repetitive manual tasks with forceful exertions, such as those performed at assembly lines, or when using hand tools, computer keyboards, and other devices or operating machinery. These tasks impose repeated stresses to the upper body,

that is, muscles, tendons, ligaments, nerve tissues, and neurovascular structures. WEUD may be classified by the type of tissue that is primarily affected. Table 7 lists a number of upper extremity MSDs by the tissue that is primarily affected.

Generally, the greater the exposure to a single risk factor or combination of factors, the greater the risk of a WMSD. Furthermore, as the number of risk factors present increases, so does the risk of injury. The interaction between risk factors is more likely to have a multiplicative rather than an additive effect. Evidence for this can be found in the investigation of the effects of repetition and force exposure by Silverstein et al. (1986, 1987). However, risk factors may pose minimal risk of injury if sufficient exposure does not occur or if sufficient recovery time is provided. It is known that changes in the levels of risk factors will result in changes in the risk of WUEDs. Therefore, a reduction in WUED risk factors should also reduce the risk for WMSDs.

### 2.3 Physical Assessments of Workers

Ranney et al. (1995) performed precise physical assessments of workers in highly repetitive jobs as part of a cross-sectional study to assess the association between musculoskeletal disorders and a set of work-related risk factors. A total of 146 female workers employed in five different industries (garment and automotive trim sewing, electronic assembly, metal parts assembly, supermarket cashiering, and packaging) were examined for the presence of potential work-related musculoskeletal disorders. The prerequisites for selection of industries

**Table 7 Upper Extremity MSDs Classified by Tissue That Is Primarily Affected**

| Tendon-Related Disorders | Nerve-Related Disorders | Muscle-Related Disorders | Circulatory/Vascular Disorders | Joint-Related Disorders | Bursa-Related Disorders |
|---|---|---|---|---|---|
| Paratenonitis, peritendonitis, tendonitis, tendinosis, tenosynovitis | Carpal tunnel syndrome | Muscle strain | Hypothenar hammer syndrome | Osteoarthritis | Bursitis |
| Epicondylitis | Cubital tunnel syndrome | Myofascial pain, trigger points, myositis, fibromyalgia, fibrositis | Raynaud's syndrome | | |
| Stenosing tenosynovitis (DeQuervain's disease; trigger finger) | Guyon canal syndrome | | | | |
| | Radial tunnel syndrome | | | | |
| Dupuytren's contracture | Thoracic outlet syndrome, digital neuritis | | | | |
| Ganglion cyst | | | | | |

*Source:* Almekinders (1998) and Buckle and Devereux (2002).
Notes (based on Almekinders (1998), NRC (2001); refer also to Figure 5 later in this chapter):
Paratenonitis: involves tendon sheath; para-, meso-, and epitenon.
Peritendonitis: involves para-, meso-, and epitenon (may also involve tendon sheath, depending on reference).
Tendonitis: involves tendon, endotenon.
Tendinosis (insertional or midsubstance): involves tendon, endotenon, tendon–bone junction; refers to situation where there are degenerative changes within the tendon without evidence of inflammatory cells.
Tenosynovitis: involves tendon sheath: para-, meso-, and epitenon: refers to inflammation of the involved tissue(s).
Stenosing tenosynovitis occurs when tendon gliding is restricted due to thickening of the sheath or tendon.

and tasks within these industries were (1) the existence of repetitive work using the upper limb, (2) at least 5–10 female workers performing the same repetitive job, (3) a range of jobs from light to demanding, (4) minimal job rotation, (5) no major change in the plant for at least one year, and (6) the support from both union or employee group and management.

The study showed that 54% of the workers had evidence of musculoskeletal disorders in the upper extremities that were judged as potentially work related. Many workers had multiple problems, and many were affected bilaterally (33% of workers). Muscle pain and tenderness were the largest problems in both the neck–shoulder area (31%) and the forearm–hand musculature (23%). Most forearm muscle problems were found on the extensor side. Carpal tunnel syndrome

(CTS) was the most common form of disorder, with 16 workers affected (7 people affected bilaterally). DeQuervain's tenosynovitis and wrist flexor tendonitis were the most commonly found tendon disorders in the distal forearm (12 workers affected for each diagnosis). In view of the study results, it was concluded that muscle tissue is highly vulnerable to overuse; that the stressors that affect muscle tissue, such as static loading, should be studied in the forearm as well as in the shoulder; and that exposure should be evaluated bilaterally. Finally, the predominance of forearm muscle and epicondyle disorders on the extensor side was linked to the dual role of these muscles for supporting the hands against gravity plus postural stability during grasping.

The criteria for establishing the work site diagnosis for various WMSDs are shown in Tables 8 and 9.

**Table 8  Minimal Clinical Criteria for Establishing Work Site Diagnoses for Work-Related Muscle or Tendon Disorders**

| Disorder | Symptoms | Examination |
|---|---|---|
| Neck myalgia | Pain in one or both sides of the neck increased by neck movement | Tender over paravertebral neck muscles |
| Trapezius myalgia | Pain on top of shoulder increased by shoulder elevation | Tender top of shoulder or medial border of scapula |
| Scapulothoracic pain syndrome[a] | Pain in scapular region increased by scapular movement | Tender over rib angles, 2, 3, 4, 5, and/or 6 |
| Rotator cuff tendonitis[b] | Pain in deltoid area or front of shoulder increased by glenohumeral movement | Rotator cuff tenderness[c] |
| Triceps tendonitis | Elbow pain increased by elbow movement | Tender triceps tendon |
| Arm myalgia | Pain in muscle(s) of the forearm | Tenderness in a specific muscle of the arm |
| Epicondylitis/tendonitis[d] | Pain localized to lateral or medial aspect of elbow | Tenderness of lateral or medial epicondyle localized to this area or to soft tissues attached for a distance of 1.5 cm |
| Forearm myalgia[e] | Pain in the proximal half of the forearm (extensor or flexor aspect) | Tenderness in a specific muscle in the proximal half of the forearm (extensor or flexor aspect) more than 1.5 cm distal to the condyle |
| Wrist tendonitis[f] | Pain on the extensor or flexor surface of the wrist | Tenderness is localized to specific tendons and is not found over bony prominences |
| Extensor finger tendonitis[f,g] | Pain on the extensor surface of the hand | Tenderness is localized to specific tendons and is not found over bony prominences |
| Flexor finger tendonitis[f,g] | Pain on the flexor aspect of the hand or distal forearm | Pain on resisted finger flexion localized to area of tendon |
| Tenosynovitis (finger/thumb) | Clicking or catching of affected digit on movement; may be pain or a lump in the palm | Demonstration of these complaints, tenderness anterior to metacarpal of affected digit |
| Tenosynovitis, DeQuervain's | Pain on the radial aspect of wrist | Tenderness over first tendon compartment and positive Finkelstein's test |
| Intrinsic hand myalgia | Pain in muscles of the hand | Tenderness in a specific muscle in the hand |

Source: Adapted from Ranney et al. (1995).
[a]Crepitation on circumduction of the shoulder.
[b]Positive impingement test.
[c]Frozen shoulder excluded.
[d]Positive Mills's test or reverse Mills's test (lateral or medial epicondylitis).
[e]Pain localized to the muscle belly of the muscle being stressed during resisted activity.
[f]Pain localized to tendon being stressed during resisted activity.
[g]Only diagnosed moderate or severe. Classification of severity of muscle/tendon problems: mild, above criteria met; moderate, pain persists more than 2 h after cessation or work but is gone after a night's sleep, or tenderness plus pain on resisted activity if localized in an anatomically correct manner, or see notes a, b, and d to f; severe, pain not completely relieved by a night's sleep.

**Table 9  Minimal Clinical Criteria for Establishing Work Site Diagnoses for Work-Related Neuritis**

| Disorder | Symptoms | Examination |
|---|---|---|
| Carpal tunnel syndrome | Numbness and/or tingling in thumb, index, and/or midfinger with particular wrist postures and/or at night | Positive Phalen's test or Tinel's sign present over the median nerve at the wrist |
| Scalenus anticus syndrome | Numbness and/or tingling on the preaxial border of the upper lip | Tender scalene muscles with positive Adson's or Wright's test |
| Cervical neuritis | Pain, numbness, or tingling following a dermatomal pattern in the upper limb | Clinical evidence of intrinsic neck pathology |
| Lateral ante-brachial neuritis | Lateral forearm pain, numbness, and tingling | Tenderness of coracobrachialis origin and reproduction of symptoms on palpation here or by resisted coracobrachialis activity |
| Pronator syndrome | Pain, numbness, and tingling in the median nerve distribution distal to the elbow | Tenderness of pronator teres or superficial finger flexor muscle, with tingling in the median nerve distribution on resisted activation of same |
| Cubital tunnel syndrome | Numbness and tingling distal to elbow in ulnar nerve distribution | Tender over ulnar nerve with positive Tinel's sign and/or elbow flexion test |
| Ulnar tunnel syndrome | Numbness and tingling in ulnar nerve distribution in the hand distal to the wrist | Positive Tinel's sign over the ulnar nerve at the wrist |
| Wartenberg's syndrome | Numbness and/or tingling in distribution of the superficial radial nerve | Positive Tinel's sign on tapping over the radial sensory nerve |
| Digital neuritis | Numbness or tingling in the fingers | Positive Tinel's sign on tapping over digital nerves |

*Source:* Adapted from Ranney et al. (1995).

Studies that include physical assessment to identify cases of a disorder are often viewed as more rigorous than those that rely exclusively on subjective recall in response to questions concerning musculoskeletal discomfort or disorder. However, there are a variety of ways in which physical examinations can be conducted and interpreted. When studying a disorder, adhering to a specific definition of the disorder is a key factor in classifying study participants as cases or noncases and thereafter determining the degree of association between the disorder and the risk factor(s) being studied. Recognizing the importance of this, criteria are put forth in order that researchers might begin to use common methods of classifying study of patients/participants. Consensus criteria for conducting physical assessments for classification of CTS in epidemiological studies are provided by Rempel et al. (1998). Sluiter et al. (2001) provided criteria for identifying 11 specific WUEDs, including CTS, and a four-step approach for determining work relatedness of a disorder once it is identified. Although these two groups took an expert consensus approach to criteria development, Helliwell et al. (2003) took a statistical approach, relying on multivariate modeling to identify the most discriminating symptoms and signs for classifying six different WUEDs. Authors of the latter two documents also address "nonspecific" upper extremity disorders as well.

Recently, David (2005) reviewed the various ergonomic methods used for assessing the risk factors of WMSDs. These methods are broadly categorized into three: (1) self-reports, (2) observational methods, and (3) direct measurements. Table 10 summarizes the methods, specific technique, main features, functions, and studies employed these techniques.

## 3  ANATOMY OF UPPER EXTREMITY

This section briefly discusses the anatomy of the hand, elbow, forearm, and shoulder following the description provided by Sommerich et al. (2006). The anatomy of the upper extremity provides for great functionality but also puts certain soft tissue components at risk of damage from repeated or sustained compressive, shear, and/or tensile loading. The shoulder complex joins the upper extremity to the axial skeleton. It provides the greatest range of motion of all the body's joints, yet this comes with several associated costs, including reduced joint stability and potential for entrapment of various soft tissues when the arm is elevated or loaded. All hand-held loads pass through the shoulder joint, but their effects are magnified increasingly the farther in the transverse plane the hands are located away from the shoulder. The elbow is a simpler joint than the shoulder yet can also be a sight of nerve entrapment. The hand is small in dimension yet is capable of producing large amounts of force. In addition, the hand is capable of configuring itself in a variety of orientations and can generate force with either the whole hand in a power grip or combinations of fingers in opposition to the thumb, as in a pinch grip. It is this very flexibility in capability that makes the upper extremity susceptible to CTDs. Refer

**Table 10  Ergonomic Methods for Assessing Exposure to Risk Factors**

| Method | Main Features/ Techniques | Function | Source |
|---|---|---|---|
| Self-reports | Ordinal scales for physical workload and musculoskeletal symptoms | Exposure assessment and prevalence of musculoskeletal symptoms | Viikari-Juntura et al. (1996) |
| | Visual analogue scales and categorical data | Estimates of the magnitude, frequency, and duration of work physical demands; assessment of risk factors; assessment of psychosocial risk factors for shoulder and neck pain | Pope et al. (1998) Spielholz et al. (1990) Holte and Westgaard (2001) |
| | Impact scales for handling work and Nordic questionnaire for MSD symptoms | Mechanical exposure estimates for the shoulder–neck region | Balogh et al. (2001) |
| | VIDAR — operator self-evaluation from video films of the work sequence | Worker ratings of load and estimations of related pain and discomfort | Kadefors and Forsman (2000) |
| | DMQ — categorical data for work load and hazardous working conditions(to provide seven indices) | Analysis of musculoskeletal workload and working conditions to identify higher risk groups | Hilderbrandt et al. (2001) |
| | Reporting of ergonomic exposures using Web-based recording method | Index of ergonomic exposures, pain, job stress, and functional limitations | Dane et al. (2002) |
| Observational methods | OWAS | Whole–body posture recording and analysis | Karhu et al. (1977) |
| | Checklist | Checklist for evaluating risk factors | Keyserling et al. (1992) |
| | RULA | Upper body and limb assessment | McAtamney and Corlett (1993) |
| | NIOSH lifting equation | Identification of risk factors and assessment | Waters et al. (1993) |
| | PLIBEL | Identification of risk factors | Kemmlert (1995) |
| | Strain index | Assessment of risk for distal upper extremity disorders | Moore and Garg (1995) |
| | OCRA | Integrated assessment scores for various types of jobs | Occhipinti (1998) |
| | QEC | Assessment of exposure of upper body and limb for static and dynamic tasks | Li and Buckle (1999a) |
| | Manual handling guidance, L23 | Checklist for identifying risk factors for manual handling | Health and Safety Executive (1998) |
| | REBA | Entire body assessment for dynamic tasks | Hignett and McAtamney (2000) |
| | FIOH risk factor checklist | Assessment of upper extremities | Ketola et al. (2001) |
| | ACGIH TLVs | Exposure assessment manual work | ACGIH Worldwide (2001) |
| | LUBA | Assessment of postural loading on the upper body and limbs | Kee and Karwowski (2001) |
| | Upper limb disorder guidance, HSG60, | Assessments of ULD risk factors | Health and Safety Executive (2002) |
| | MAC | Assessment of risk factors for individual and team manual handling tasks | Monnington et al. (2003) |

*(continued overleaf)*

**Table 10  (*Continued*)**

| Method | Main Features/ Techniques | Function | Source |
|---|---|---|---|
| Direct measurements | Video analysis | Posture assessment of hand/finger, computerized estimation of repetitiveness, body postures, force and velocity, assessment of dynamic and static tasks, measurement of trunk angles and angular velocities, various manual tasks | Armstrong et al. (1986) Yen and Radwin (1995) Fallentin et al. (2001) Speilholtz et al. (2001) Neuman et al. (2001) |
| | ROTA | Assessment of dynamic and static tasks | Ridd et al. (1989) |
| | TRAC | Assessment of dynamic and static tasks | Van der Beck et al. (1992) |
| | HARBO | Long–duration observation of various types of jobs | Wiktorin et al. (1995) |
| | PEO | Various tasks performed during period of job | Frasson et al. (1995) |
| | PATH | Nonrepetitive work | Buchholtz et al. (1996) |
| | SIMI motion | Assessment of dynamic movement of upper body and limbs | Li and Buckle (1999b) |
| | Biomechanical models | Estimation of internal exposures during task performance | Chaffin et al. (1999) |

*Source*: Adapted from David (2005).
Acronyms:
VIDAR – Video-och Datorbaserad Arbetsanalys
DMQ – Dutch Musculoskeletal Questionnaire
OWAS – Ovako Working Posture Analysis System
OCRA – The Occupational Repetitive Actions
QEC – Quick Exposure Checklist
FIOH – Finnish Institute of Occupational Health
TRAC – Task Recording and Analysis on Computer
RULA – Rapid Upper Limb Assessment
NIOSH – National Institute for Occupational Safety & Health
PLIBEL – A Method for Identification of Ergonomic Hazards
REBA – Rapid Entire Body Assessment
ACGIH – American Congress of Government Industrial Hygienists
LUBA – Loading on the Upper Body Assessment
MAC – Manual Handling Assessment Charts
HARBO – Hands Relative to the Body
PEO – Portable Ergonomic Observation

to Figures 1–3 for some views of the anatomy of the upper extremity.

### 3.1  Anatomy of Hand

The anatomy of the hand is illustrated in Figure 1. To achieve a variety of functions, the hand is constructed so that it contains numerous small muscles which facilitate fine, precise positioning of the hand and fingers but few power-producing muscles. One of the only power-producing muscles in the hand is a group of three muscles that form the thenar group, which flex, abduct, and position the thumb for opposition. Strong grasping is produced by extrinsic finger flexor muscles that are located in the forearm. Force is transmitted to the fingers through a network of long tendons (tendons attach muscles to bone). These tendons pass from the muscles in the forearm through the wrist (and through the carpal canal), through the hand, and to the fingers. These tendons are secured at various points along this path with ligaments that keep the tendons in close proximity

to the bones. The transverse carpal ligament forms the palmar boundary of the carpal tunnel, and the carpal bones form the remainder of the boundary. The tendons of the upper extremity are also encased in a sheath, which assists in the sliding of the tendons that occurs in concert with muscle contraction and prevents the tendons from sliding directly over the carpal bones or the transverse carpal ligament. A common sheath envelops the nine tendons passing through the carpal tunnel. For the fingers to generate force, a great deal of tension must be passed through these tendons. Since there are various possible combinations of tendons experiencing tension depending on the configuration of the fingers, type of grip, and grip force required, many of these tendons experience friction. This frictional component can be exacerbated by several factors, including position of the wrist and fingers, motion of the wrist and fingers, and insufficient rest periods. The extrinsic finger flexor muscles are paired with a set of extrinsic finger extensors on the dorsal side of the forearm and hand.
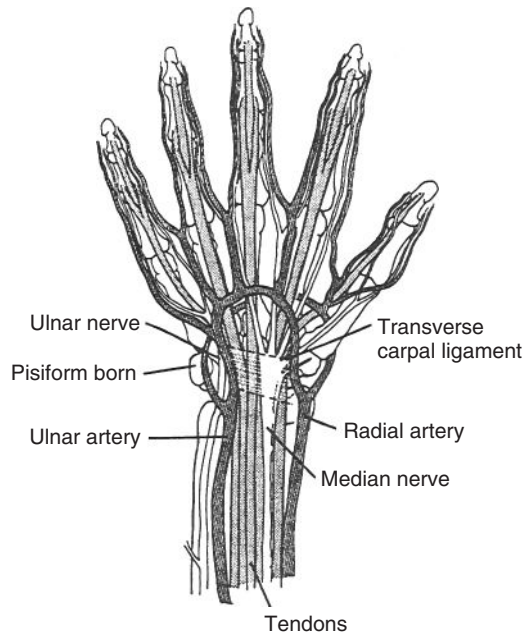
**Figure 1** Anatomy of the hand.



**Figure 2** Superficial muscles of the anterior forearm. The two sets of extrinsic muscles that flex the fingers are part of the intermediate and deep layers of anterior forearm muscles. (Modified from Gray's Anatomy, 1918.)

Other structures in the hand are also important to the development of cumulative trauma in the distal portion of the upper extremity. As shown in Figure 1, two major blood vessels pass through the hand. Both the radial artery and the ulnar artery provide the tissues and structures of the hand with a blood supply. One of the key structures of the hand that is often involved with cumulative-trauma experiences is the nerve structure. The median nerve enters the lower arm and passes through the carpal canal. Once it passes through the carpal canal, the median nerve becomes superficial at the base of the wrist and then branches off to serve the thumb, index finger, middle finger, and radial side of the ring finger. This nerve also serves the palmar surface of the hand connected to these fingers as well as the dorsal portion up to the first two knuckles on the fingers mentioned above as well as the thumb up to the first knuckle.

### 3.2 Anatomy of Forearm and Elbow

Located on the humerus, the medial epicondyle is the attachment site for the primary wrist flexor muscles and the extrinsic finger flexor muscles; the lateral epicondyle is the attachment site for the primary wrist extensor muscles and extrinsic finger extensor muscles. Repeated activation of either of these groups of muscles has been associated with development of epicondylitis at the relevant epicondyle. At the elbow, the ulnar nerve passes between the olecranon process of the ulna and the medial epicondyle. The space is referred to as the cubital tunnel. Cubital tunnel syndrome may develop as a result of compressive loading of the ulnar nerve as it passes through that tunnel when the elbow is flexed, either

repeatedly or over a sustained period. Direct pressure can also be applied to the nerve when part of the body's weight is supported by the elbows, depending on the shallowness of the space. The extrinsic finger extensor muscles, located on the posterior side of the forearm, are a common site of discomfort, as are their tendons and the tendons of the thumb's extrinsic extensor and abductor muscles. An illustration of the anterior view of the forearm appears in Figure 2.

### 3.3 Anatomy of Shoulder

The glenohumeral joint, where the head of the humerus partially contacts the glenoid fossa of the scapula, is what most people think of when the term shoulder is used. However, there are four articulations that make up the shoulder complex; the other three are the acromioclavicular, sternoclavicular, and scapulothoracic joints. The four joints work in a coordinated manner to provide the wide range of motion possible in a healthy shoulder. The shoulder complex is also composed of approximately 16 muscles and numerous ligaments. The extensive range of motion at the glenohumeral joint is afforded, in part, because of the minimal contact made between the humerus and scapula. The connection is secured by a group of muscles referred to as the rotator cuff (teres minor, infraspinatus, supraspinatus, and subscapularis muscles). They create a variety of torques about the joint and also help protect against subluxation (incomplete dislocation). The supraspinatus tendon is thought to be particularly susceptible to injury (tears and tendonitis) for three reasons: (1) It may have an avascular zone near its insertion, (2) it is placed under significant tension when the arm is elevated, and (3) it passes through a confined space above the humeral
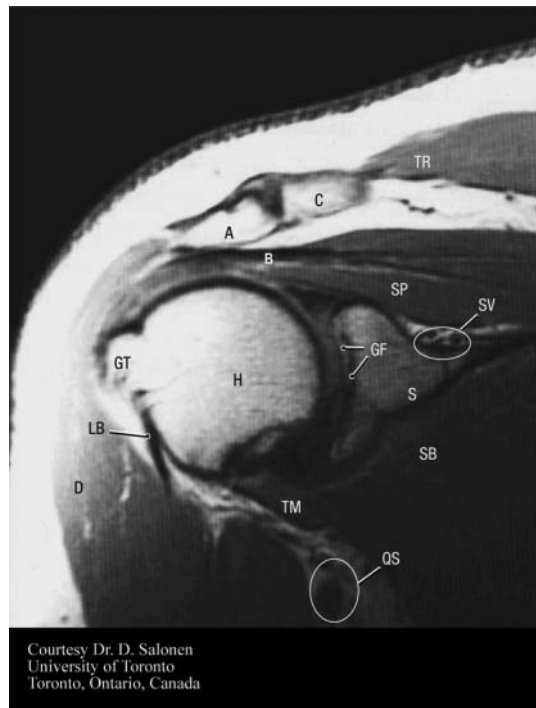
A    Acromnion
C    Clavicle
H    Head of humerus
GT   Greater tuberosity
     on humerus
TR   Trapezius muscle

SP   Supraspinatus muscle
D    Deltoid muscle
TM   Teres minor
LB   Tendon of the long head
     of the biceps muscle
B    Subdeltoid bursa

**Figure 3** Frontal plane view of the shoulder. (Modified from LifeART, 2000.)

head and below the acromion. Scapular anatomy also makes the supraspinatus tendon prone to compression between the humeral head and coracoacromial arch. The coracoacromial arch is formed by the acromion of the scapula and the coracoacromial ligament, which joins the corticoid process and acromion. The arch is a structure above the supraspinatus that can apply a compression force to the tendon if the humeral head migrates superiorly (Soslowsky et al., 1994). The subdeltoid and subacromial bursas are subject to compressive forces when the humerus is elevated, as is the tendon of the long head of the biceps. That tendon is also subject to frictional forces as it moves relative to the humerus within the bicipital groove, when the humerus is elevated.

Other structures within the shoulder complex are also important to the development of cumulative trauma in the proximal portion of the upper extremity. These include the brachial plexus (the anterior primary rami of the last four cervical spinal nerves and the first thoracic nerve, which go on to form the radial, median, and ulnar nerves in the upper extremity), the subclavian artery, and the subclavian vein. These structures all pass over the

first cervical rib, in close proximity to it. The artery and plexus pass in between the anterior and medial scalene muscles (which attach to the first cervical rib), and the vein lies anterior to the anterior scalene muscle, which separates the vein from the artery. The plexus, artery, and nerve all pass underneath the pectoralis minor muscle. These structures can be compressed by the muscles or bones in proximity to them when the humerus is elevated or when the shoulders are loaded directly (such as when wearing a backpack) or indirectly (as when holding a load in the hands). Additionally, the muscles of the shoulder, particularly trapezius, infraspinatus, supraspinatus, and levator scapula, are also common sites of pain and tenderness (Norregaard et al., 1998). The anterior view of the shoulder, from magnetic resonance imaging (MRI), is illustrated in Figure 3.

## 4  CAUSATION MODELS FOR DEVELOPMENT OF DISORDERS

### 4.1  Conceptual Model

Armstrong et al. (1993) developed a conceptual model for the pathogenesis of work-related musculoskeletal disorders which is not specific to any particular disorder. The model is based on the set of four cascading and interacting state variables of exposure, dose, capacity, and response, which are measures of the system state at any given time. The response at one level can act as the dose at the next level (see Figure 4). Furthermore, it is assumed that a response to one or more doses can diminish or increase the capacity for responding to successive doses. This conceptual model for development of WUEDs reflects the multifactorial nature of these disorders and the complex nature of the interactions among exposure, dose, capacity, and response variables. The model also reflects the complexity of interactions among the physiological, mechanical, individual, and psychosocial risk factors.

In the proposed model, exposure refers to the external factors (i.e., work, requirements) that produce the internal dose (i.e., tissue loads and metabolic demands and factors). Workplace organization and hand tool design characteristics are examples of such external factors, which can determine work postures and define loads on the affected tissues or the velocity of muscular contractions. Dose is defined by a set of mechanical, physiological, or psychological factors that in some way disturb an internal state of the affected worker. Mechanical disturbance factors may include tissue forces and deformations produced as a result of exertion or movement of the body. Physiological disturbances are such factors as consumption of metabolic substrates, or tissue damage, whereas psychological disturbance factors are those related to, for example, anxiety about work or inadequate social support.

Changes in the states of variables of the worker are defined as responses. A response is an effect of the dose caused by exposure. The model also allows for a given response to constitute a new dose, which then produces
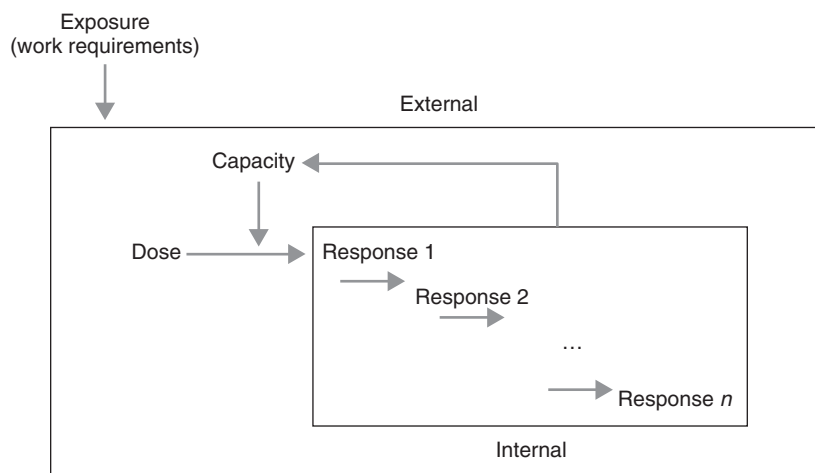
**Figure 4** Conceptual model for development of WMSDs proposed by Armstrong et al. (1993).

a secondary response (called the tertiary response). For example, hand exertion can cause elastic deformation of tendons and changes in tissue composition and/or shape, which in turn may result in hand discomfort (Armstrong et al., 1993). The dose–response time relationship implies that the effect of a dose can be immediate or the response may be delayed for a long period of time.

The model requires that system changes (responses) can also result in either increased dose tolerance (adaptation) or reduced dose tolerance lowering the system capacity. Capacity is defined as the worker's ability (physical or psychological) to resist system destabilization resulting from various doses. Whereas capacity can be reduced or enhanced by previous doses and responses, it is assumed that most people are able to adapt to certain types and levels of physical activity. Table 11 shows characterization of WMSDs with respect to exposure–dose relationship, the worker's capacity, and the response model proposed by Armstrong et al. (1993).

The main purpose of the dose–response model is to account for the factors and processes that result in WMSDs to specify acceptable limits with respect to work design parameters for a given person. The exposure, dose, response, and capacity variables need to be measured and quantified. Exposure can be measured using the job title or job classification, questionnaires on possible risk factors, job checklists, or direct measurements. Dose can be measured by estimating muscle forces and joint positions. Worker capacity can be measured using anthropometry, muscle strength, and psychological characteristics. The model proposed should be useful in the design of studies on the etiology and pathomechanisms of WMSDs. The model should also complement epidemiological studies that focus on associations between the physical workload, psychological demands, and environmental risk factors of work at one end and the manifestations of symptoms, diseases, or disabilities at the other.

## 4.2 Pathomechanical and Pathophysiological Models

Numerous epidemiological studies are consistent in their finding of statistically significant associations between workplace exposures to various physical risk factors and the incidence and/or prevalence of upper extremity MSDs in workers. Experimental studies on humans (in vivo) and investigations that utilize cadavers (in vitro) provide more direct evidence to support hypotheses regarding the internal responses to exposure doses of which Armstrong et al. (1993) wrote. However, the types of experimental studies that can establish direct causal links are those based on animal models, where the exposure dose and potential confounding factors can be strictly controlled and the effects measured over time to provide a view of the natural history of a disorder's progression. A number of reviews have been published recently which piece together information from these various types of studies in order to examine, from all sides, the patterns of evidence in support of workplace physical exposures as causes of musculoskeletal disorders in workers. These include NRC (2001) and Barr and Barbe (2002), which reviewed studies concerning the mechanobiology and pathophysiology of tendons, muscles, peripheral nerves, and other tissues that are involved in MSDs; Buckle and Devereux (2002), who conducted a similar review for the European Commission; Visser and van Dieen (2006), who focused on upper extremity muscle disorders; and Viikari-Juntura and Silverstein (1999), who focused on CTS. A sampling of key details and excerpts from their conclusions are provided herein. Readers are encouraged to read the full reviews and original research studies for a deeper appreciation of the strength of the evidence provided by these studies.

### 4.2.1 Tendons

Tendons are a complex composite material consisting of collagen fibrils embedded in a matrix of proteoglycans.

**Table 11　Characterization of Work-Related Musculoskeletal Disorders in General and Muscle, Tendon, and Nerve Disorders in Particular According to Sets of Cascading Exposure and Response Variables as Conceptualized in Model**

| Exposure Dose | Worker's Capacity | Response |
|---|---|---|
| Musculoskeletal system | | |
| Work load | Body size and shape | Joint position |
| Work location | Physiological state | Muscle force |
| Work frequency | Psychological state | Muscle length |
| | | Muscle velocity |
| | | Frequency |
| Muscle disorders | | |
| Muscle force | Muscle mass | Membrane permeability |
| Muscle velocity | Muscle anatomy | Ion flow |
| Frequency | Fiber type and composition | Membrane action potentials |
| Duration | Enzyme concentration | Energy turnover (metabolism), muscle |
| | Energy stores | enzymes, and energy stores |
| | Capillary density | Intramuscular pressure |
| | | Ion imbalances |
| | | Reduced substrates |
| | | Increased metabolites and water |
| | | Increase in blood pressure, heart rate, |
| | | cardiac output, muscle blood flow |
| | | Muscle fatigue |
| | | Pain |
| | | Free radicals |
| | | Membrane damage |
| | | Z-disk ruptures |
| | | Afferent activation |
| Tendon disorders | | |
| Muscle force | Anthropometry | Stress |
| Muscle length | Tendon anatomy | Strain (elastic and viscous) |
| Muscle velocity | Vascularity | Microruptures |
| Frequency | Synovial tissue | Necrosis |
| Joint position | | Inflammation |
| Compartment | | Fibrosis |
| pressure | | Adhesions |
| | | Swelling |
| | | Pain |
| Nerve disorders | | |
| Muscle force | Anthropometry | Stress |
| Muscle length | Nerve anatomy | Strain |
| Muscle velocity | Electrolyte status | Ruptures in perineural tissues |
| Frequency | Basal compartment pressure | Protein leakage |
| Joint position | | Ruptures in perineural tissue |
| Compartment | | Protein leakage in nerve trunks |
| pressure | | Edema |
| | | Increased pressure |
| | | Impaired blood flow |
| | | Numbness, tingling, conduction block |
| | | Nerve action potentials |

*Source:* Adapted from Armstrong et al. (1993).

A schematic representation of tendon structure is provided in Figure 5.

***Biomechanics***　Tendons transmit tensile loads between the muscles and the bones to which they are attached. However, they are also subjected to compressive and frictional/shear loads from adjacent structures (bone, other tendons, tendon sheaths, muscle, etc.).

Estimates of these loads have been made by modeling tendons as mechanical pulley systems (Armstrong and Chaffin, 1979). When loads on tendons are excessive (in magnitude, duration, repetition, or in some combination of these), damage is thought to occur. Moore (2002) presented a biomechanical model which incorporated the pulley tendon model to express the relative importance of load duration and repetition in the development of tendon entrapment at the dorsal wrist compartments. The most common of these is de Quervain's disease (stenosing tenosynovitis), which involves the tendons of the abductor pollicis longus and
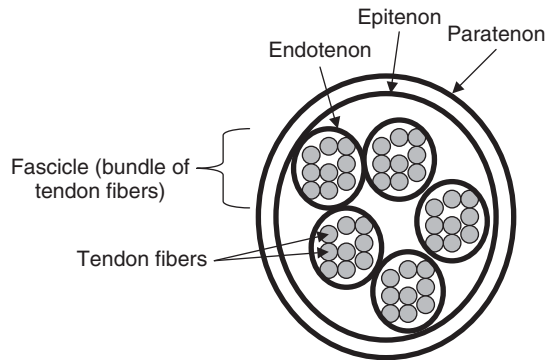
**Figure 5** Tendon structure. The paratenon is essentially doubled for those tendons that are surrounded by a synovial sheath (not shown in the figure). The mesotenon (not shown in the figure) are synovial layers that pass from the tendon to the wall of the sheath.

extensor pollicis brevis (two extrinsic thumb-moving muscles).

***Pathophysiology and Pathomechanics*** The NRC (2001) review of the mechanobiology of tendons concluded that "basic science studies support the conclusion that repetitive motion or overuse loading can cause chronic injury to tendon tissues." The review noted that fibrocartilaginous tissue is found within tendons, where they wrap around bone (e.g., the long head of biceps tendon within the bicipital groove of the humerus). The review also cited animal studies in which excessive repetitive loading was shown to result in degenerative changes to tendons and edema, increased numbers of capillaries, fibrosis, and other changes to the paratenon (the loose connective tissue that surrounds tendons that are not surrounded by a sheath). These changes are similar to those associated with peritendonitis and tendinosis in humans. Recently, Barbe et al. (2003) found evidence of tendon fraying at the muscle–tendon junction within the reaching limb of rats that performed a voluntary, low-force, repetitive reaching task for six to eight weeks. They also found progressive increases, over the course of the study, in the number of infiltrating macrophages in tendons and other tissues in both forward limbs but more so in the reaching limb. [Macrophages are large cells that possess the property of ingesting bacteria, foreign particles, and other cells.

Infiltrating macrophages are typically found at sites of inflammation. Inflammation is a fundamental pathological process that occurs in response to an injury or abnormal stimulation caused by a physical, chemical, or biological agent (McDonough, 1994).] Effects were also seen in tendons that were not directly related to the reaching task, indicating not just a local effect but also a systemic effect of the repetitive hand/paw-use intensive task.

### 4.3 Muscle

Muscle is a composite structure made up of muscle cells, organized networks of nerves and blood vessels, and extracellular connective matrix. Cells are fused together to form each muscle fiber, the basic structural element of skeletal muscle. Muscle is unique among the tissues in the body for its ability to contract in response to a stimulus from a motoneuron. A motor unit is made up of a single alpha motoneuron and the muscle fibers it innervates. There are essentially three types of fibers or motor units. All the fibers within a motor unit are of the same type. Type I fibers are small and are recruited first and at low levels of contraction. They have a high capillary density and are fatigue resistant. Type II fibers are larger and are recruited later, when more force is required. They have low capillary density and are the most fatiguable of the three types. Type IIA have a mix of properties of the other two types. Muscles usually contain all three types in varying proportions, based on the role of the muscle as well as the construction of the person.

***Biomechanics*** Muscle fiber damage can occur when external loads exceed the tolerance of the active contractile components and the passive connective tissue. Nonfatiguing muscle exertions require energy and oxygen, supplied by the blood. The flow of blood can be reduced when a muscle contracts if the intramuscular pressure increases beyond about 30 mm Hg (capillary closing pressure). Järvholm et al. (1988, 1991) demonstrated reductions in blood flow in rotator cuff muscles as a function of arm position (abduction or flexion) and weight held in the hand (0–2 kg). Even mild elevations, $30°$ of abduction combined with $45°$ of flexion, increased intramuscular pressure to 70 mm Hg. Contraction of the supraspinatus to only 10% of maximum capability increased intramuscular pressure to 50 mm Hg. Elevated intramuscular pressure may lead to localized muscle fatigue or more serious outcomes. Studies have shown that intramuscular pressure sustained at 30 mm Hg for 8 h resulted in muscle fiber atrophy, splitting, necrosis, and other damage (NRC, 2001).

***Pathophysiology and Pathomechanics*** The NRC (2001) review of the mechanobiology of skeletal muscle concluded that "the scientific studies reviewed support the conclusion that repetitive mechanical strain exceeding tolerance limits ... results in chronic skeletal muscle injury." In addition to finding changes in tendon, Barbe et al. (2003) also found infiltrating macrophages in the muscles of the reaching and nonreaching limb of test rats which performed the repetitive, low-force reaching task. Muscles in the paw and the distal forearm were affected, as were muscles that were involved in the task only indirectly (forearm extensors, upper arm and shoulder muscles) or not at all (tibial muscles). Heat-shock protein (HSP) cells increased, first in the intrinsic hand muscles and then in the distal forelimb flexor muscles. Heat-shock proteins have a protective role in the cell. Cells increase their production of HSPs when they experience acute or chronic stress. Precursors that stimulate HSP production include inflammation, ischemia, and nerve crush as well as other stimulating factors.

Visser and van Dieen (2006) reviewed several hypotheses concerning the pathogenesis of work-related

upper extremity muscle disorders and concluded that no complete proof existed in the literature for any of them but that some of them were likely to interact or follow one another in a downward spiral of damage. It appears that the selective and sustained motor unit recruitment (of small type I fibers) in combination with homeostatic disturbances possibly due to limitations in blood supply and metabolite removal offers a plausible basis for the pathogenesis of muscle disorder in low-intensity tasks. The bulk of the findings from the biopsy studies reviewed, which indicate mitochondrial dysfunction of type I fibers in myalgic muscles, could also be accounted for by such a mechanism. As a response to the release of metabolites in the muscle, the circulation increases. Sympathetic activation (stress) might lead to a reduction of circulation and an increase of muscle activation. Sustained exposure can result in an accumulation of metabolites, stimulating nociceptors. This process can be enhanced in subjects with relatively large type I fibers and low capillarization, which paradoxically may have developed as an adaptation to the exposure. Nociceptor activation can disturb the proprioception and thereby the motor control most likely leading to further increased disturbance of muscle homeostasis. In addition, in the long run a reduction of the pain threshold and an increase of pain sensitivity can develop. It is worth noting that initial nociceptor stimulation may be a response to metabolite accumulation and not to tissue damage.

### 4.3.1 Peripheral Nerves

Peripheral nerves are composed of nerve fibers, connective tissue, and blood vessels. A nerve fiber is a long process that extends from a nerve cell body. The term *nerve* refers to a bundle of axons, some of which send information from the spinal cord to the periphery (e.g., muscles, vessels) and some of which send information from peripheral tissues (e.g., skin, muscle, tendon) to the spinal cord. Figure 6 illustrates a single nerve cell (motor neuron) and a nerve.

***Biomechanics*** Peripheral nerves are well vascularized in order to supply energy needs for impulse transmission and axonal transport (transportation of nutrition and waste products within a nerve cell). If pressure is elevated within the nerve (due to edema or external compression), blood flow is reduced, and as a result, impulse transmission and axonal transport can be slowed or disrupted. Structural damage may also occur.

***Pathophysiology and Pathomechanics*** Numerous studies of nerve function have identified threshold levels of pressure ranging from 20 to 40 mm Hg, above which nerve function, including circulation, axonal transport, and impulse conduction, is compromised (Rydevik et al., 1981; Szabo and Gelberman, 1987). Damage to nerves comes in the form of loss of myelin (Mackinnon et al., 1984), large myelinated fibers (Hargens et al., 1979), and damage to small unmyelinated fibers (Hargens et al., 1979). The physiological and histological effects of pressure on a nerve depend on the amount of pressure, how it is applied (dispersed or focal), and duration of application. Vibration can also cause damage to peripheral nerves, including breakdown of myelin, interstitial and perineural fibrosis, and axonal loss, based on evidence from biopsies from humans exposed to vibrating hand tools and empirical animal studies (NRC, 2001).

### 4.3.2 Carpal Tunnel Syndrome

Carpal tunnel syndrome is the most commonly encountered peripheral neuropathy (Falkiner and Myers, 2002; Werner and Andary, 2002). The term describes a constellation of symptoms that result from localized compression of the median nerve within the carpal tunnel. The carpal tunnel is a fibro-osseous canal created by the carpal bones (floor and walls) and the flexor retinaculum (ceiling). The contents include the median nerve, the eight extrinsic flexor tendons of digits 2–5 [flexor digitorum superficialis and profundus (FDS and FDP), respectively], and tendons of the flexor pollicis longus (FPL) and flexor carpi radialis (FCR) (the latter is somewhat separated from the other contents, in its own subtunnel). Part of the volume of the carpal tunnel is also taken up by the synovial sheaths that surround the tendons.

The contents of the tunnel are not static. The nerve moves transversely within the tunnel, and relative to the
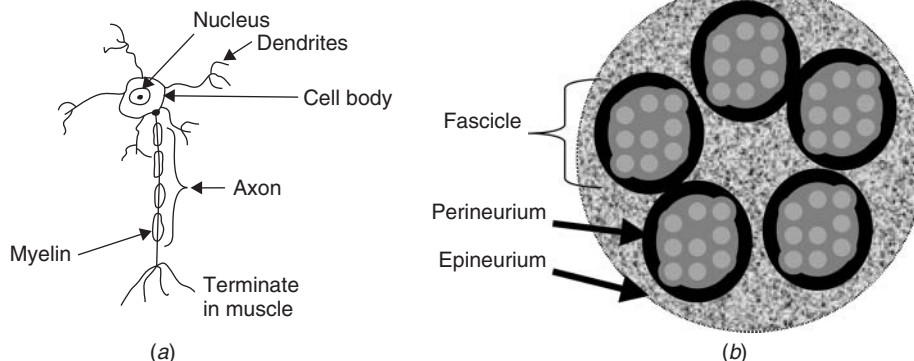


**Figure 6** (*a*) Neuron and (*b*) nerve.

tendons, when the extrinsic finger flexors are activated isometrically (Nakamichi and Tachibana, 1992), when the fingers are flexed or extended (Ham et al., 1996), and when the wrist is flexed or extended (Zeiss et al., 1989), and change in the nerve's location seems to be somewhat variable from one person to the next. The extrinsic finger flexor tendons slide proximally with finger flexion and distally with finger extension. Further, in some people lumbrical muscles may enter the tunnel during pinching (Ditmars, 1993) and appear consistently to enter the tunnel with full finger flexion (Siegel et al., 1995; Ham et al., 1996). Distal fibers of the FDS and FDP may enter during wrist extension (Keir and Bach, 2000). The cross-sectional area of the tunnel also changes, being smaller when the fingers are fully extended than when flexed (Ham et al., 1996). The area of the tunnel is reduced in wrist flexion and extension compared with a neutral posture (Keir, 2001).

***Biomechanics*** Soft tissues appear to be subjected to mechanical stress within the carpal tunnel. The median nerve takes on an oval or somewhat flattened shape within the carpal tunnel (Robbins, 1963; Zeiss et al., 1989). In a study of cadaver hands (donor health history unknown), Armstrong et al. (1984) observed increased subsynovial and adjacent connective tissue densities, increased synovial hyperplasia, and arteriole and venule muscular hypertrophy, with changes being most pronounced near the distal wrist crease. The authors suggested that the types of changes similar to those seen in CTS biopsy specimens and the extent to which their severity corresponded to their proximity to the distal wrist crease indicated that repeated flexion and extension at the wrist imposes mechanical stress on the tissues that cross that joint and that their alterations were in direct response to that stress. They also suggested that highly repetitive use of the wrist might bring about alterations that would be severe enough to elicit CTS symptoms. In patients undergoing CTS release surgery, Schuind et al. (1990) found fibrous hyperplasia of the flexor tendon synovium and increased amounts of collagen fibers (irregular and disorganized). More advanced lesions contained necrotic areas as well. The authors concluded that these histological lesions were "typical of a connective tissue undergoing degeneration under repeated mechanical stresses."

***Pathophysiology and Pathomechanics*** The median nerve can be damaged by direct force (damaging myelin and other structural components) and by elevated hydrostatic pressure (ischemic response). Schuind et al. (1990) hypothesized that, following an initial mechanical stress upon the flexor tendon synovium, a vicious cycle develops in which changes in synovium increase frictional loads on the tendons as they move back and forth through the tunnel, which causes further irritation to the synovium. Effects on the median nerve are increased pressure due to a reduction of free volume within the carpal tunnel and possibly restrictions on the nerve's freedom to move within the tunnel during wrist movement (contact stress). This is consistent with the cascade model of work-related upper limb musculoskeletal disorders proposed by Armstrong et al. (1993). Another cascade model theorizes that elevated pressure within the carpal tunnel leads to ischemia in the tendons, sheaths, and median nerve. This is followed by tissue swelling, which further increases the pressure in the tunnel and can result in physiological and histological changes in the median nerve (Lluch, 1992).

Studies of nerve function have identified threshold levels of pressure ranging from 20 to 40 mm Hg, above which nerve function, including circulation, axonal transport, and impulse conduction, is compromised (Rydevik et al., 1981; Szabo and Gelberman, 1987). Short-term laboratory-based studies have also shown that carpal tunnel pressure (CTP) in healthy people exceeds these levels with nonneutral wrist postures (Keir et al., 1998b) or in using the hand in ordinary ways, such as pressing with the fingertip or pinching (Keir et al., 1998a) or typing (Sommerich et al., 1996). Even with hands inactive and wrists in a neutral posture, CTP is elevated in people with CTS (Rojviroj et al., 1990). Cyclic movement of the wrist was shown to induce sustained, elevated CTP in patients with early or intermediate CTS (Szabo and Chidgey, 1989).

Viikari-Juntura and Silverstein (1999) examined the pattern of evidence of the role of physical factors in the development of CTS by reviewing studies from various areas: epidemiological, experimental, cadaver, and animal. One of the consistent and unifying threads they found was the connection (manifestation) of these external, physical factors (posture, force, repetition, and external pressure) as CTP or as having an effect on CTP. They concluded that there is sufficient evidence that duration, frequency, or intensity of exposure to forceful repetitive work and extreme wrist postures is likely to be related to the occurrence of CTS in working populations. Recent work by Clark et al. (2003, 2004) provides additional insight into the effects of performing highly repetitive tasks in rats performing voluntary hand/paw use-intensive tasks. Effects of a low-force, repetitive reaching task included increased numbers of infiltrating macrophages (a sign of inflammation) in the median nerve, with a greater increase in the reach than in the nonreach limb, and myelin degradation and fibrosis, particularly in the epineurium at the wrist and just distal to the carpal ligament (associated with increased compression) (Clark et al., 2003). Effects of a high-force, repetitive grasping task included infiltrating macrophages in all connective tissue associated with the nerve and increased collagen type I (fibrosis) in perineurium, epineurium, and surrounding tissues. For both of these effects, there was no difference between the reach and nonreach limbs (Clark et al., 2004). These bilateral effects are important, because bilateral presentation of CTS in workers is taken by some to be an indication that work is not the cause of the condition.

## 4.4 Causal Models Summary

Evidence of the role of physical factors in the development of upper extremity MSDs from epidemiological, experimental, cadaver, and animal studies seems to support the conceptual model of Armstrong et al. (1993).

Additional animal models, to replicate and extend the work of Clark and colleagues, will serve to clarify and begin to quantify the relationships between external dose and internal response. Eventually, such work will lead to quantitative, preventive models that can be used to reduce risk of WUED in human workers.

## 5 QUANTITATIVE MODELS FOR CONTROL OF DISORDERS

### 5.1 Challenges

Today, only a few quantitative models that are based on the physiological, biomechanical, or psychophysical data and that relate the specific job risk factors for musculoskeletal disorders to increased risk of developing such disorders have been developed. As discussed by Moore and Garg (1995), this is mainly because (1) the dose–response (cause–effect) relationships are not well understood; (2) measurement of some task variables, such as force and even posture, is difficult in an industrial setting; and (3) the number of task variables is very large. However, it is generally recognized that biomechanical risk factors such as force, repetition, posture, recovery time, duration of exposure, static muscular work, use of the band as a tool, and type of grasp are important for explaining the causation mechanism of WUEDs (Armstrong and Lifshiz, 1987; Keyserling et al., 1993). Given the foregoing knowledge, even though limited in scope and subject to epidemiological validation, a few methodologies that allow discrimination between safe and hazardous jobs in terms of workers being at increased risk of developing the WUEDs have been developed and reported in the subject literature. Some of the quantitative data and models for evaluation and prevention of WUEDs available today are described below.

### 5.2 Maximum Acceptable Forces for Repetitive Wrist Motions for Females

Snook et al. (1995) utilized the psychophysical methodology to determine the maximum acceptable forces for various types and frequencies of repetitive wrist motion, including (1) flexion motion with a power grip (handle diameter 40 mm, handle length 135 mm); (2) flexion motion with a pinch grip (handle thickness 5 mm, handle length 55 mm); and (3) extension motion with a power grip (handle diameter 40 mm, handle length 135 mm). Subjects were instructed to work as if they were on an incentive basis, getting paid for the amount of work that they performed. They were asked to work as hard as they could (i.e., against as much resistance as they could) without developing unusual discomfort in the hands, wrists, or forearms.

Fifteen women worked 7 h each day, 2 days per week, for 40 days in the first experiment. Repetition rates of 2, 5, 10, 15, and 20 motions per minute were used with each flexion and extension task. Maximum acceptable torque was determined for the various motions, grips, and repetition rates without dramatic changes in wrist strength, tactile sensitivity, or number of symptoms. Fourteen women worked in the second experiment, performing a wrist flexion motion

(power grip) 15 times per minute, 7 h per day, 5 days per week, for 23 days. In addition to the four dependent variables—maximum acceptable torque, maximum isometric wrist strength, tactile sensitivity, and symptoms—performance errors and duration of force were measured. The most common health symptom reported was muscle soreness (55.3%), located mostly in the hand and wrist (51.8%). Numbness in the palmar side of the fingers and thumb (69.1%) and stiffness on the dorsal (back) side of the fingers and thumb (30.6%) were also reported. The number of symptoms increased consistently as the day progressed. The number of symptoms reported was two to three times higher after the seventh hour of work than after the first hour of work (similar to the two to four times higher rate in the two-days-per-week exposure). Symptoms reports were 4.1 times higher at the end of the day than at the beginning of the day before testing began.

The maximum acceptable torque determined during the five-days-per-week exposure was 36% lower than the task performed only two days per week. Based on the assumption that maximum acceptable torque decreases 36.3% for the other repetition rates used during the two-days-per-week exposure, and using the adjusted means and coefficients of variation from the two-days-per-week exposure, the maximum acceptable torques were estimated for different repetitions of wrist flexion (power grip) and different percentages of the population. Torques were then converted into forces by dividing each torque by the average length of the handle lever (0.081 m). The estimated maximum acceptable forces for female wrist flexion (power grip) are shown in Table 12. Tables 13 and 14 show the estimated maximum acceptable forces for female wrist flexion (pinch grip) and wrist extension (power grip), respectively. The torques were converted into forces by dividing by 0.081 m for the power grip and 0.123 m for the pinch grip.

### 5.3 Hand Activity Level

In 2002 the American Congress of Governmental Industrial Hygienists (ACGIH) adopted the hand activity level (HAL) within the section on ergonomics in its annual publication of threshold limit values (TLVs). This annual publication contains recommendations and guidelines "to assist in the control of potential workplace health hazards" (ACGIH, 2002). The TLV considers the dual exposures of average hand activity level and peak hand force for monotask jobs performed for four

**Table 12 Maximum Acceptable Forces for Female Wrist Flexion (Power Grip) (N)**

| Percentage of Population | Repetition Rate | | | | |
|---|---|---|---|---|---|
| | 2/min | 5/min | 10/min | 15/min | 20/min |
| 90 | 14.9 | 14.9 | 13.5 | 12.0 | 10.2 |
| 75 | 23.2 | 23.2 | 20.9 | 18.6 | 15.8 |
| 50 | 32.3 | 32.3 | 29.0 | 26.0 | 22.1 |
| 25 | 41.5 | 41.5 | 37.2 | 33.5 | 28.4 |
| 10 | 49.8 | 49.8 | 44.6 | 40.1 | 34.0 |

*Source:* Adapted from Snook et al. (1995).

**Table 13  Maximum Acceptable Forces for Female Wrist Flexion (Pinch Grip) (N)**

| Percentage of Population | Repetition Rate | | | | |
|---|---|---|---|---|---|
| | 2/min | 5/min | 10/min | 15/min | 20/min |
| 90 | 9.2 | 8.5 | 7.4 | 7.4 | 6.0 |
| 75 | 14.2 | 13.2 | 11.5 | 11.5 | 9.3 |
| 50 | 19.8 | 18.4 | 16.0 | 16.0 | 12.9 |
| 25 | 25.4 | 23.6 | 20.6 | 20.6 | 16.6 |
| 10 | 30.5 | 28.3 | 24.6 | 24.6 | 19.8 |

*Source:* Adapted from Snook et al. (1995).

**Table 14  Maximum Acceptable Forces for Female Wrist Extension (Power Grip) (N)**

| Percentage of Population | Repetition Rate | | | | |
|---|---|---|---|---|---|
| | 2/min | 5/min | 10/min | 15/min | 20/min |
| 90 | 8.8 | 8.8 | 7.8 | 6.9 | 5.4 |
| 75 | 13.6 | 13.6 | 12.1 | 10.9 | 8.5 |
| 50 | 18.9 | 18.9 | 16.8 | 15.1 | 11.9 |
| 25 | 24.2 | 24.2 | 21.5 | 19.3 | 15.2 |
| 10 | 29.0 | 29.0 | 25.8 | 23.2 | 18.3 |

*Source:* Adapted from Moore and Garg (1995).

or more hours per day. Peak force is normalized to the strength capability of the workforce for the activity being assessed. Combinations of force and activity either fall within an acceptable range (below the action limit line), which means that nearly all workers could be exposed repeatedly without adverse health effects, a midrange that is delineated by the action limit line below and the threshold limit line above in which some workers might be at risk, and a region that exceeds the TLV in which jobs with similar characteristics have been shown to be associated with elevated risk of upper extremity MSDs (UEMSDs). The TLV does not specifically account for contact stress, nonneutral postures, cold, gloves, or vibration, and the ACGIH urges use of professional judgment when modifying factors such as these are also present in the task. More specific information about the HAL can be found at http://www.aegih.org/ and other locations on the Web.

### 5.4  Strain Index

#### 5.4.1  Model Structure

Moore and Garg (1995) developed a self-styled semi-quantitative job analysis methodology for identifying industrial jobs associated with distal upper extremity disorders (elbow, forearm, wrist, and hand). An existing body of knowledge and theory of the physiology, biomechanics, and epidemiology of distal upper extremity disorders was used for that purpose. The following major principles were derived from the physiological model of localized muscle fatigue:

1.  The primary task variables are intensity of exertion, duration of exertion, and duration of recovery.

2.  Intensity of exertion refers to the force required to perform a task one time and is characterized as a percentage of maximal strength.

3.  Duration of exertion describes how long an exertion is applied. The sum of duration of exertion and duration of recovery is the cycle time of one exertional cycle.

4.  Wrist posture, type of grasp, and speed of work are considered by means of their effects of maximal strength.

5.  The relationship between strain on the body (endurance time) and intensity of exertion is nonlinear.

The following major principles derived from the biomechanical model of the viscoelastic properties of components of a muscle–tendon unit were utilized for the model:

1.  The primary task variables for the viscoelastic properties are intensity and duration of exertion, duration of recovery, number of exertions, wrist posture, and speed of work.

2.  The primary task variables for intrinsic compression are intensity of exertion and nonneutral wrist posture.

3.  The relationship between strain on the body and intensity of effort is nonlinear.

Finally, the major principles derived from the epidemiological literature and used for the purpose of model development were as follows:

1.  The primary task variables associated with an increased prevalence or incidence of distal upper extremity disorders are intensity of exertion (force), repetition rate, and percentage of recovery time per cycle.

2.  Intensity of exertion is the most important task variable related to disorders of the muscle–tendon unit.

3.  Wrist posture may not be an independent risk factor because it may contribute to an increased incidence of distal upper extremity disorders when combined with intensity of exertion.

4.  The roles of other task variables have not been clearly established epidemiologically.

Moore and Garg (1995) compared exposure factors for jobs associated with WUEDs to jobs without prevalence of such disorders. They found that the intensity of exertion, estimated as a percentage of maximal strength and adjusted for wrist posture and speed of work, was the major discriminating factor. The relationship between the incidence rate for distal upper extremity disorder and job risk factors was defined as

$$IR = (30 \times F^2)/RT^{0.6} \qquad (1)$$

where IR is the incidence rate (per 100 workers per year), $F$ the intensity of exertion (% of maximum strength), and RT the recovery time (% of cycle time).

### 5.4.2 Elements of Strain Index

The strain index (Sl) proposed by Moore and Garg (1995) is the product of six multipliers that correspond to six task variables: (1) intensity of exertion, (2) duration of exertion, (3) exertions per minute, (4) hand−wrist posture, (5) speed of work, and (6) duration of task per day. An ordinal rating is assigned for each of the variables according to the exposure data. The ratings that are applied to model variables are presented in Table 15. The multipliers for each task variable related to these ratings are shown in Table 16. The strain index score is defined as follows:

$$
\begin{aligned}
\text{SI} = \ & (\text{intensity of exertion multiplier}) \\
& \times (\text{duration of exertion multiplier}) \\
& \times (\text{exertions per minute multiplier}) \\
& \times (\text{posture multiplier}) \\
& \times (\text{speed of work multiplier}) \\
& \times (\text{duration per day multiplier}) \quad (2)
\end{aligned}
$$

*Intensity of exertion*, the most critical variable of the SI, is defined as the percentage of maximum strength required to perform a task once. The intensity of exertion is estimated by an observer using verbal descriptors (see Table 15) and assigned corresponding rating values (1, 2, 3, 4, or 5). The multiplier values (Table 16) are defined based on the rating score raised to a power of 1.6 to reflect the nonlinear nature of the relationship between intensity of exertion and manifestations of strain according to the psychophysical theory. The multipliers for other task variables are modifiers to the intensity of the exertion multiplier.

*Duration of exertion* is defined as the percentage of time that an exertion is applied per cycle. The terms *cycle* and *cycle time* refer to the exertional cycle and average exertional cycle time, respectively. The duration of recovery per cycle is equal to the exertional cycle time minus the duration of exertion per cycle. The duration of exertion is the average duration of exertion per exertional cycle (calculated by dividing all durations of a series of exertions by the number of exertions observed). The percentage duration of exertion is calculated by dividing the average duration of exertion per cycle by the average exertional cycle time, then multiplying the result by 100:

%duration of exertion

$$
= (\text{average duration of exertion per cycle/average}
$$

$$
\text{exertional cycle time}) \times 100 \quad (3)
$$

The percentage duration of exertion calculated is compared to the ranges in Table 15 and assigned the appropriate rating. The corresponding multipliers are identified using Table 16.

*Effort per minute* is the number of exertions per minute (e.g., repetitiveness) and is synonymous with frequency. Efforts per minute are measured by counting the number of exertions that occur during a representative observation period (as described for determining the average exertional cycle time). The results measured are compared to the ranges shown in Table 15 and given the corresponding ratings. The multipliers are defined in Table 16.

*Posture* refers to the anatomical position of the wrist or hand relative to the neutral position and

**Table 15  Rating Criteria for Strain Index**

| Rating | Intensity of Exertion | Duration of Exertion (% of Cycle) | Efforts per Minute | Hand−Wrist Posture | Speed of Work | Duration per Day (h) |
|---|---|---|---|---|---|---|
| 1 | Light | <10 | <4 | Very good | Very slow | ≤1 |
| 2 | Somewhat hard | 10−29 | 4−8 | Good | Slow | 1−2 |
| 3 | Hard | 30−49 | 9−14 | Fair | Fair | 2−4 |
| 4 | Very hard | 50−79 | 15−19 | Bad | Fast | 4−8 |
| 5 | Near maximal | ≥80 | ≥20 | Very Bad | Very fast | ≥8 |

*Source:* Adapted from Moore and Garg (1995).

**Table 16  Multipliers for Strain Index**

| Rating | Intensity of Exertion | Duration of Exertion (% of Cycle) | Efforts per Minute | Hand−Wrist Posture | Speed of Work | Duration per Day (h) |
|---|---|---|---|---|---|---|
| 1 | 1 | 0.5 | 0.5 | 1.0 | 1.0 | 0.25 |
| 2 | 3 | 1.0 | 1.0 | 1.0 | 1.0 | 0.50 |
| 3 | 6 | 1.5 | 1.5 | 1.5 | 1.0 | 0.75 |
| 4 | 9 | 2.0 | 2.0 | 2.0 | 1.5 | 1.00 |
| 5 | 13 | 3.0[a] | 3.0 | 3.0 | 2.0 | 1.50 |

*Source:* Adapted from Moore and Garg (1995).
[a] If duration of exertion is 100%, the efforts/minute multiplier should be set to 3.0.

can be rated qualitatively using verbal anchors. As shown in Table 16, posture has four relevant ratings. Postures that are "very good" or "good" are essentially neutral and have multipliers of 1.0. As hand or wrist postures progressively deviate beyond the neutral range to extremes, they are graded as "fair," bad," and "very bad."

*Speed of work* refers to the perceived pace of the task or job and can be estimated subjectively. Once a verbal anchor is selected, a rating is assigned according to Table 15. Duration of task per day is defined as the total time that a task is performed per day. As such, this variable reflects the beneficial effects of task diversity such as job rotations and the adverse effects of prolonged activity such as overtime. Duration of task per day is measured in hours and assigned a rating according to Table 15.

Application of the SI involves five steps: (1) collecting data, (2) assigning rating values, (3) determining multipliers, (4) calculating the SI score, and (5) interpreting the results. The values of intensity of exertion, wrist posture, and speed of work can be estimated using the verbal descriptors in Table 15. The values of percentage duration of exertion per cycle, efforts per minute, and duration per day are based on measurements and counts. These values are then compared to the appropriate column in Table 16 and assigned a rating. The SI multipliers are determined from Table 16. Table 17 shows the numerical example for calculating the strain index.

### 5.4.3 Application

In a preliminary test of the ability of the SI to distinguish between jobs that were high or low risk for distal UEMSDs, the SI was found to have a sensitivity of 0.92 (able to identify correctly 11 of 12 known positive jobs) and a specificity of 1.0 (able to identify correctly 13 of 13 known negative jobs) when 25 jobs in a pork processing facility were assessed. Three different studies conducted by Moore and colleagues have shown that the SI is capable of distinguishing between safe and hazardous jobs for upper extremity disorders with an odds ratio (OR) of 114 (95% confidence interval: 24–545), sensitivity of 0.90, specificity of 0.93, positive predictive value of 0.93 and negative predictive value of 0.91 (Moore and Garg, 1995; Knox and Moore, 2001; Moore et al., 2001, 2006). Findings from these studies suggested using an SI score of 5 as a threshold to distinguish between safe and hazardous jobs. Rucker and Moore (2002) performed another assessment of the SI

on 28 jobs from two different manufacturing companies. The sensitivity, specificity, positive predictive value, and negative predictive value were 1.00, 0.91, 0.75, and 1.00, respectively, providing additional evidence of the validity of the tool for assessment of single tasks. Other recent studies utilized the SI method to examine its reliability and validity (Spielholz et al., 2008; Stephens et al., 2006), its possible application on cumulative-trauma reduction (Smith, 2007) and in multiple task jobs assessment (Bao et al., 2009)

### 5.4.4 Limitations

The proposed SI methodology aims to discriminate between jobs that expose workers to risk factors (task variables) that cause WUEDs and jobs that do not. However, according to Moore and Garg (1995), the SI is not designed to identify jobs associated with an increased risk of any specific disorder. It is anticipated that jobs identified by the SI to be in the high-risk category will exhibit higher levels of WUEDs among workers who currently perform or historically performed those jobs believed to be hazardous. Finally, the authors caution that large-scale studies are needed to validate and update the methodology proposed.

The SI has the following primary limitations in terms of its application:

1. It is designed primarily to predict distal upper extremity disorders involving muscle–tendon units and CTS rather than all UEMSDs [e.g., hand–arm vibration syndrome (HAYS), ganglion cysts. osteoarthritis].

2. The SI has not been developed to predict disorders beyond the distal upper extremity, such as disorders of the shoulder, shoulder girdle, neck, or back.

3. No method has been developed for using the SI to assess multiple tasks.

## 6 ERGONOMICS EFFORTS TO CONTROL DISORDERS

### 6.1 Strategies for Prevention of Musculoskeletal Injuries

Facing the growing challenges of musculoskeletal injuries in the contemporary workplace, *Proposed National Strategies for the Prevention of Leading Work-Related Diseases and Injuries* (NIOSH, 1986) identified environmental hazards and human biological hazards among the four main factors that contribute to

**Table 17 Example to Demonstrate Procedure for Calculating the Score**

|  | Intensity of Exertion | Duration of Exertion | Efforts per Minute | Posture | Speed of work | Duration perDay (h) |
|---|---|---|---|---|---|---|
| Exposure dose | Somewhat hard | 60% | 12 | Fair | Fair | 4–8 |
| Rating | 2 | 4 | 3 | 3 | 3 | 4 |
| Multiplier | 3.0 | 2.0 | 1.5 | 1.5 | 1.0 | 1.0 |
|  | SI Score = 3.0 × 2.0 × 1.5 × l.5 × l.0 × 1.0 = 13.5 | | | | | |

*Source:* Adapted from Moore and Garg (1995).

human diseases. *Environmental hazards* to the muscu-loskeletal system associated with work were described as workplace traumatogens (i.e., a source of biomechan-ical stress from job demands that exceed the worker's strength or endurance, such as heavy lifting or repeti-tive and forceful manual exertions). Traumatogens can be measured by determining the frequency, magnitude, and direction of forces imposed on the body in rela-tion to posture and the point of application. *Hunan biological factors* include the anthropometric or innate attributes that influence a worker's capacity for perform-ing a job safely. Examples include the worker's physical size, strength, range of motion, and work endurance. These factors account partially for variability in per-formance capability in the population and the poten-tial for a mismatch between the worker and job that can be addressed by applying ergonomic principles of work design. To reduce the extent of work-related mus-culoskeletal injuries, progress in four methodological areas was expected (NIOSH, 1986): (1) identifying the biomechanical hazards accurately, (2) developing effec-tive health promotion and hazard control interventions, (3) changing management concepts and operational poli-cies with respect to expected work performance, and (4) devising strategies for disseminating knowledge on con-trol technology and promoting their application through incentives.

With the issuance of *National Occupational Research Agenda*, the NIOSH (1996) built upon these pre-vious strategies by including musculoskeletal disor-ders and several related areas (e.g., organization of work, indoor environment, special populations at risk, exposure assessment methods, intervention effective-ness research, risk assessment methods, and surveillance research methods) in its list of 21 declared priority areas, defined as areas with the highest likelihood of reducing workplace injuries/illnesses. Consistent with NIOSH's vision are the recommendations from the NRC (2001), which included encouraging the institution or extension of ergonomic and other preventive, science-based strategies and identifying areas that need fur-ther research (improved tools for exposure assessment, improved measures of outcome and case definition for use in epidemiological and intervention studies, further quantification of exposure–outcome relationships).

## 6.2 Applying Ergonomic Principles and Processes

Ergonomic job design (and redesign) efforts focus on fitting characteristics of the job to capabilities of workers. In simple terms, this can be accomplished, for example, by reducing excessive strength requirements and exposure to vibration, improving the design of hand tools and work layouts, designing out unnatural postures at work, or addressing the problem of work–rest requirements for jobs with high production rates. From the occupational safety and health perspective, the current state of ergonomics knowledge should allow for management of WUEDs in order to minimize human suffering, potential for disability, and the related worker compensation costs. Application of ergonomics can help to (1) identify working conditions under which the

WUEDs might occur, (2) develop engineering design measures aimed at elimination or reduction of the known job risk factors, and (3) identify the affected worker population and target it for early medical and work intervention efforts.

Workplace and work design–related risk factors, which often overlap, typically involve a combina-tion of poor work methods, inadequate workstations and hand tools, and high production demands. A risk factor is defined as an attribute or exposure that increases the probability of the disease or disorder (Putz-Anderson, 1993). As discussed before, the biome-chanical risk factors for WUEDs include repetitive and sustained exertions, awkward postures, and application of high mechanical forces. Vibration and cold environ-ments may also accelerate the development of WUEDs. Tools that can be used to identify the potential for development of WUEDs include plant walk-throughs and/or more detailed work–methods analyses. Check-lists, analytical tools [such as the SI (Moore and Garg, 1995) or HAL (ACGIH, 2002)], and/or expertise of the analyst are utilized to identify undesirable work site conditions or worker activities that can contribute to injury.

Since job redesign decisions may require some design trade-offs (Putz-Anderson, 1993), the ergonomic intervention process should follow these steps: (1) per-form a thorough job analysis to determine the nature of specific problems, making sure to identify the root causes for the problem; (2) evaluate and select the most appropriate intervention(s), based on the root cause and other factors relevant to the particular circumstance; (3) develop and apply conservative treatment (imple-ment the intervention) on a limited scale if possible; (4) monitor progress to ensure that the intervention has the intended effect and no adverse consequences; and (5) adjust or refine the intervention as needed.

## 6.3 Administrative and Engineering Controls

The control of WUEDs requires consideration of the following aspects of this complex problem: (1) WUED diagnosis, (2) treatment, (3) rehabilitation and return to work, (4) WUED surveillance, (5) surveillance and control of risk factors at the micro- and macrolevels, (6) training and education, and (7) management and leadership with regard to WUED-related organizational and social aspects (Hagberg et al., 1995). The spe-cific recommendations for prevention of WUEDs can be classified as being either primarily administrative (i.e., focusing on personnel solutions) or engineering (i.e., focusing on redesigning tools, workstations, and jobs) (Putz-Anderson, 1993). In general, administrative con-trols are those actions taken by the management that are intended to limit the potentially harmful effects of a physically stressful job on individual workers. Admin-istrative controls, which are focused on the workers, are modifications of existing personnel functions such as worker training, job rotation, and matching employees to job assignments. A summary of selected ergonomics measures that aim to control the incidence of WUEDs is given in Table 18.

**Table 18 Ergonomic Measures to Control Common WMSDs**

| Disorder | Avoid in General | Avoid in Particular | Recommendation | Design Considerations |
|---|---|---|---|---|
| Carpal tunnel syndrome | Rapid, often repeated finger movements, wrist deviation | Dorsal and palmer flexion, pinch grip, vibrations between 10 and 60 Hz | | Workplace design |
| Cubital tunnel syndrome | Resting forearm on sharp edge or hard surface | | | Workplace design |
| DeQuervain's syndrome | Combined forceful gripping and hard twisting | | | Workplace design |
| Epicondylitis | ''Bad tennis backhand'' | Dorsiflexion, pronation | | Workplace design |
| Pronator syndrome | Forearm pronation | Rapid and forceful pronation, strong elbow and wrist flexion | Use large muscles but infrequently and for short time | Design of work object |
| Shoulder tendonitis, rotator cuff syndrome | Arm elevation | Arm abduction, elbow elevation | Let wrists be in linewith forearm | Design of job task |
| Tendonitis | Often repeated movements, particularly with force exertion; hard surface in contact with skin, vibrations | Frequent motions of digits, wrists, forearm shoulder | Let shoulder and upper arm be relaxed | Design of hand tools (''bend tool, not the wrist'') |
| Tenosynovitis, DeQuervain's syndrome, ganglion | Finger flexion, wrist deviation | Ulnar deviation, dorsal and palmar flexion, radial deviation with firm grip | Let forearms be horizontal or more declined | Design for round corners, use pad |
| Thoracic outlet syndrome | Arm elevation, carrying loads | Shoulder flexion, arm hyperextension | | Design of work object placement |
| Trigger finger or thumb | Digit flexion | Flexion of distal phalanx alone | | Workplace design |
| Ulnar artery aneurysm | Pounding and pushing with heel of the hand | | | Workplace design |
| Ulnar nerve entrapment | Wrist flexion and extension | Wrist flexion and extension, pressure of hypothenar eminence | | Workplace design |
| White finger, vibration syndrome | Vibrations, tight grip, cold exposure | Vibrations between 40 and 125 Hz | | Workplace design |
| Neck tension syndrome | Static head posture | Prolonged static head–neck posture | Alternative head–neck postures | Workplace design |

*Source:* Adapted from Kroemer et al. (1994).

With respect to biomechanical risk factors, prevention and control efforts for WUEDs should be directed toward fulfilling several recommendations based on ergonomics principles for workplace design, work methods, and work organization. As discussed by Putz-Anderson (1993), these may include, for example, the following recommendations: (1) permit several different working postures; (2) place controls, tools, and materials between waist and shoulder heights for ease of reach and operation; (3) use jigs and fixtures for holding purposes; (4) resequence jobs to reduce the repetition; (5) automate highly repetitive operations; (6) allow self-pacing

of work whenever feasible; and (7) allow frequent (voluntary and mandatory) rest breaks.

Furthermore, with respect to hand tools used at work, the following general work design guidelines are provided:

1.  Make sure that the center of gravity of the tool is located close to the body and the tool is balanced.
2.  Use power tools to reduce the force and repetition required.
3.  Consider redesigning the straight tool handle; bend it as necessary to preserve the neutral posture of the wrist.
4.  Use tools with pistol grip or straight grips, respectively, where the axis in use is horizontal or vertical (or when the direction of force is perpendicular to the workplace).
5.  Avoid tools that require working with a flexed wrist and extended arm at the same time or tools that call for the flexion of distal phalanges (last joints) of the fingers.
6.  Minimize the tool weight; suspend all tools heavier than 20 N (or 2 kg) of force by a counterbalancing harness.
7.  Align the tool's center of gravity with the center of the grasping hand.
8.  Use special-purpose tools that facilitate fitting the task to the worker (avoid standard off-the-shelf tools for specific repetitive operations).
9.  Design tools so that workers can use them with either hand.
10. Use a power grip where power is needed and a precision grip for precise tasks.
11. The handles and grips should be cylindrical or oval with a diameter between 3.0 and 4.5 cm (for precise operations the recommended diameter is from 0.5 to 1.2 cm).
12. The minimum handle length should be 10.0 cm, whereas a 11.5–2.0-cm handle is preferable.
13. A handle span of 7.5–8.0 cm can be used by male and female workers for plier-type handles.
14. Triggers on power tools should be at least 5.1 cm wide, allowing their activation by two or three fingers.
15. Avoid form-fitting handles that cannot be adjusted easily.
16. Provide handles that are nonporous, nonslip, and nonconductive (thermally and electrically).

## 6.4  Ergonomics Programs for Prevention

An important component of WUED management efforts is development of well-structured and comprehensive ergonomics programs. According to Alexander and Orr (1992), the basic components of such a program should include the following: (1) health and risk factor surveillance, (2) job analysis and improvement, (3) medical management, (4) training, and (5) program evaluation. An excellent program should include participation of all levels of management; medical, safety, and health personnel; labor unions; engineering; facility planners; and workers and contain the following elements:

1.  Routine (monthly or quarterly) reviews of the OSHA log (injury records) for patterns of injury and illness (dedicated computer programs can be used to identify problem areas)
2.  Workplace audits for ergonomic problems that are a routine part of the organization's culture (more than one audit annually for each operating area) and timely interventions as a response to the problems identified
3.  A knowledge by management and workers regarding the list of most critical problems (i.e., jobs with the job title clearly identified)
4.  Application of both engineering solutions and administrative controls, with engineering solutions treated as long-term solutions
5.  Awareness of ergonomic considerations by design engineering that utilizes them in new or reengineered designs (people are an important design consideration)
6.  Frequent refresher training in ergonomics, including short courses and seminars for site-appointed "ergonomists"

## 6.5  Employer Benefit from Ergonomic Programs

In 1997, the U.S. Government Accountability Office (GAO) issued a report in response to a charge to (1) identify core elements of effective ergonomics programs and describe how these are operationalized within companies, (2) examine whether or not such programs have proven beneficial to companies and employees where they have been implemented, and (3) address the implications for employers who have not adopted such programs (GAO, 1997). The core elements they identified were consistent with those listed above. The parts of the report that are particularly interesting are the five case studies that are included, which provide details of the ergonomic program experiences of five different companies, in different sectors of industry, and how each tailored the generic components of a successful program to fit their circumstances, company culture, and so on. Examples of specific benefits realized by the companies are also provided and include reductions in numbers of lost workdays, workers' compensation costs (total and per case), and MSD incidence rates. Methods for supporting the case for ergonomics based on economics are provided by a number of authors (Andersson, 1992a,b; Simpson and Mason, 1995; Oxenburgh, 1997).

On a smaller scale, specific interventions may require cost–benefit analyses to justify them if a substantial initial investment is required. Seeley and Marklin (2003) employed the cost–benefit analysis method described by

Rouse and Boff (1997) to calculate the expected benefit to an electric utility company as a result of replacing the manual cutters and presses their linemen used with battery-powered models. Based on the quantification of expected benefits (reductions in medical and workers' compensation costs due to UEMSDs, costs to replace injured workers, training costs for replacement workers, and additional medical expenses for employees who postpone reporting injuries until they become severe), they determined a payback period of only four months for the $300,000 cost of the new tools. Details of their methodology are provided in Seeley and Marklin (2003).

## 6.6 Surveillance

### 6.6.1 Surveillance System

To evaluate the extent of WUEDs in a working population, a surveillance system should be used. Surveillance refers to the ongoing systematic collection, analysis, and interpretation of health and exposure data. Relevant to this chapter, this refers to the process of describing and monitoring work-related MSD occurrence. Surveillance is used to determine which jobs need further evaluation and where ergonomic interventions may be warranted. Surveillance data are used to determine the need for occupational safety and health action and to plan, implement, and evaluate ergonomic interventions and programs (Klaucke et al., 1988). Health and hazard (job risk factor) surveillance provides employers and employees with a means of evaluating WUEDs and workplace ergonomic risk factors systematically by monitoring trends over time. This information can also be of benefit for planning, implementing, and evaluating ergonomic interventions.

Although the climate for standards making has cooled in the United States, the final draft of the standard for management of WMSDs from the American National Standards Institute (ANSI) Z365 Committee (2002) is still a valuable source of information on the elements of a management program, including the surveillance component. The draft standard describes surveillance as including (1) review and analysis of existing records on worker injury and illness (OSHA 300 logs, company medical records, etc.), (2) worker reports concerning MSD symptoms or potential risk factors in the workplace, and (3) job surveys (cursory or screening-level reviews of jobs conducted to identify potential risk factors and the degree of risk they might pose to workers). The goal of surveillance, and of ergonomics programs in general, is to reduce or eliminate MSD risk factor exposure through approaches that are both reactive (after MSDs or their symptoms develop in workers) and proactive (identify hazards before workers who are exposed to the hazard develop a problem).

### 6.6.2 Worker Health Data

Analysis of existing records will be used to estimate the potential magnitude of the problem in the workplace. The number of employees in each job, department, or similar population needs to be determined first. Then the incidence rates can be calculated on the basis of hours worked:

Incidence (new case) rate (IR)

$$= \text{(no. of new cases during time} \times 200,000)/\text{work hours} \qquad (4)$$

where time refers to the time period of interest, typically one year, and work hours refers to the total number of hours worked by all employees in the group for which the rate is being calculated. The incidence rate is equivalent to the number of new cases per 100 full-time workers (assuming that each works 40 h per week and 50 weeks per year). Workplace-wide incidence rates (IRs) can be calculated for all WUEDs classified by body location for each department, process, or type of job. If specific work hours are not readily available, the number of full-time equivalent employees in each area multiplied by 2000 h can be used to estimate the denominator. Another important calculation is that of severity, used to describe the number of lost workdays. One way to calculate severity is to substitute the number of lost workdays for the number of new cases in equation (4). Another would be to examine the number of lost workdays per case. Prevalence refers to the number of existing cases relative to the number of employees in the group. These numbers can be compared within a company among different departments to see where problems exist. They can also be compared with data from the BLS to provide a sense of where a company's statistics are relative to those of its industry or business sector. That information can be found on the BLS website (http://www.bls.gov/iif/oshcdnew.htm). [Note that BLS incidence rate data are provided per 10,000 workers, not per 100, as in equation (4).]

In addition to making use of existing records, information about current symptoms can be sought through use of employee surveys. These usually provide employees with diagrams and other means by which to indicate where they are experiencing symptoms as well as the intensity and frequency of occurrence. In their chapter on surveillance, Hagberg et al. (1995) provide some examples of symptom surveys. Once a symptom survey has been conducted, the employer must be prepared to follow up with job analysis if problems are identified through the symptom survey.

### 6.6.3 Job Surveys

Job surveys are performed to identify specific jobs and processes that may put employees at risk of developing WMSDs. Conduct surveys of all jobs, a representative sample, or jobs that have been identified as potential problems through some other method (such as jobs with excessive turnover or absenteeism or when a substantial change is made to a job). Job surveys may include walkthroughs; conversations with employees, supervisors, and/or company health personnel; use of checklists; and other basic methods.

### 6.6.4 Data Collection Instruments

The surveillance system aims to link the occurrence of WMSDs to work-related risk factors. Ideally, the surveillance should make it possible to identify workplace risk factors before symptoms develop. Surveillance data collection instruments can be passive or active in nature (Hagberg et al., 1995). A summary of active and passive surveillance methods is given in Table 19. The passive surveillance process relies on information collected from existing databases and records (e.g., company dispensary logs, insurance records, workers' compensation records, accident reports, and absentee records) to identify the WRMD cases and patterns and potential problem jobs. Passive surveillance records are often useful in helping to determine the frequency with which active surveillance tools should be used and the interventions required or in assessing the effectiveness of ergonomics programs. In addition, a brief job analysis or physical demand analysis to assess the suitability of a job for the return to work of an injured worker can also be used for passive risk factor surveillance.

Active surveillance uses specifically designed tools and information, such as checklists and job analysis. As shown in Table 20, there can be both health active surveillance and risk factor active surveillance. Since

**Table 20  Summary of Tools Used in Surveillance**

| Approach | Tools |
|---|---|
| Health | |
|   Passive | Existing records |
|   Active, level 1 | Symptoms surveys or questionnaires (self or group administered) |
|   Active, level 2 | Health-related interviews and/or brief physical exams |
| Risk factors | |
|   Active, level 1 | Quick checklists of risk factors |
|   Active, level 2 | In-depth job analysis |

*Source:* Adapted from Kuorinka and Forcier (1995).

most musculoskeletal disorders produce some symptoms of pain or discomfort, health questionnaires are useful in identifying new or incipient problems as well as for assessing the effectiveness of medical interventions and ergonomic controls. In addition to symptom questionnaires, medical interviews and examination can also be used in active health surveillance (Table 21).

### 6.6.5 Analysis and Interpretation of Data

The surveillance data can be analyzed and interpreted to study possible associations between the WMSD surveillance data and the risk factor surveillance data (Hagberg et al., 1995). The two principal goals of the analysis are (1) to help identify patterns in the data that reflect differences between jobs or departments and (2) to target and evaluate intervention strategies. This analysis can be done on the number of existing WMSD cases (cross-sectional analysis) or during a specific period of time on the number of new WMSD cases in a retrospective and prospective fashion (retrospective and prospective analysis).

**Table 19  Passive and Active Surveillance Methods**

| Passive Surveillance[a] | Active Surveillance |
|---|---|
| Information source and method already exist and are usually designed for other administrative purposes | Information source and method specifically designed for surveillance |
| Relatively inexpensive | Modest to quite expensive |
| Usually requires additional coding of information for the purpose [e.g., surrogate(s) of exposure, such as job titles] | Since tools are "tailor made," includes at least job title information and other data considered important by surveillance analyst; will include data for linking of information between risk factor and WMSD data |
| *Examples:* health and safety logs, medical department logs, workers' compensation data, early retirement, medical insurance, absenteeism and transfer records, accident reports, product quality, productivity | *Examples:* for WMSD surveillance: confidential questionnaires without personal identifiers, questionnaire interviews, physical examinations; for risk factor surveillance: workplace walk-throughs, job checklists, postural discomfort surveys |

*Source:* Adapted from Kuorinka and Forcier (1995).
[a]Used mostly for health surveillance since, in practice, no existing records have been used to obtain information on risk factors associated with WMSDs.

**Table 21  Examples of Tools for WMSD Surveillance**

| Focus of Surveillance | Methods of Surveillance | |
|---|---|---|
| | Passive | Active |
| Health (WMSDs) | Company dispensary logs Insurance records Workers compensation records Accident reports Transfer requests Absentee records Grievances | Checklists Questionnaires Interviews Physical exams |
| Workplace risk factors (associated with WMSDs) | Not really used for WMSD risk factor yet[a] | Checklists Questionnaires Job analysis |

*Source:* Adapted from Kuorinka and Forcier (1995).
[a]The use of surrogate measures for exposure (e.g., job title of firm's department) could be viewed as "passive surveillance."

**Table 22  Examples of Odds Ratio Calculations for a Firm of 140 Employees**

| Risk Factor (e.g., Overhead Work for More Than 4 h) Is: | WMSDs Are:[a] | | |
|---|---|---|---|
| | Present | Not Present | Total |
| Present | 15 (A) | 25 (B) | 40 (A + B) |
| Not present | 15 (C) | 85 (D) | 100 (C + D) |
| Total | 30 (A + C) | 110 (B + D) | 140 (N) |

*Source:* Adapted from Kuorinka and Forcier (1995).
[a]Number in each cell indicates the count of employees with or without WMSD and the risk factor. Odds ratio (OR) = (A x D) / (B x C) = (15 x 85) / (25 x 15) = 3.4.

One of the simplest ways to assess the association between risk factors and WMSDs is to calculate the ORs (see Table 22). For this example, the prevalence data obtained in health surveillance are linked with the data obtained in risk factor surveillance. In the example shown in Table 22 (for more details, see Hagberg et al., 1995), one risk factor is selected at a time (e.g., overhead work for more than 4 h). Using the data obtained in surveillance, the following numbers of employees are counted:

- Employees with WMSDs exposed to more than 4 h of overhead work (15 workers)
- Employees with WMSDs not exposed to more than 4 h of overhead work (15 workers)
- Employees without WMSDs exposed to more than 4 h of overhead work (25 workers)
- Employees without WMSDs not exposed to more than 4 h of overhead work (85 workers)

The overall prevalence for the company is 30/140, or 21.4%. The prevalence for those exposed to the risk factor is 37.5% (15/40) compared with 15.0% (15/100) for those not exposed. The risk of having a WMSD depending on exposure to the risk factor, the OR, can be calculated using the number of existing cases of WMSD (prevalence). In the example above, those exposed to the risk factor have 3.4 times the odds of having the WMSD than those not exposed to the risk factor. An OR greater than l indicates an elevated risk. Such ratios can be monitored over time to assess the effectiveness of the ergonomics program in reducing the risk of WMSD, and a variety of statistical tests can be used to assess the patterns seen in the data.

## 6.7  Procedures for Job Analysis

Detailed job analysis typically consists of analyzing the job at the element or microlevel. Job surveys, on the other hand, can be used for establishing work relatedness, for prioritizing jobs for further analysis, or for proactive risk factor surveillance. Job analysis involves breaking down the job into component actions, measuring and quantifying risk factors, and identifying the problems and conditions contributing to each risk factor. Tools that might be employed to perform

a job analysis include videotape, tape measure, scale to weigh tools and parts, stopwatch to measure exposure duration, and possibly more sophisticated tools (electrogoniometers to measure wrist joint posture and motion; electromyographic equipment to assess muscle activity; vibration analysis equipment for assessing a powered hand tool's vibration characteristics). Exposures are characterized by magnitude, duration, and rate of repetition. Work organization factors, such as number of hours in a shift (8, 10, 12 h), job rotation schedule if applicable, and pay system (hourly, incentive, etc.). These data may be examined relative to existing research findings, regarding levels of exposure associated with elevated risk, or they may be used as input to tools such as the SI (Moore and Garg, 1995), HAL (ACGIH, 2002), or others, to determine the degree of risk posed by the hazards (risk factors). The website of the Washington State Department of Labor and Industries provides assessment tools as well (http://www.Ini.wa.gov/Safety/Topics/Ergonotnics/Services Resources/Tools/default.asp).

The job analysis should be performed at a sufficient level of detail to identify potential work-related risk factors associated with WMSDs and include the following steps: (1) collection of pertinent information about the job (number of employees on the job, which jobs precede and follow it, cycle time, tools used, etc.), (2) interview of a representative sample of workers, (3) breakdown of the job into tasks or elements, (4) description of the component actions of each task or element, (5) measurement and quantification of WMSD risk factors, (6) identification of the risk factors for each task element, (7) identification of the problems contributing to risk factors (root-cause analysis), and (8) summary of the problems and needs for intervention for the job. If intervention is required, once it has been developed, with input from workers and others, and put in place, follow-up assessments should be performed to ensure that the intervention is effective in dealing with the former problem and that no new problems are inadvertently introduced with the intervention. It is also important to document all of these steps within the analysis process in order to track progress, provide justification for changes, and share information with others about successful interventions.

## 6.8  Medical Management

Both scholarly academic work and various federal agencies and their activities recognize that WUEDs are responsible for a significant proportion of loss of work and also the fact that not all WUEDs that manifest at work can be prevented. Thus, it is extremely important that the management of WUEDs takes the appropriate approach.

### 6.8.1  Basic Activities

The primary objective of medical management in occupational health and safety programs is the prevention of work-related disorders and injuries (Hagberg et al., 1995). The specific goals of occupational health programs relevant to prevention of musculoskeletal

disorders were specified by the American Medical Association (AMA, 1972) as follows: (1) protecting employees against health and safety hazards in their work situation; (2) evaluating workers' physical, mental, and emotional capacity before job placement; (3) ensuring that employees can perform the work with an acceptable degree of efficiency and without endangering their own health and safety or that of others; (4) ensuring adequate medical care and rehabilitation for the occupationally ill or injured: and (5) encouraging and assisting with measures for personal health maintenance, including the acquisition of a personal physician whenever possible.

Medical management of WUEDs includes medical diagnosis, treatment, rehabilitation and return to work, and work hardening (Karwowski and Kasdan, 1988). In addition to these activities, medical management should also be involved in both passive and active health surveillance, job skills training programs, and ergonomic task force activities (Hagberg et al., 1995). As discussed in Section 6.1, the use of injury reports for health surveillance purposes is a form of passive health surveillance. The effective passive health surveillance requires data that have a high sensitivity for WUEDs. Injury reports should be followed up by workplace visits and an evaluation. In a population of workers or in a specific job category where there is a high risk of WUEDs, it may be necessary to perform the active health surveillance (i.e., periodic medical evaluations to identify workers in the early stages of a disease) and to target these workers for early secondary prevention efforts (i.e., medical treatment).

### 6.8.2 Medical Treatment

In general, the medical treatment efforts for WUEDs in the acute phase are similar to the treatments used for non-work-related disorders. As discussed by Hagberg et al. (1995), the general therapeutic objectives for WUEDs should include the following: (1) promotion of rest for the anatomical structures affected, (2) diminished spasms and inflammation, (3) reduction of pain, (4) increase in strength and endurance, (5) increase in range of motion, (6) alteration of mechanical and neurological structures, (7) increase in functional and physical work capacity, and (8) modification of work content and social environment.
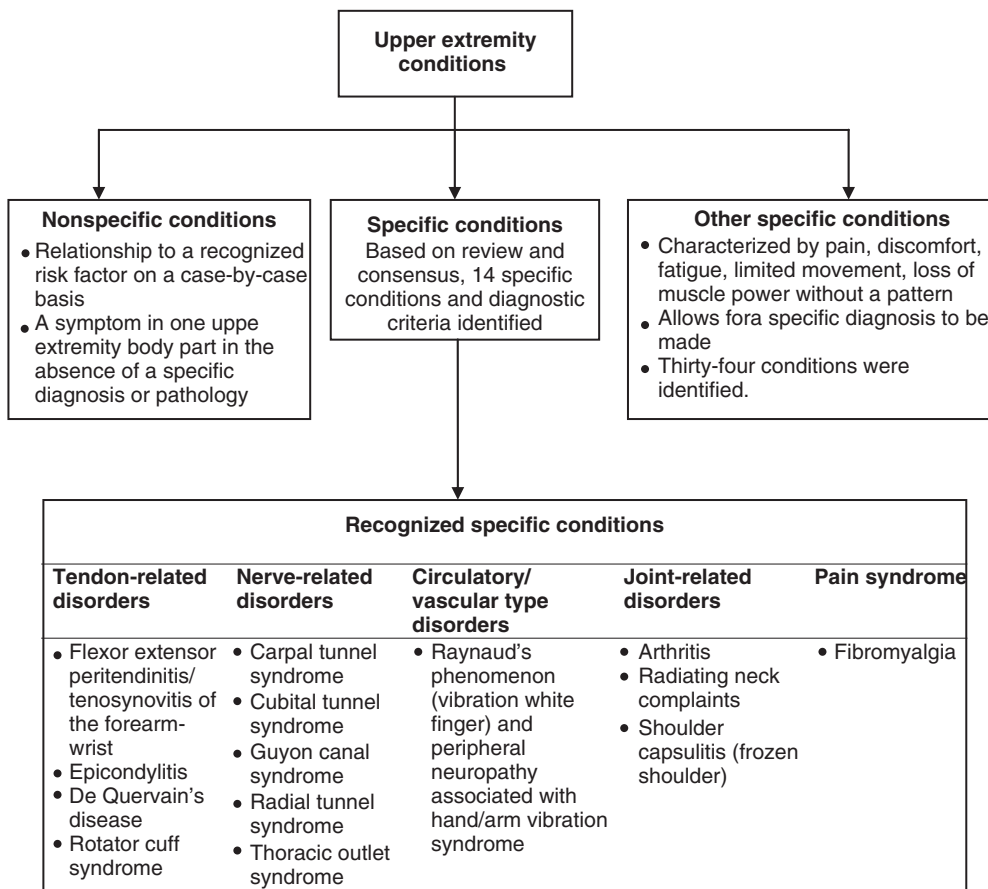


**Figure 7** Proposed model for classifying work-related upper extremity conditions.

**Table 23 Strength of Various WMSD Management Components**

| Criterion | Evidence | Strength of Evidence | Source |
|---|---|---|---|
| Classification and diagnosis | Classification and diagnosis are problematic. Lack of agreement on diagnostic criteria, even for the specific conditions (e.g., tenosynovitis, epicondylitis and rotator cuff syndrome), and Inconsistent application, both in the clinic and in the workplace, lead to misdiagnosis, incorrect labeling, and difficulties in interpretation of research findings. | Strong | Beaton et al. (2007) Huisstede et al. (2006) Van Eerd et al. (2003) Walker-Bone et al. (2003) |
| | The scientific basis for descriptive classification terms implying a uniform etiology is weak or absent (e.g., RSI or CTD). They are inconsistently applied/understood, and there is an argument that such terms should be avoided. | Moderate | Szabo (2006) Hagberg (2005) Lucire (2003) Bonde et al. (2003) |
| Epidemiology | There is a very high background prevalence of upper limb pain and neck symptoms in the general population (e.g., the 1-week prevalence in general population can be as high as 50%), Estimates of the prevalence rates of specific diagnoses are less precise but are considerably lower than for nonspecific complaints, and rates vary depending on region, population, country, case definition, and question asked. | Strong | Huisstede et al. (2006) Palmer and Smedley (2007) Eltayeb et al. (2007) Silverstein et al. (2006) |
| | Upper limb pain is often recurrent and frequently experienced in more than one region at the same time (both bilaterally and at anatomically adjacent sites). | Moderate | Macfarlane et al. (2000) Walker-Bone et al. (2004a) Silverstein et al. (2006) |
| | WUEDs often lead to difficulty with normal activities and to sickness absence, yet most workers with WUEDs can and do remain at work. | Strong | Walker-Bone et al. (2004b) Silverstein et al. (2006) Lee et al. (2006) Baldwin and Butler (2006) |
| Association and risks | Published scientific reviews (which included much cross-sectional data) concluded that there were strong associations between biomechanical occupational stressors (e.g. repetition, force) and WUEDs. Supported by plausible mechanisms from the biomechanics literature, the association was generally considered to be causative, particularly for prolonged or multiple exposures (however, a dose–response relationship generally was not evident). | Moderate | NIOSH (1997) NRC (1999) Industrial Injuries Advisory Council (2006) |
| | More recent longitudinal epidemiological studies also suggest an association between physical exposures and development of WUEDs, but they report the effect size to be rather modest and largely confined to intense exposures. The predominant outcome investigated (primary causation, symptom expression, or symptom modification) is inconsistent across studies and remains a subject of debate. This is true for regional complaints and (with few exceptions) most of the specific diagnoses. | Strong | Walker-Bone and Cooper (2005) Palmer and Smedley (2007) Coggon et al. (2000) Ijmker et al. (2007) |
| | The evidence that cumulative exposure to typical work is the cause of most reported upper limb injury is limited and inconsistent. | Weak | Macfarlane et al. (2000) NIOSH (1997) Hadler (2005) Dembe (1996) |

**Table 23  (Continued)**

| Criterion | Evidence | Strength of Evidence | Source |
|---|---|---|---|
| | Workplace psychosocial factors (beliefs, perceptions, and work organization) have consistently been found to be associated with various aspects of WUEDs, These include symptom expression, care seeking, sickness absence, and disability. | Strong | Walker-Bone and Cooper (2005) Bongers et al. (2006) Woods (2005) Burton et al. (2005) |
| | Individual psychological factors (such as anxiety, distress, and depression) have consistently been found to be associated with various aspects of WUEDs, including symptom expression, care seeking, sickness absence, and disability. | Strong | Mallen et al. (2007) Alizadehkhaiyat et al. (2007) Coutu et al. (2007) Henderson et al. (2005) |
| Interventions for MSDs in general | General management principles are to provide advice that promotes self-management, such as staying active and engaging in productive activity (with appropriate modifications). Pain modulation and control should be directed toward allowing appropriate levels of activity. | Weak | ARMA (2007) Breen et al. (2007) |
| | Programs using cognitive–behavioral approaches are effective and cost-effective at reducing pain and increasing productive activity in both the earlier and the later phases. | Strong | Hanson et al. (2006) Meijer et al. (2005) Marhold et al. (2001) |
| | Multimodal integrated interventions that address both biomechanical and psychosocial aspects at the same time should be useful for managing musculoskeletal problems in the workplace. | Weak | Waddell and Burton (2004) Cole et al. (2006) Selander et al. (2002) Feuerstein et al. (2003) |
| Interventions specifically with respect to WUEDs | Pain management programs using cognitive behavioral principles and multidisciplinary occupational rehabilitation for people with WUEDs can improve occupational outcomes in the short term and significantly reduce days away from work in the longer term. Earlier intervention appears to yield better results. | Moderate | Crawford and Laiou (2007) Feuerstein et al. (1999) |
| | There is a conceptual case that rehabilitation should be started early and that long periods of rest or sick leave are generally counterproductive. | Weak | Hagberg (2005) Helliwell and Taylor (2004) NHMRC (2004) Waddell and Burton (2004) |
| | Ergonomic work redesign directed at equipment or organization has not been shown to have a significant effect on incidence and prevalence rates of WUEDs. Ergonomic interventions can improve worker comfort that can in principle contribute positively to multimodal interventions. | Moderate | Boocock et al. (2007) Szabo (2006) Hadler (2005) Pransky et al. (2002) |
| | There is limited evidence that ergonomic adjustments (e.g., mouse/keyboard design) can reduce upper limb pain in display screen workers Insufficient evidence for equipment interventions among manufacturing workers. | Weak | Boocock et al. (2007) Verhagen et al. (2006) Williams et al. (2004) |
| | In general resting injured upper limbs delays recovery, | Weak | Nash et al. (2004) Melhorn (2005) |

**Table 23** (*Continued*)

| Criterion | Evidence | Strength of Evidence | Source |
|---|---|---|---|
| | Early activity improves pain and stiffness and can speed return to work yet does not increase complications or residual symptoms and may lead to less treatment consumption. | | Haahr and Andersen (2003) Cheng et al. (2007) |
| | There is wide consensus that early return to work is an important goal which should be facilitated by multimodal interventions, including provision of accurate information, pain relief, and encouragement of activity. An integrative approach by all the players (i.e., employer, worker, and health professional) is conceptually a fundamental requirement. | Weak | Cheng and Hung (2007) Lee and Higgins (2006) Breen et al. (2007) Meijer et al. (2005) |
| | Although the components of return-to-work interventions vary, there is emerging evidence that integrative approaches can be effective for MSDs in general and possibly for WUEDs. Facilitation of return to work through temporary transitional work arrangements (modified work) seems to be an important component. | Moderate | Breen et al. (2007) McCluskey et al. (2006) Cole et al. (2006) Franche et al. (2005) |
| Nonspecific complaints and specific diagnoses | There is insufficient robust evidence to identify reliable prognostic indicators that are applicable across the WUED spectrum (specific diagnoses and regional complaints). | Weak | N⊘rregaard et al. (1999) Hagberg (2005) Kuijpers et al. (2004) Ryall et al. (2007) |
| | There is inconsistent and conflicting evidence on whether and to what extent certain specific diagnoses and regional complaints should be conceived differently in terms of overall management targeted at vocational outcomes. | Weak | Hadler (2005) Melhorn (2005) Derebery et al. (2006) Staal et al. (2007) |

*Source:* After Burton et al. (2009).

Medical treatment of WUEDs begins with classification and diagnosis. As Boocock et al. (2009) reviewed, to date, there is much confusion with (1) operational definition of MDSs and (2) accurate classification and diagnosis of conditions appropriate to a particular case. The problem initiates with the nomenclature of the general term of MDS that varied within and between countries with about 14 different terms. The subsequent challenge is identification of conditions. Following the review, Boocock et al. (2009) suggested three broad categories of conditions: (1) specific conditions, (2) nonspecific conditions, and (3) other specific conditions. With these three identification conditions, Boocock et al. (2009) proposed the classification model shown in Figure 7.

### 6.8.3  Rehabilitation, Return to Work, and Work-Hardening Programs

A program that promotes healing and helps an injured worker to return to work and specifies appropriate job placement conditions based on different job tasks and work requirements is called occupational rehabilitation. Since the injury may not always have only a physical basis, psychosocial (at work and outside work) and psychological disability aspects are essential parts of the rehabilitation process. According to the Commission on Accreditation of Rehabilitation Facilities (CARF, 1989), a work-hardening program is a highly structured, goal-oriented, individualized treatment program designed to

maximize a person's ability to return to work. Such a program uses a set of conditioning tasks that are graded progressively in the quest to improve biomechanical, neuromuscular, cardiovascular, and psychosocial functions with real or simulated work activities.

In a recent review, Burton et al. (2009) evaluated various components of WMSD management: (1) classification and diagnosis, (2) epidemiology, (3) association and risks, (4) generic interventions, (5) specific interventions, (6) return to work, and (7) nonspecific complaints and specific diagnoses. Three levels of strength of evidence were used: (1) strong (consistent findings provided by multiple studies), (2) moderate (consistent findings provided by fewer and lower quality studies), and (3) weak (single study, general consensus/guidance). The findings of the review are provided in Table 23.

### 6.8.4  Preemployment and Preplacement Screening

Although there is no scientific evidence that screening can predict the development of WUEDs, preemployment and preplacement screening may be an important part of medical management activities (Hagberg et al., 1995). According to the American College of Occupational and Environmental Medicine Committee on Occupational Medical Practice (ACOM, 1990), screening refers to the application of at least one test (or examination) to

workers in order to identify apparently healthy workers who are at high risk of developing a specific WUED from those workers who are not. Although the screening tests are not diagnostic, preemployment screening and examination are typically performed before any offer of employment can be made. On the other hand, a preplacement screening process is an examination of an employee who has already received an offer of employment and addresses a question of employee placement in a specific job.

## 7  SUMMARY

### 7.1  Balancing Work System for Ergonomics Benefits

As pointed out by Hagberg et al. (1995), there are no perfect jobs or perfect workplaces that are free of all work-related hazards and provide ideal psychosocial conditions for complete satisfaction for all employees. Therefore, one must consider the trade-offs between competing needs for ergonomic improvements at the workplace and establish a basis for identifying the most critical workplace characteristics for design or redesign. Such trade-offs between the biomechanical factors, personal factors, and work organizational factors, including work stress, coping strategies, and organizational practices, require one to balance various ergonomic needs to achieve the solution that will have the greatest benefits for employee health and productivity.

The balance theory-based model proposed by Smith and Sainfort (1989) takes a systems approach by focusing on the interactions between the worker, including the physical characteristics, perceptions, personality and work behavior; the physical and social environments; and the organizational structure that defines the nature and level of worker involvement, interaction, control, and supervision. The capabilities of technologies available to a worker to perform a specific job affect task performance and the worker's skills and knowledge needed for their effective use. Task requirements affect the required skills and knowledge of the worker. Both the tasks and technologies affect the content of the job and physical demands. The balance theory-based model can be used to establish relationships between interacting elements such as job demands, job design factors, and ergonomic loads. Demands that are placed on the worker create loads that can be healthy or harmful. Harmful loads may lead to physical and psychological stress responses that can produce adverse health effects such as WUEDs. It should be noted that a number of personal considerations may also contribute to the physical and psychological effects. These include the strength and health of the worker, previous musculoskeletal or nerve injury, personality, perceptual–motor skills and abilities, physical conditioning, prior experience and learning, motives, goals, and needs and intelligence.

### 7.2  Ergonomics Guidelines

The expected benefits of managing WUEDs in industry are improved productivity and quality of work products, enhanced safety and health of the employees, higher employee morale, and accommodation of people with various degrees of physical abilities. Strategies for managing the WUEDs at work should focus on prevention efforts and should include, at the plant level, employee education, ergonomic job redesign, and other early intervention efforts, including engineering design technologies such as workplace reengineering and active and passive surveillance. At the macrolevel, management of the WUEDs should aim to provide adequate occupational health care provisions, legislation, and industry-wide standardization.

Already widely recognized in Europe (Wilson, 1994), ergonomics has to be seen as a vital component of the value-adding activities of a company. Even in strictly financial terms, the benefits of an ergonomics management program will outweigh the costs of the program. A company must be prepared to engage in a participative culture and to utilize participative techniques. The ergonomics-related problems and consequent interventions should go beyond engineering solutions and must include design for manufacturability, total quality management, work organization, workplace redesign, and worker training. Only then will the promise of ergonomics in managing the WUEDs at work be fulfilled.

In the absence of generally applicable guidelines and criteria on minimizing and/or optimizing risk factor actor exposure, two complementary approaches have merit for the prevention of WUEDs: (1) general guidelines that describe in general terms the principles and policies to be adopted in preventing WUEDs and (2) specific guidelines that aim at the design and redesign of work and tasks that are known in detail (Hagberg et al., 1995). Since the specific guidelines draw on both scientific knowledge and the collective industrial experience, they may be much more detailed and often contain quantitative data.

Most of the current guidelines for control of the biomechanical risk factors for WUEDs at work aim to (1) reduce the exposure to highly repetitive and stereotyped movements, (2) reduce excessive force levels, and (3) reduce the need for sustained postures. For example, to control the extent of force required to perform a task, one should (1) reduce the force required through tool and fixture redesign, (2) distribute the application of force, or (3) increase the mechanical advantage of the (muscle) lever system. Because of the neurophysiological needs of the working muscles, adequate rest pauses (determined based on the scientific knowledge of the physiology of muscular fatigue and recovery) should be scheduled to provide relief for the most active muscles used on the job. Furthermore, reduction in task repetition can be achieved, for example, by (1) task enlargement (increasing variety of tasks to perform), (2) increase in the job cycle time, and (3) work mechanization and automation.

Finally, it should be noted that many of the recommendations offered by ergonomics may be difficult to implement in practice without full understanding of the production processes, plant layouts, or quality requirements and total commitment from all management levels and workers of the company. This is because many of the guidelines are not specific and define what to

avoid (e.g., avoid high contact forces and static loading, avoid extreme or awkward joint positions, avoid repetitive finger action, and avoid tool vibration) but do not define how to avoid these risk factors. In view of the above, involvement of professional ergonomists (i.e., those who are certified by the Board of Certification in Professional Ergonomics), along with engineering personnel and production workers in a truly participative manner, is critical to the success of ergonomic intervention efforts. Furthermore, ergonomics must be treated with the same level of attention and significance as other business functions of the plant (e.g., quality management control) and be accepted as the cost of doing business rather than add-on activity requiring action only when problems arise.

## REFERENCES

Alexander, D. C., and Orr, G. B. (1992), "The Evaluation of Occupational Ergonomics Programs," in *Proceedings of the Human Factors Society 36th Annual Meeting*, Santa Monica, CA, pp. 697–701.

Alizadehkhaiyat, O., Fisher A. C., Kemp G. J., and Frostick S. P. (2007), "Pain, Functional Disability, and Psychologic Status in Tennis Elbow," *Clinical Journal of Pain*, Vol. 23, pp. 482–489.

Almekinders, L. C. (1998), "Tendinitis and Other Chronic Tendinopathies," *Journal of the American Academy of Orthopaedic Surgeons*, Vol. 6, No. 3, pp. 157–164.

American College of Occupational and Environmental Medicine Committee on Occupational Medical Practice (ACOM) (1990), "Preplacement/Preemployment Physical Examinations," *Journal of Occupational Medicine*, Vol. 32, pp. 295–299.

American Congress of Government Industrial Hygienists (ACGIH) (2002), *TLVs and BEls*, ACGIH, Cincinnati, OH.

American Congress of Government Industrial Hygienists (ACGIH) Worldwide (2001), *Threshold Limit Values for Chemical Substances and Physical Agents*, ACGIH, Cincinnati, OH.

American Medical Association (AMA) (1972), *Scope, Objectives, and Functions of Occupational Health Programs*, AMA, Chicago.

American National Standards Institute (ANSI) Z365 Committee (2002), "Management of Work-Related Musculoskeletal Disorders," ANSI/National Safety Council, available: http://www.nsc.org/ehc/z365.htm, accessed 2004.

Anderson, E. R. (1992a), "Economic Evaluation of Ergonomic Solutions, 1: Guidelines for the Practitioner," *International Journal of Industrial Ergonomics*, Vol. 10, pp. 161–171.

Anderson, E. R. (1992b), "Economic Evaluation of Ergonomic Solutions, II: The Scientific Basis," *International Journal of Industrial Ergonomics*, Vol. 10, pp. 173–178.

Arthritis and Musculoskeletal Alliance (ARMA) (2007), *Standards for People with Regional Musculoskeletal Pain*, ARMA, London.

Armstrong, T. J. (1986), "Upper-Extremity Posture: Definition, Measurement and Control," in *The Ergonomics of Working Postures*, N. Corlett, J. Wilson, and I. Manenica, Eds., Taylor & Francis, London, pp. 57–73.

Armstrong, T. J., and Chaffin, D. B. (1979), "Some Biomechanical Aspects of the Carpal Tunnel," *Journal of Biomechanics*, Vol. 12, pp. 567–570.

Armstrong, T. J., and Lifshiz, Y. (1987), "Evaluation and Design of Jobs for Control of Cumulative Trauma Disorders," in *Ergonomic Interventions to Prevent Musculoskeletal Injuries in Industry*, Lewis Publishers, Chelsea, England, pp. 73–85.

Armstrong, T. J., Castelli, W. A., Evans, F. G., and Diaz-Perez, R. (1984), "Some Histological Changes in Carpal Tunnel Contents and Their Biomechanical Implications," *Journal of Occupational Medicine*, Vol. 26, pp. 197–201.

Armstrong, T. J., Radwin, R. G., Hansen, D. J., and Kennedy, K. W. (1986), "Repetitive Trauma Disorders: Job Evaluation and Design," *Human Factors*, Vol. 28, pp. 325–336.

Armstrong, T. J., Buckle, P., Fine, L. J., Hagberg, M., Jonsson, B., Kilbom, A., Kuorinka, I. A., Silverstein, B. A., Siogaard, G., and Viikari-Juntura, E. R. (1993), "A Conceptual Model for Work-Related Neck and Upper-Limb Musculoskeletal Disorders," *Scandinavian Journal of Work, Environment and Health*, Vol. 19, pp. 73–84.

Baldwin, M. L., and Butler, R. J. (2006), "Upper Extremity Disorders in the Workplace: Costs and Outcomes Beyond the First Return to Work," *Journal of Occupational Rehabilitation*; Vol. 16, pp. 303–323.

Balogh, I., et al. (2001), "Questionnaire-Based Mechanical Exposure Indices for Large Population Studies: Reliability, Internal Consistency and Predictive Validity," Scandinavian Journal of Work Environment and Health}, Vol. 27, pp. 41–48.

Bao, S., Spielholtz, P., Howard, N., and Silverstein, B. (2009), "Application of the Strain Index in Multiple Task Jobs," *Applied Ergonomics*, Vol. 40, pp. 56–68.

Barbe, M. F., Barr, A. E., Gorzelany, I., Amin, M., Gaughan, J. P., and Safadi, F. F. (2003), "Chronic Repetitive Reaching and Grasping Results in Decreased Motor Performance and Widespread Tissue Responses in a Rat Model of MSD," *Journal of Orthopaedic Research*, Vol. 21, pp. 167–176.

Barr, A. E., and Barbe, M. F. (2002), "Pathophysiological Tissue Changes Associated with Repetitive Movement: A Review of the Evidence," *Physical Therapy*, Vol. 82, pp. 173–187.

Beaton, D. E., Bombardier, C., Cole, D. C., Hogg-Johnson, S., and Van Eerd, D. (2007), "A Pattern Recognition Approach to the Development of a Classification System for Upper-Limb Musculoskeletal Disorders of Workers," *Scandinavian Journal of Work Environmental and Health*, Vol. 33, pp. 131–139.

Bernard, B. P. (1997), *Musculoskeletal Disorders and Workplace Factors*, DHHS (NIOSH) Publication 97-141, U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Washington, DC.

Bonde, J. P., et al. (2003), "Prognosis of Shoulder Tendonitis in Repetitive Work: A Follow Up Study in a Cohort of Danish Industrial and Service Workers," *Occupational and Environmental Medicine*, Vol. 60, 649–654.

Bongers, P., Ijmker, S., den Heuvel, S. V., and Blatter, B. (2006), "Epidemiology of Work Related Neck and Upper Limb Problems: Psychosocial and Personal Risk Factors (Part I) and Effective Interventions from a Bio Behavioural Perspective (Part II)," *Journal of Occupational Rehabilitation*, Vol. 16, pp. 272–295.

Bongers, P. M., Kremer, A. M., and ter Laak, J. (2002), "Are Psychosocial Factors, Risk Factors for Symptoms and Signs of the Shoulder, Elbow, or Hand/Wrist? A Review of the Epidemiological Literature," *American Journal of Industrial Medicine*, Vol. 41, pp. 315–342.

Boocock, M. G., et al. (2007), "Interventions for the Prevention and Management of Neck/Upper Extremity Musculoskeletal Conditions: A Systematic Review," *Occupational and Environmental Medicine*, Vol. 64, pp. 291–303.

Boocock, M. G., Collier, J. M. K., McNair, P. J., Simmonds, M., Larmer, P. J., and Armstrong, B. (2009), "A Framework for the Classification of Diagnosis of Work-Related Upper Extremity Conditions: Systematic Review," *Seminars in Arthritis and Rheumatism*, Vol. 38, pp. 296–311.

Breen, A., Langworthy, J., and Baghust, J. (2007), "Improved Early Pain Management for Musculoskeletal Disorders," in A. Breen, J. Langworthy, and J. Bagust (Eds.), *Health & Safety Executive*, RR 399, Dorset, UK.

Buchholtz, B., Pacquet, V., Punnett, L., Lee, D., Moir, S. (1996), "PATH: A Work Sampling-Based Approach to Ergonomics Job Analysis for Construction and Other Non-Repetitive Work," *Applied Ergonomics*, Vol. 27, pp. 177–187.

Buckle, P. W., and Devereux, J. J. (2002), "The Nature of Work-Related Neck and Upper Limb Musculoskeletal Disorders," *Applied Ergonomics*, Vol. 33, pp. 207–217.

Bureau of Labor Statistics (BLS) (2002), *Bureau of Labor Statistics: Annual. Survey of Occupational Injuries and Illnesses*, U.S. Department of Labor, available: http://www.bls.gov/iif/ oshcdnew.htm, accessed 2004.

Bureau of Labor Statistics (BLS) (2008), Bureau of Labor Statistics: Nonfatal occupational injuries and illnesses, USDL 08-1716, U.S Department of Labor.

Burton, A. K., Bartys, S., Wright, I. A., and Main, C. J. (2005), *Obstacles to Recovery from Musculoskeletal Disorders in Industry*, RR 323, HSE Books, London, available: www.hse.gov.uk/research/rrhtm.

Burton, A. K., Kendall, N. A. S., Pearce, B. G., Birrell, L. N., and Bainbridge, C. (2009), "Management of Work-Relevant Upper Limb Disorders: A Review," *Occupational Medicine*, Vol. 59, pp. 44–52.

Chaffin, D., Andersson, G., and Martin, B. (1999), *Occupational Biomechanics*, 3rd ed., Wiley, New York.

Cheng, A. S. K., Hung, L. K. (2007), "Randomized Controlled Trial Of Workplace-Based Rehabilitation for Work-Related Rotator Cuff Disorder," *Journal of Occupational Rehabilitation*, Vol. 17, pp. 487–503.

Clark, B. D., Barr, A. E., Safadi, F. F., Beitman, L., AlShatti, T., Amin, M., Gaughan, J. P., and Barbe, M. F. (2003), "Median Nerve Trauma in a Rat Model of Work-Related Musculoskeletal Disorder," *Journal of Neurotrauma*, Vol. 20, pp. 681–695.

Clark, B. D., Al-Shatti, T. A., Barr, A. E., Amin, M., and Barbe, M. F. (2004), "Performance of a High-Repetition, High-Force Task Induces Carpal Tunnel Syndrome in Rats," *Journal of Orthopaedic and Sports Physical Therapy*, Vol. 34, pp. 244–253.

Cochrane Database of Systematic Reviews (2006), Issue 3, Art. No. CD003471.

Coggon, D., Palmer, K. T., and Walker-Bone, K. (2000), "Occupation and Upper Limb Disorders," *Rheumatology*, Vol. 39, pp. 1057–1059.

Cole, D. C., Van Eerd, D., Bigelow, P., and Rivilis, I. (2006), "Integrative Interventions for MSDS: Nature, Evidence, Challenges & Directions," *Journal of Occupational Rehabilitation*, Vol. 16, pp. 359–374.

Commission on Accreditation of Rehabilitation Facilities (CARF) (1989), *Standards Manual for Organizations Serving People with Disabilities*, CARF, Tucson, AZ.

Coutu, M. F., Durand, M. J., Loisel. P., Goulet. C., and Gauthier. N. (2007), "Level of Distress among Workers Undergoing Work Rehabilitation for Musculoskeletal Disorders," *Journal of Occupational Rehabilitation*, Vol. 17, pp. 289–303.

Crawford, J. O., and Laiou, E. (2007), "Conservative Treatment of Work Related Upper Limb Disorders—A Review," *Occupational Medicine*, London, Vol. 57, pp. 4–17.

Da Costa, B. R., and Vieira, R. (2010), "Risk Factors for Work-Related Musculoskeletal Disorders: A Systematic Review of Recent Longitudinal Studies," *American Journal of Industrial Medicine*, Vol. 53, pp. 285–323.

Dane, D., Feuerstein, M., Huang, G., Dimberg, L., Ali, D., and Lincoln, A. (2002), "Measurement Properties of a Self-Report Index of Ergonomic Exposures for Use in an Office Environment," *Journal of Occupational Rehabilitation*, Vol. 44, pp. 73–81.

David, G. C. (2005), "Ergonomic Methods For Assessing Exposure to Risk Factors for Work-Related Musculoskeletal Disorders," *Occupational Medicine*, Vol. 55, pp. 190–199.

Dembe, A. E. (1996), "Occupation and Disease: How Social Factors Affect the Conception of Work-Related Disorders", Yale University Press, New Haven, CT.

Derebery, J., Kadan, M, L., and Gonzalez, R. (2006), *Clinics in Occupational and Environmental Medicine*, Vol. 5, W. B. Saunders Company, New York.

Ditmars, D. M., Jr. (1993), "Patterns of Carpal Tunnel Syndrome," *Hand Clinics*, Vol. 9, pp. 241–252.

Eltayeb, S., Staal, B., Kennes, J., Lamberts, P. H. G., de Bie, R. A. (2007), "Prevalence of Complaints of Arm, Neck and Shoulder among Computer Office Workers and Psychometric Evaluation of a Risk Factor Questionnaire," *BMC Musculoskeletal Disorders*, Vol. 8

Engholm, G., and Holmstrom, E. (2005), "Dose–Response Associations between Musculoskeletal Disorders and Physical and Psychosocial Factors among Construction Workers," *Scandinavian Journal of Work Environment and Health*, Vol. 31, No. 2, pp. 57–67.

Falkiner, S., and Myers, S. (2002), "When Exactly Can Carpal Tunnel Syndrome Be Considered Work-Related?" *ANZ Journal of Surgery*, Vol. 72, pp. 204–209.

Fallentin, N., et al. (2001), "Physical Exposure Assessment in Monotonous Repetitive Work—The PRIM Study," *Scandinavian Journal of Work Environment and Health*, Vol. 27, pp. 21–29.

Feuerstein, M., Burrell, L. M., Miller, V. I., Lincoln, A., Huang, G. D., and Berger, R. (1999), "Clinical Management of Carpal Tunnel Syndrome: A 12-Year Review of Outcomes," *American Journal of Industrial Medicine*, Vol. 35, pp. 232–245.

Feuerstein, M., Huang, G. D., Ortiz, J. M., Shaw, W. S., Miller, V. I., and Wood, P. M. (2003), "Integrated Case Management for Work-Related Upper-Extremity Disorders: Impact of Patient Satisfaction on Health and Work Status," *Journal of Occupational & Environmental Medicine*, Vol. 45, pp. 803–812.

Franche, R. L., et al. (2005), "Workplace-Based Return-to-Work Interventions: A Systematic Review of the Quantitative Literature," *Journal of Occupational Rehabilitation*, Vol. 15, pp. 607–631.

Frasson, H. C., Gloria, R., Kilbom, A., Winkel, J., Karlqvist, L., and Wiktorin, C. (1995), "A Portable Ergonomic Observation Method (PEO) for Computerised On Line Recording of Postures and Manual Handling," *Applied Ergonomics*, Vol. 26, pp. 93–100.

Haahr, J. P., and Andersen, J. H. (2003), "Prognostic Factors in Lateral Epicondylitis: A Randomized Trial with One-Year Follow-Up in 266 New Cases Treated with Minimal Occupational Intervention or the Usual Approach in General Practice," *Rheumatology*, Vol. 42, pp. 1216–1225.

Haber, L. D. (1971), "Disabling Effects of Chronic Disease and Impairment," *Journal of Chronic Disease*, Vol. 24, pp. 269–487.

Hadler, N. M. (2005), *Occupational Musculoskeletal Disorders*, 3rd ed.., Lippincott Williams & Wilkins, Philadelphia, PA.

Hagberg, M. (1992), "Exposure Variables in Ergonomic Epidemiology," *American Journal of Industrial Medicine*, Vol. 21, pp. 91–100.

Hagberg, M. (2005), "Clinical Assessment, Prognosis and Return to Work with Reference to Work Related Neck and Upper Limb Disorders," *Giornale Italiano di Medicina del Lavoro ed Ergonomia*, Vol. 27, pp. 51–57.

Hagberg, M., Silverstein, B., Wells, R., Smith, M. J., Hendrick, H. W., Carayon, P., and Peruse, M. (1995), in *Work Related Musculoskeletal Disorders (WMSDs): A Reference Book for Prevention*, I. Kourinka and L. Forcier, Eds., Taylor & Francis, London.

Ham, S. J., Kolkman, W. F. A., Heeres, J., den Boer, J. A., and Vierhout, P. A. M. (1996), "Changes in the Carpal Tunnel Due to Action of the Flexor Tendons: Visualization with Magnetic Resonance Imaging," *Journal of Hand Surgery* [*American Volume*], Vol. 21, pp. 997–1003.

Hanson, M. A., Burton, A. K., and Kendall, N. A. S. (2006), in *The Costs and Benefits of Active Case Management and Rehabilitation for Musculoskeletal Disorders*, RR 493, R. J. Lancaster and A. Pilkington, Eds., HSE Books, London.

Hargens, A. R., Romine, J. S., Sipe, J. C., Evans, K. L., Mubarak, S. J., and Akeson, W. H. (1979), "Peripheral Nerve-Conduction Block by High Muscle-Compartment Pressure," *Journal of Bone and Joint Surgery*, Vol. 61-A, pp. 192–200.

Harrington, C. B., Siddiqui, A., and Feuerstein, M. (2009), "Work Style as a Predictor of Pain and Restricted Work Associated with Upper Extremity Disorders: A Prospective Study," *Journal of Hand Surgery*, Vol. 34, No. 4, pp. 724–731.

Hart, D. A., Archambault, J. M., Kydd, A., Reno, C., Frank, C. B., and Herzog, W. (1998), "Gender and Neurogenic Variables in Tendon Biology and Repetitive Motion Disorders," *Clinical Orthopaedics*, June, Vol. 351, pp. 44–56.

Health and Safety Executive (1998), *Manual Handling Operations Regulations. 1992 Guidance on Regulations L23*, HSE Books, Suffolk.

Health and Safety Executive (2002), *Upper Limb Disorders in the Workplace HSG60*, HSE Books, Suffolk.

Helliwell, P. S., and Taylor, W. J. (2004), "Repetitive Strain Injury," *Postgraduate Medical Journal* Vol. 80, pp. 438–443.

Helliwell, P. S., Bennett, R. M., Littlejohn, G., Muirden, K. D., and Wigley, R. D. (2003), "Towards Epidemiological Criteria for Soft-Tissue Disorders of the Arm," *Occupational Medicine* (*London*), Vol. 53, pp. 313–319.

Henderson, M., Kidd, B. L., Pearson, R. M., and White, P. D. (2005), "Chronic Upper Limb Pain: An Exploration of the Biopsychosocial Model," *The Journal of Rheumatology*, Vol. 32, pp. 118–122.

Hignett, S., and McAtamney, L. (2000), "Rapid Entire Body Assessment (REBA)," *Applied Ergonomics*, Vol. 31, pp. 201–205.

Hilderbrandt, V., Bongers, P., van, D. F., Kemper, H., and Dul, J. (2001), "Dutch Musculoskeletal Questionnaire: Description and Basic Qualities," *Ergonomics*, Vol. 44, pp. 1038–1055.

Holte, K., and Westgaard, R. (2001), "A Comparison of Different Methods for Assessment of Psychosocial Risk Factors for Shoulder and Neck Pain," in *Exposure Assessment in Epidemiology and Practice*, M. Hagberg, B. Knave, L. Lillienberg, and H. Westberg, Eds., Vol. 79, National Institute for Working Life S-112, Stockholm, Sweden, pp. 381–382.

Huisstede, B. M. A., Bierma-Zeinstra, S. M. A., Koes, B. W., and Verhaar, J. A. N. (2006), "5Incidence and Prevalence of Upper-Extremity Musculoskeletal Disorders. A Systematic Appraisal of the Literature," *BMC Musculoskeletal Disorder*, Vol. 7.

Ijmker, S., Huysmans, M. A., Blatter, B. M., van der Beek, A. J., van Mechelen, W., and Bongers, P. M. (2007), "Should Office Workers Spend Fewer Hours at Their Computer? A Systematic Review of the Literature," *Occupational and Environmental Medicine*, Vol. 64, pp. 211–222.

Industrial Injuries Advisory Council (2006), *Prescribed Diseases*, available: www.iiac.org.uk/about/index.asp, accessed February 11, 2006.

J?rvholm, U., Palmerud, G., Styf, J., Herberts, P., and Kadefors, R. (1988), "Intramuscular Pressure in the Supraspinatus Muscle," *Journal of Orthopaedic Research*, Vol. 6, pp. 230–238.

J?rvholm, U., Palmerud, G., Karlsson, D., Herberts, P., and Kadefors, R. (1991), "Intramuscular Pressure and Electromyography in Four Shoulder Muscles," *Journal of Orthopaedic Research*, Vol. 9, pp. 609–619.

Kadefors, R., and Forsman, M. (2000), "Ergonomic Evaluation of Complex Work: A Participative Approach Employing Video Computer Interaction, Exemplified in a Study of Order Picking," *International Journal of Industrial Ergonomics*, Vol. 25, pp. 435–445.

Karhu, O., Kansi, P., and Kuorinka, I. (1977), "Correcting Working Postures in Industry: A Practical Method for Analysis," *Applied Ergonomics*, Vol. 8, pp. 199–201.

Karwowski, W., and Kasdan, M. L. (1988), "The Partnership of Ergonomics and Medical Intervention in Rehabilitation of Workers with Cumulative Trauma Disorders of the Hand," in *Ergonomics in Rehabilitation*, A. Mital and W. Karwowski, Eds., Taylor & Francis, London, pp. 35–53.

Kee, D., Karwowski, W. (2001), "LUBA: An Assessment Technique for Postural Loading on the Upper Body Based on Joint Motion Discomfort and Maximum Holding Time," *Applied Ergonomics*, Vol. 32, pp. 357–366.

Keir, P. J. (2001), "Magnetic Resonance Imaging as a Research Tool for Biomechanical Studies of the Wrist," *Seminars in Musculoskeletal Radiology*, Vol. 5, pp. 241–250.

Keir, P. J., and Bach, J. M. (2000), "Flexor Muscle Incursion into the Carpal Tunnel: A Mechanism for Increased Carpal Tunnel Pressure? *Clinical Biomechanics* (*Bristol, Avon*), Vol. 15, pp. 301–305.

Keir, P., Bach, J. M., and Rempel. D. (1998a), "Fingertip Loading and Carpal Tunnel Pressure: Differences between a Pinching and a Pressing Task," *Journal of Orthopaedic Research*, Vol. 16, pp. 112–115.

Keir, P. J., Bach, J. M., and Rempel, D. M. (1998b), "Effects of Finger Posture on Carpal Tunnel Pressure during Wrist Motion," *Journal of Hand Surgery* [*American Volume*], Vol. 23, pp. 1004–1009.

Kemmlert, K. (1995) "A Method Assigned for the Identification of Ergonomics Hazards—PLIBEL," *Applied Ergonomics*, Vol. 26, pp. 199–211.

Ketola, R., Toivonen, R., and Viikari-Juntura, E. (2001), "Inter-Observer Repeatability and Validity of an Observation Method to Assess Physical Loads Imposed on the Upper Extremities," *Ergonomics*, Vol. 44, pp. 119–131.

Keyserling, M., Brouwer, M., and Silverstein, B. (1992), "A Checklist for Evaluating Ergonomic Risk Factors Resulting from Awkward Postures of the Legs, Trunk, and Neck," *International Journal of Industrial Ergonomics*, Vol. 9, pp. 283–301.

Keyserling, W. M., Stetson, D. S., Silverstein, B. A., and Brouwer, M. L. (1993), "A Checklist for Evaluating Ergonomic Risk Factors Associated with Upper Extremity Cumulative Trauma Disorders," *Ergonomics*, Vol. 36, pp. 807–831.

Klaucke, D. N., Buehler, J. W., Thacker, S. B., Parrish, R. G., Trowbridge, R. L., and Berkelman, R. L. (1988), "Guidelines for Evaluating Surveillance Systems," *Morbidity and Mortality Weekly Report*, Vol. 37, No. 5, pp. 1–18.

Kleinbaum, D. G., Kupper, L. L., and Morgenstern, H. (1982), *Epidemiologic Research*, Van Nostrand Reinhold, New York.

Knox, K. and Moore, J.S., (2001), "Predictive Validity of the Strain Index in Turkey Processing," *Journal of Occupational and Environmental Medicine*, Vol. 43, No. 5, pp. 451–462.

Kroemer, K., Kroemer, H., and Kroemer-Elbert, K. (1994), *Ergonomics: How to Design for Ease and Efficiency*, Prentice Hall, Englewood Cliffs, NJ.

Kuijpers, T., van der Windt, D. A. W. M., van der Heijden, G. J. M. G., and Bouter, L. M. (2004), "Systematic Review of Prognostic Cohort Studies on Shoulder Disorders," *Pain*, Vol. 109, pp. 420–431.

Kuorinka, I., and Forcier, L., Eds. (1995), *Work Related Musculoskeletal Disorders: A Reference Book for Prevention*, Taylor & Francis, London.

Larsson, B., Sogaard, K., and Rosendal, L. (2007), "Work Related Neck–Shoulder Pain: A Review on Magnitude, Risk Factors, Biochemical Characteristics, Clinical Picture and Preventive Interventions," *Clinical Rheumatology*, Vol. 21, No. 3, pp. 447–463.

Leino, P. (1989), "Symptoms of Stress Predict Musculoskeletal Disorders," *Journal of Epidemiology and Community Health*, Vol. 43, pp. 293–300.

Lee, R., and Higgins, G. (2006), "Report on the Proceedings of HSE's Musculoskeletal Disorder and Return to Work Workshop (WPS/06/02)," Health & Safety Laboratory, Buxton.

Li, G., Buckle, P. (1999a), *Evaluating Change in Exposure to Risk for Musculoskeletal Disorders—A Practical Tool*, CRR251, HSE Books, Suffolk.

Li, G., Buckle, P. (1999b), "Current Techniques for Assessing Physical Exposure to Work-Related Musculoskeletal Risks, with Emphasis on Posture-Based Methods," *Ergonomics*, Vol. 42, pp. 674–695.

LifeART (2000), "Grant's Atlas 10th Edition Images" [computer file], Lippincott Williams & Wilkins, Baltimore. MD.

Lluch, A. L. (1992), "Thickening of the Synovium of the Digital Flexor Tendons: Cause or Consequence of the Carpal Tunnel Syndrome?" *Journal of Hand Surgery* [*British Volume*], Vol. 17, pp. 209–212.

Lucire, Y. (2003), *Constructing RSI: Belief and Desire*, University of New South Wales Press, Sydney.

Lundberg, U. (2002), "Psychophysiology of Work: Stress, Gender, Endocrine Response, and Work-Related Upper Extremity Disorders," *American Journal of Industrial Medicine*, Vol. 41, pp. 383–392.

Lundberg, G., Dahlin, L. B., Hansson, H. A., Kanje, M., and Necking, L. E. (1990), "Vibration Exposure and Peripheral Nerve Fiber Damage," *Journal of Hand Surgery* [*American Volume*], Vol. 15, pp. 346–351.

Luopajarvi, T., Kuorinka, I., Viroiamen, M., and Holmberg, M. (1979), "Prevalence of Tenosynovitis and Other Injuries of the Upper Extremities in Repetitive Work," *Scandinavian Journal of' Work, Environment and Health*, Vol. 5, Suppl. 3, pp. 48–55.

MacDonald, L. A., Karasek, R. A., Punnets, L., and Scharf, T. (2001), "Covariation between Workplace Physical and Psychosocial Stressors: Evidence and Implications for Occupational Health Research and Prevention," *Ergonomics*, Vol. 44, pp. 696–718.

Macfarlane, G. J., Hunt, I. M., and Silman, A. J. (2000), "Role of Mechanical and Psychosocial Factors in the Onset of Forearm Pain: Prospective Population Based Study," *British Medical Journal*, Vol. 321, pp. 1–5.

Mackinnon, S. E., Dellon, A. L., Hudson, A. R., and Hunter, D. A. (1984), "Chronic Nerve Compression: An Experimental Model in the Rat," *Annals of Plastic Surgery*, Vol. 13, pp. 112–120.

Mallen, C. D., Peat, G., Thomas, E., Dunn, K. M., and Croft, P. R. (2007), "Prognostic Factors for Musculoskeletal Pain in Primary Care: A Systematic Review," *British Journal of General Practice*, Vol. 57, pp. 655–661.

Marhold, C., Linton, S. J., and Melin, L. (2001), "A Cognitive-Behavioral Return-to-Work Program: Effects on Pain Patients with a History of Long-Term Versus Short-Term Sick Leave," *Pain*, Vol. 91, pp. 155–163.

Marras, W. S., and Schoenmarklin, R. W. (1993), "Wrist Motions in Industry," *Ergonomics*, Vol. 36, pp. 341–351.

McAtamney, L., and Corlett, E. (1993), "RULA: A Survey Method for the Investigation of Work-Related Upper Limb Disorders," *Applied Ergonomics*, Vol. 24. pp. 91–99.

McCluskey, S., Burton, A. K., and Main, C. J. (2006), "The Implementation of Occupational Health Guidelines Principles for Reducing Sickness Absence Due to Musculoskeletal Disorders," *Occupational Medicine* (*London*), Vol. 56, pp. 237–242.

McDonough, J. T., Jr., Ed. (1994), *Stedman's Concise Medical Dictionary*, Williams & Wilkins, Baltimore.

Meijer, E. M., Sluiter, J. K., and Frings-Dresen, M. H. W. (2005), "Evaluation of Effective Return-to-Work Treatment Programs for Sick Listed Patients with Non-Specific Musculoskeletal Complaints: A Systematic Review," *International Archives of Occupational and Environmental Health*, Vol. 78, pp. 523–532.

Melhorn, J. M. (2005), "Working with Common Upper Extremity Problems" in *A Physician's Guide to Return to Work*, J. B. Talmage and J. M. Melhorn, Eds., AMA Press, Chicago, IL, pp. 181–213.

Monnington, S., Quarrie, C., Pinder, A., and Morris, L. (2003), "Development of Manual Handling Assessment Charts (MAC) for Health and Safety Inspectors," in *Contemporary Ergonomics*, T. McCabe, Ed., Taylor & Francis, London, pp. 3–8.

Moore, A., Wells, R., and Ranney, D. (1991), "Quantifying Exposure in Occupational Manual Tasks with Cumulative Trauma Disorder Potential," *Ergonomics*, Vol. 34, pp. 1433–1453.

Moore, J. S. (2002), "Biomechanical Models for the Pathogenesis of Specific Distal Upper Extremity Disorders," *American Journal of Industrial Medicine*, Vol. 41, pp. 353–369.

Moore, J. S., and Garg, A. (1995), "The Strain Index: A Proposed Method to Analyze Jobs for Risk of Distal Upper Extremity Disorders," *American Industrial Hygiene Association Journal*, Vol. 56, pp. 443–458.

Moore, J. S., Rucker, N. P., and Knox, K. (2001), "Validity of Generic Risk Factors and the Strain Index for Predicting Non-Traumatic Distal Upper Extremity Morbidity," *American Industrial Hygiene Association Journal*, Vol. 62, No. 2, pp. 229–235.

Moore, J.S., et al. (2006), "The Validity and Reliability of the Strain Index," in *Proceedings of IEA 2006: 16th World Congress on Ergonomics*, Maastricht, Netherlands.

Nakamichi, K., and Tachibana, S. (1992), "Transverse Sliding of the Median Nerve Beneath the Flexor Retinaculum," *Journal of Hand Surgery* [*British Volume*], Vol. 17, pp. 213–216.

Nash, C. E., Mickan, S. M., Del Mar, C. B., and Glasziou, P. P. (2004), "Resting Injured Limbs Delays Recovery: A Systematic Review," *The Journal of Family Practice*, Vol. 53, pp. 706–712.

National Health and Medical Research Council (NHMRC) (2004), "Evidence-based Management of Acute Musculoskeletal Pain: A Guide for Clinicians," NHMRC, Australian Academic Press, Bowen Hills, Queensland.

National Institute for Occupational Safety and Health (NIOSH) (1977), *National Occupational Hazard Survey, 1972-1974*, Publication 78-114, NIOSH, Washington, DC.

National Institute for Occupational Safety and Health (NIOSH) (1986), *Proposed National Strategies for the Prevention of Leading Work-Related Diseases and injuries*, Part 1, PB87-114740, NIOSH, Washington, DC.

National Institute for Occupational Safety and Health (NIOSH) (1996), *National Occupational Research Agenda*, NIOSH, available: http://www2a.cdc.gov/NORA/default.html, accessed July 18, 2004.

National Institute for Occupational Safety and Health (NIOSH) (1997), "Musculoskeletal Disorders and Workplace Factors: A Critical Review of Epidemiologic Evidence for Work-Related Musculoskeletal Disorders of the Neck, Upper Extremity, and Low Back," NIOSH, Cincinnati, OH.

National Institute for Occupational Safety and Health (NIOSH) (2001), "National Occupational Research Agenda for Musculoskeletal Disorders", NIOSH, available: http://www.cdc.gov/niosh/docs/2001-117/pdfs/2001-117pdf, accessed October 26, 2011.

National Research Council (NRC) (1999), "Work-Related Musculoskeletal Disorders" in *Report, Workshop Summary and Workshop Papers*, National Academy Press, Washington, DC.

National Research Council (NRC) (2001), *Musculoskeletal Disorders and the Workplace*, National Academy of Sciences, Washington, DC.

Neuman, W., Wells, R., Norman, R., Kerr, M., Frank, J., and Shannon, H. (2001), "Trunk Posture: Reliability, Accuracy, and Risk Estimates of Low Back Pain from a Video Based Assessment Method," *International Journal of Industrial Ergonomics*, Vol. 28, pp. 355–365.

Norregaard, J., Jacobsen, S., and Kristensen, J. H. (1998), "A Narrative Review on Classification of Pain Conditions of the Upper Extremities," *Scandinavian Journal of Rehabilitation and Medicine*, Vol. 31, pp. 153–164.

Norregaard, J., Jacobsen, S., and Kristensen, J. H. (1999), "A Narrative Review on Classification of Pain Conditions of the Upper Extremities," *Scandinavian Journal of Rehabilitation and Medicine*, Vol. 31, pp. 153–164.

Occhipinti, E. (1998), "OCRA: A Concise Index for the Assessment of Exposure to Repetitive Movements of the Upper Limb," *Ergonomics*, Vol. 41, pp. 1290–1311.

Oxenburgh, M. S. (1997), "Cost-Benefit Analysis of Ergonomics Programs," *American Industrial Hygiene Association Journal*, Vol. 58, pp. 150–156.

Palmer, K. T., and Smedley, J. (2007), "Work Relatedness of Chronic Neck Pain with Physical Findings," *Scandinavian Journal of Work, Environment & Health*, Vol. 33, pp. 165–191.

Pope, D., Silman, A., Cherry, N., Pritchard, C., and MacFarlane, G. (1998), "Validity of a Self Completed Questionnaire Measuring the Physical Demands of Work," *Scandinavian Journal of Work Environment and Health*, Vol. 24, pp. 376–385.

Pransky, G., Robertson, M. M., and Moon, S. D. (2002), "Stress and Work Related Upper Extremity Disorders: Implications for Prevention and Management," *American Journal of Industrial Medicine*, Vol. 41, pp. 443–455.

Putz-Anderson, V. (1993), *Cumulative Trauma Disorders: A Manual for Musculoskeletal Diseases of the Upper Limbs*, Taylor & Francis, London.

Ranney, D., Wells, R., and Moore, A. (1995), "Upper Limb Musculoskeletal Disorders in Highly Repetitive Industries: Precise Anatomical Physical Findings," *Ergonomics*, Vol. 38, pp. 1408–1423.

Rempel, D., Evanoff, B., Amadio, P. C., de Krom, M., Franklin, G., Franzblau, A., Gray, R., Gerr, F., Hagberg. M., Hales, T., Katz, J. N., and Pransky, G. (1998), "Consensus Criteria for the Classification of Carpal Tunnel Syndrome in Epidemiologic Studies," *American Journal of Public Health*, Vol. 88, pp. 1447–1451.

Ridd, J., Nicholson, A., and Montan, A. (1989), "A Portable Microcomputer Based System of 'On-Site' Activity and Posture Recording" in *Contemporary Ergonomics*, E. Megaw, Ed., Taylor & Francis, London, pp. 366–371.

Robbins, H. (1963), "Anatomical Study of the Median Nerve in the Carpal Tunnel and Etiologies of the Carpal-Tunnel Syndrome," *Journal of Bone and Joint Surgery*, Vol. 45-A, pp. 953–966.

Rojviroj, S., Sirichativapee, W., Kowsuwon, W., Wongwiwattananon, J., Tamnanthong, N., and Jeeravipoolvarn, P. (1990), "Pressures in the Carpal Tunnel: Comparison between Patients with Carpal Tunnel Syndrome and

Normal Subjects," *Journal of Bone and Joint Surgery* [*British.*], Vol. 72, pp. 516–518

Rouse, W. B., and Boff, K. R. (1997), "Assessing Cost/Benefits of Human Factors," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, NewYork, pp. 1617–1633.

Rucker, N., and Moore, J. S. (2002), "Predictive Validity of the Strain Index in Manufacturing Facilities," *Applied Occupational and Environmental Hygiene*, Vol. 17, pp. 63–73.

Ryall, C., Coggon, D., Peveler, R., Poole, J., and Palmer K. T. (2007), "A Prospective Cohort Study of Arm Pain in Primary Care and Physiotherapy—Prognostic Determinants," *Rheumatology*, Vol. 46, pp. 508–515.

Rydevik, B., Lundborg, G., and Bagge, U. (1981), "Effects of Graded Compression on Intraneural Blood Flow," *Journal of Hand Surgery*, Vol. 6, pp. 3–12.

Sauni, R., Paakkonen, R., Virtema, P., Toppila, E., and Uitti, J. (2009), "Dose-Response Relationship between Exposure to Hand-Arm Vibration and Health Effects among Metalworkers," *Annals of Occupational Hygiene*, Vol. 53, No. 1, pp. 55–62.

Schoenmarklin, R. W., Marras, W. S., and Leurgans, S. E. (1994), "Industrial Wrist Motions and Incidence of Hand/Wrist Cumulative Trauma Disorders," *Ergonomics*, Vol. 37, pp. 1449–1459.

Schuind, F., Ventura, M., and Pasteels, J. L. (1990), "Idiopathic Carpal Tunnel Syndrome: Histologic Study of Flexor Tendon Synovium," *Journal of Hand Surgery* [*American Volume*], Vol. 15, pp. 497–503.

Seeley, P. A., and Marklin, R. W. (2003), "Business Case for Implementing Two Ergonomic Interventions at an Electric Power Utility," *Applied Ergonomics*, Vol. 34, pp. 429–439.

Selander, J., Marnetoft, S. U., Bergroth, A., and Ekholm, J. (2002), "Return to Work Following Vocational Rehabilitation for Neck, Back and Shoulder Problems: Risk Factors Reviewed," *Disability and Rehabilitation*, Vol. 24, pp. 704–712.

Siegel, D. B., Kuzman. G., and Eakins, D. (1995), "Anatomic Investigation of the Role of the Lumbrical Muscles in Carpal Tunnel Syndrome," *Journal of Hand Surgery* [*American Volume*], Vol. 20, pp. 860–863.

Silverstein, B. A., Fine, L. J., and Armstrong, T. J. (1986), "Hand Wrist Cumulative Trauma Disorders in Industry," *British Journal of Industrial Medicine*, Vol. 43, pp. 779–784.

Silverstein, B. A., Fine, L. J., and Armstrong, T. J. (1987), "Occupational Factors and Carpal Tunnel Syndrome," *American Journal of Industrial Medicine*, Vol. 11, pp. 343–358.

Silverstein B. A., Viikari-Juntura, E., Fan, Z. J., Bonauto, D. K., Bao, S., and Smith, C. (2006), "Natural Course of Nontraumatic Rotator Cuff Tendinitis and Shoulder Symptoms in a Working Population," *Scandinavian Journal of Work*, Environment & Health, Vol. 32, pp. 99–108.

Simpson, G., and Mason, S. (1995), "Economic Analysis in Ergonomics," in *Evaluation of Human Work*, 2nd ed., J. R. Wilson and E. N. Corlett, Eds., Taylor & Francis, London, pp. 1017–1037.

Sluiter, J. K., Rest, K. M., and Frings-Dresen, M. H. (2001), "Criteria Document for Evaluating the Work-Relatedness of Upper-Extremity Musculoskeletal Disorders," *Scandinavian Journal of Work, Environment and Health*, Vol. 27, pp. 1–102.

Smith, G. C. (2007), "Identifying Prevention Methods to Reduce Carpal Tunnel Syndrome Incidents: Based on Analysis of Employee Perception and by Utilizing the Strain Index," unpublished PhD Dissertation, Southern Illinois University, Carbondale, IL.

Smith, M. J., and Sainfort, P. C. (1989), "A Balance Theory of Job Design for Stress Reduction," *International Journal of Industrial Ergonomics*, Vol. 4, pp. 67–79.

Snook, S. H., Vaillancourt, D. R., Ciriello, V. M., and Webster, B. S. (1995), "Psychophysical Studies of Repetitive Wrist Flexion and Extension," *Ergonomics*, Vol. 38, pp. 1488–1507.

Sommerich, C. M., Marras, W. S., and Parnianpour, M. (1996), "A Quantitative Description of Typing Biomechanics," *Journal of Occupational Rehabilitation*, Vol. 6, pp. 33–55.

Sommerich, C. M., Marras, W. S., and Karwowski, W. (2006), "Work-Related Upper Extremity Musculoskeletal Disorders," in *Handbook of Human Factors and Ergonomics*, 3rd ed., G. Salvendy, Ed., Wiley, Hoboken, NJ, pp. 855–888.

Soslowsky, L. J., An, C. H., Johnston, S. P., and Carpenter, J. E. (1994), "Geometric and Mechanical Properties of the Coracoacromial Ligament and Their Relationship to Rotator Cuff Disease," *Clinical Orthopaedics*, Vol. 304, July, pp. 10–17.

Spielholz, P., Silverstein, B., and Stuart, M. (1990), "Reproducibility of a Self-Report Questionnaire for Upper Extremity Musculoskeletal Disorder Risk Factors," *Applied Ergonomics*, Vol. 30. pp. 429–433.

Spielholtz, P., Silverstein, B., Morgan, M., Checkoway, H., and Kaufman, J. (2001), "Comparison of Self- Report Video Observation and Direct Measurement Methods for Upper Extremity Musculoskeletal Risk Factors," *Ergonomics*, Vol. 44, pp. 588–613.

Spielholz, P., Bao, S., Howard, N., Silverstein, B., Fan, J., Smith, C., and Salazar, C. (2008), "Reliability and Validity Assessment of the Hand Activity Level Threshold Limit Value and Strain Index Using Expert Ratings of Mono-Task Jobs," *Journal of Occupational and Environmental Hygiene*, Vol. 5, No. 4, pp. 250–257.

Staal, J. B., de Bie, R. A., and Hendriks, E. J. M. (2007), "Aetiology and Management of Work-Related Upper Extremity Disorders," *Best Practice & Research Clinical Rheumatology*, Vol. 21, pp. 123–133.

Stephens, J. P., Vos, G. A., Stevens, E. M., and Moore, J. S. (2006), "Test–Retest Repeatability of the Strain Index," *Applied Ergonomics*, Vol. 37, pp. 275–281.

Stock, S.R. (1991), "Workplace ergonomic factors and the development of musculoskeletal disorders of the neck and upper limbs: a meta-analysis," *American Journal of Industrial Medicine*, Vol. 19, pp.87–107.

Szabo, R. M. (2006), "Determining Causation of Work-Related Upper Extremity Disorders," *Clinics in Occupational and Environmental Medicine*, Vol. 5, pp. 225–234.

Szabo, R. M., and Chidgey, L. K. (1989), "Stress Carpal Tunnel Pressures in Patients with Carpal Tunnel Syndrome and Normal Patients," *Journal of Hand Surgery*, Vol. 14A, pp. 624–627.

Szabo, R. M., and Gelberman, R. H. (1987), "The Pathophysiology of Nerve Entrapment Syndromes," *Journal of Hand Surgery*, Vol. 12A, pp. 880–884.

Taylor, W., and Pelmear, P. L. (1976), "Raynaud's Phenomenon of Occupational Origin: An Epidemiological Survey," *Acta Chirurgica Scandinavica. Supplementum*, Vol. 465, pp. 27–32.

Treaster, D. E., and Burr, D. (2004), "Gender Differences in Prevalence of Upper Extremity Musculoskeletal Disorders," *Ergonomics*, Vol. 47, pp. 495–526.

U.S. General Accountability Office (GAO) (1997), *Worker Protection: Private Sector Ergonomics Programs Yield Positive Results*, GAO/HERS-97-163, GAO, Washington. DC.

van der Beek, A., van Galen, L., and Frings-Dresen, M. (1992), "Working Postures and Activities of Lorry Drivers: A Reliability Study of On-Site Observation and Recording on a Pocket Computer," *Applied Ergonomics*, Vol. 23, pp. 331–336.

Van Eerd, D., Beaton, D., Cole, D., Lucas, J., Hogg-Johnson, S., and Bombardier, C. (2003), "Classification Systems for Upper-Limb Musculoskeletal Disorders in Workers: A Review of the Literature," *Journal of Clinical Epidemiology*, Vol. 56, pp. 925–936.

Verhagen, A. P., et al. (2006), "Ergonomic and Physiotherapeutic Interventions for Treating Work-Related Complaints of the Arm, Neck or Shoulder in Adults," *Cochrane Database of Systematic Reviews*, Issue 3. Art. No. CD003471.

Viikari-Juntura, E., and Silverstein, B. (1999), "Role of Physical Load Factors in Carpal Tunnel Syndrome," *Scandinavian Journal of Work, Environment and Health*, Vol. 25, pp. 163–185.

Viikari-Juntura, E., Rauas, S., Martikainen, R., Kumosa, E., Riihimaki, H., and Saarenmaa, K. (1996), "Validity of Self-Reported Physical Work Load in Epidemiological Studies on Musculoskeletal Disorders," *Scandinavian Journal of Work Environment and Health*, Vol. 22, pp. 251–259.

Visser, B., and van Dieen, J. H. (2006), "Pathophysiology of Upper Extremity Muscle Disorders," *Journal of Electromyography and Kinesiology*, Vol. 16, pp. 1–16.

Waddell, G., and Burton, A. K. (2004), "Concepts of Rehabilitation for the Management of Common Health Problems," The Stationery Office, Norwich.

Walker-Bone, K., and Cooper, C. (2005), "Hard Work Never Hurt Anyone: Or Did It? A Review of Occupational Associations with Soft Tissue Musculoskeletal Disorders of the Neck and Upper Limb," *Annals of the Rheumatic Diseases*, Vol. 64, pp. 1391–1396.

Walker-Bone, K. E., Palmer, K. T., Reading, I., and Cooper, C. (2003), "Criteria for Assessing Pain and Nonarticular Soft-Tissue Rheumatic Disorders of the Neck And Upper Limb," *Seminars in Arthritis and Rheumatism*, Vol. 33, pp. 168–184.

Walker-Bone, K., Reading, I., Coggon, D., Cooper, C., and Palmer, K. T. (2004a), "The Anatomical Pattern and Determinants of Pain in the Neck and Upper Limbs: An Epidemiologic Study," *Pain*, Vol. 109, pp. 45–51.

Walker-Bone, K., Palmer, K. T., Reading, I., Coggon, D., and Cooper, C. (2004b), "Prevalence and Impact of Musculoskeletal Disorders of the Upper Limb in the General Population," *Arthritis Rheumatism*, Vol. 51, pp. 642–651.

Wasserman, D. E., Badger, D. W., Doyle, T. E., and Margoilies, L. (1974), "Industrial Vibration: An Overview," *ASSE Journal*, Vol. 19, pp. 38–42.

Waters, T., Putz-Anderson, V., Garg, A., and Fine. L. (1993), "Revised NIOSH Equation for the Design and Evaluation of Manual Lifting Tasks," *Ergonomics*, Vol. 36, pp. 749–766.

Werner, R. A., and Andary, M. (2002), "Carpal Tunnel Syndrome: Pathophysiology and Clinical Neurophysiology," *Clinical Neurophysiology*, Vol. 113, pp. 1373–1381.

Wiktorin, C., Mortimer, M., Ekenval, L., Kilbom, A., and Hjelm, E. (1995), "HARBO, a Simple Computer-Aided Observation Method for Recording Work Postures," *Scandinavian Journal of Work*, Environment & Health, Vol. 21, pp. 440–449.

Williams, R. M., Westmorland, M. G., Schmuck, G., and Mac-Dermid, J. C. (2004), "Effectiveness of Workplace Rehabilitation Interventions in the Treatment of Work-Related Upper Extremity Disorders: A Systematic Review," *Journal of Hand Therapy*, Vol. 17, pp. 267–273.

Wilson, J. R. (1994), "Devolving Ergonomics: The Key to Ergonomics Management Programs," *Ergonomics*, Vol. 37, pp. 579–594.

Woods, V. (2005), "Work-Related Musculoskeletal Health and Social Support," *Journal of Occupational Medicine (London)*, Vol. 55, pp. 177–189.

World Health Organization (WHO) (1985), "Identification and Control of Work-Released Diseases," Technical Report 174, WHO, Geneva.

Yen, T., and Radwin, R. (1995), "A Video Based System for Acquiring Biomechanical Data Synchronized with Arbitrary Events and Activities," *IEEE Transactions on Biomedical Engineering*, Vol. 42, pp. 944–948.

Zeiss, J., Skie, M., Ebraheim, N., and Jackson, W. T. (1989), "Anatomic Relations between the Median Nerve and Flexor Tendons in the Carpal Tunnel: MR Evaluation in Normal Volunteers," *American Journal of Roentgenology*, Vol. 153, pp. 533–536.

# CHAPTER 29

# WARNINGS AND HAZARD COMMUNICATIONS

**Michael S. Wogalter**
**Department of Psychology**
**North Carolina State University**
**Raleigh, North Carolina**

**Kenneth R. Laughery, Sr.**
**Department of Psychology**
**Rice University**
**Houston, Texas**

**Christopher B. Mayhorn**
**Department of Psychology**
**North Carolina State University**
**Raleigh, North Carolina**

## 1 INTRODUCTION

Warnings are safety communications that are used to inform people about hazards and to provide instructions so as to avoid or minimize undesirable consequences such as injury of death. Warnings are used in a variety of contexts to address environmental and product-related hazards.

In the United States, interest in warnings is also associated with litigation concerns. The adequacy of warnings has become a prevalent issue in product liability and personal injury litigation. According to the Restatement of Torts (second) and to the Theory of Strict Liability, if a product needs a warning and the warning is absent or defective, then the product is defective (see, e.g., Madden, 1999).

Regulations, standards, and guidelines as to when and how to warn have been developed more extensively in the last three decades. Also, there has been a substantial increase in research activity on the topic during this time. Human factors specialists, or ergonomists, have played a major role in the research and the technical literature that has resulted.

This chapter reviews some of the major concepts and findings regarding factors that influence warning effectiveness. Most of the research review is presented in the context of a communication–human information processing (C-HIP) model. The model not only is useful for organizing research findings but also provides a predictive and investigative tool. Following the presentation of the model and the review of major concepts and findings, a collection of recommendations for designing warnings in applications is presented.

## 2 BACKGROUND

In this section several terms will be defined and the role of warnings in the broader context of hazard control will be discussed.

## 2.1 Definitions

It is important to establish a few definitions for terms that will be used in this chapter, particularly the concepts of hazard and danger. These terms are sometimes used in different ways with different meanings; hence, we want to be clear as to their meaning in this context.

*Hazard* is defined as a set of circumstances that can result in injury, illness, or property damage. Such circumstances may include characteristics of the environment, of equipment, and of a task someone is performing. From a human factors perspective, it is important to note that circumstances also include characteristics of the people involved. These people characteristics encompass abilities, limitations, and knowledge.

*Danger* is a term that is used in a variety of ways. In this chapter it is viewed as the product of hazard and likelihood; that is, if one has quantified values of hazard and likelihood, multiplying these quantities would give a value for danger. Note that an implication of this definition is that if either value is zero, there is no danger. If the hazard and its consequence are serious but will not occur, there is no danger. Similarly, if the probability of an event occurring is high but there will be no resulting undesirable consequences, there is no danger. Note, however, people commonly use the words hazard and danger interchangeably.

## 2.2 Hierarchy of Hazard Control

In the field of safety there is a concept of hazard control that includes the notion of a hierarchy (Sanders and McCormick, 1993). This hierarchy defines a sequence of approaches to dealing with hazards in order of preference. The sequence is (1) design it out, (2) guard against it, and (3) warn about it. The notion of a design solution is that the first preference is to eliminate the hazard through alternative designs. If a nonflammable solution can be used effectively for a cleaning task, then such a solution is preferable to wearing protective equipment or warning about avoiding an ignition source due to flammability. Of course, often it is not possible to eliminate hazards. Guarding, whether physical or procedural, is a second line of defense and has as its purpose preventing contact between people and the hazard. Barriers and protective equipment are examples of physical barriers while designing tasks in such a way as to keep people out of a hazard zone is an example of a procedural guard. However, like alternative designs, guarding is not always a feasible solution, and the third line of defense is warning. Warnings are third in the priority sequence because influencing behavior is sometimes difficult and seldom foolproof. There is another implication of this priority scheme; namely, warnings are not a substitute for good design or adequate guarding. Indeed, warnings are properly viewed as a supplement, not a substitute, to other approaches to safety (Lehto and Salvendy, 1995).

In addition to the above three-part hierarchy, there are other approaches that may be effective in dealing with hazards (see, e.g., Laughery and Wogalter, 2011). Generally, they fall into the same category as warnings in that they are means of influencing the behavior of people. Training and personnel selection are examples. Another approach that includes elements similar to procedural guarding and warnings is supervisory control. These latter approaches are particularly applicable to hazards in the context of employment and job performance.

## 3 WARNINGS

In this section the purpose(s) of warnings and some general criteria for warnings are discussed.

## 3.1 Purpose of Warnings

The purpose of warnings can be explained at several levels. Most generally, warnings are intended to improve safety, that is, to decrease accidents or incidents that result in injury, illness, or property damage. At another level, warnings are intended to influence or modify people's behavior in ways that will improve safety. At still another level, warnings are intended to provide information that enables people to understand hazards, consequences, and appropriate behaviors that in turn enable them to make informed decisions. This latter point places warnings as a type of communication.

There are two additional points associated with the purposes of warnings. First, warnings are sometimes used as a means of shifting or assigning responsibility for safety to people in the system, the product user, the worker, and so on, in situations where hazards cannot be designed out or adequately guarded. This point is not to say that people do not have safety responsibilities independent of warnings; of course they do. Rather, a purpose of warnings is to provide the information necessary to enable them to carry out such responsibilities. Whether responsibility has been shifted depends at least in part on the effectiveness of the communications. The second point regarding warnings' communication purpose concerns an issue that has received little attention in the technical literature, namely, people's right to know. This notion makes the point that, even in situations where the likelihood of warnings being effective may not be high, people have the right to be informed about safety problems confronting them. This aspect of warnings relates to personal, societal, and legal concerns.

## 3.2 General Criteria for Warnings

The most important general criterion for warnings is that their design should be viewed as an integral part of the overall system design process. Frantz et al. (1999) address this issue in a chapter on developing product warnings. While safety warnings are a third line of defense behind design and guarding, they should not be considered for the first time after the design (including guards) of the environment or product has already been set and established. Too many warnings are developed at this late stage of design, as an afterthought, and their quality and effectiveness often reflect it. Further, warnings based on unrealistic and untested assumptions or expectations about the target audience are destined to be inadequate.

### 3.2.1 When/What to Warn?

There are several principles or rules that guide when a warning should be used. They include:

1. A significant hazard exists.
2. The hazard, consequences, and appropriate safe modes of behavior are not known by the people exposed to the hazard.
3. The hazards are not open and obvious; that is, the appearance and function of the environment or product do not convey them.
4. A reminder is needed to assure awareness of the hazard at the proper time. This concern is especially important in situations of high task loading or potential distractions.

### 3.2.2 Who to Warn

The general principle regarding who should be warned is that it should include everyone who may be exposed to the hazard and everyone who may be able to do something about it. There are occasions when people in the latter category may not themselves be exposed to the hazard. An example would be the industrial toxicologist who receives warning information regarding a product to be used by employees and who then defines job procedures and/or protective equipment to be employed in handling the material. The physician who prescribes medications with side-effect hazards is another example.

There are, of course, situations and products where the target audience is the general public and that includes nearly everyone. Hazards in the public environment or products on the shelf of a drugstore or hardware store are examples. Other warnings may be directed to a very specific audience. Warnings about the risk of birth defects associated with taking a prescription medication would be directed primarily to women of child-bearing age; although others such as spouses or parents might also receive the warning (Mayhorn and Goldsworthy, 2007). Likewise, as noted above, health care professionals such as physicians or pharmacists should receive the warnings regarding potential birth defects when treating patients who are or may become pregnant. If warnings are to be effective, the characteristics of the target audience should be taken into account.

## 4 COMMUNICATION-HUMAN INFORMATION PROCESSING (C-HIP) MODEL

In this section a theoretical context is presented that will serve as an organizing framework or model for reviewing some of the major concepts and findings regarding factors that influence warning effectiveness. Specifically, a C-HIP model is described (Wogalter, 2006a). To place this model in context, a few general comments about communications and human information processing are in order.

***Communications*** Warnings are a form of safety communications. Communication models have been around for most of the last century (Lasswell, 1948;

Shannon and Weaver, 1949). A typical, very basic model shows a sequence starting with a source who encodes a message into a channel that is transmitted to a receiver who receives a decoded version of that message. Noise may enter into the system at several points in the sequence, reducing the correspondence between the message sent and the one received. The warning sender may be a product manufacturer, government agency, employer, and so on. The receiver is the user of the product, the worker, or any other person at risk. The message, of course, is the safety information to be communicated. The medium refers to the channels or routes through which information gets to the receiver from the sender. Understanding and improving these components of a safety communication system increases the probability that the message will be successfully conveyed.

However, the communication of warnings is seldom as simple as implied by a sequential communication model. Frequently more than one medium or channel may be available and/or involved; multiple messages in different formats and/or containing different information may be called for; and the receiver or target audience may include different subgroups with varying characteristics. An example of such a warning situation would occur when a product with associated hazards is being used in a work environment. Figure 1 illustrates a communication model that might be applicable. It shows the distribution of safety information from several entities to the receiver and that feedback may influence the kind of safety information given. It also shows that in addition to the sender (manufacturer) and receiver (end user), other people or entities may be involved such
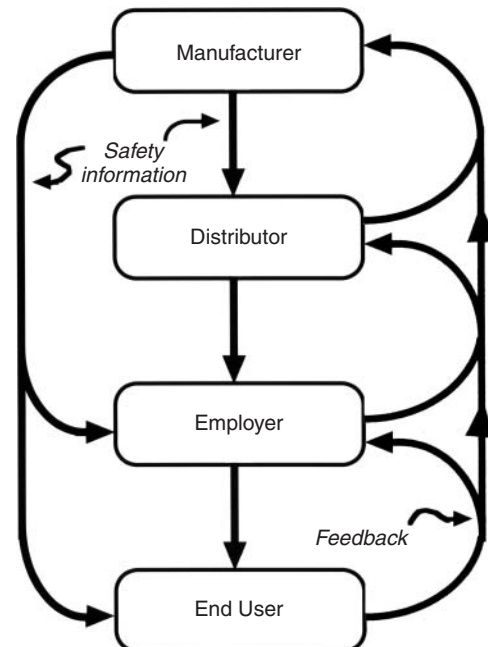


**Figure 1**   Distribution of safety information and feedback.

as distributors and employers. Further, each of these entities may be both receivers and senders of safety information. There are also more routes through which warnings may travel, such as from the manufacturer to the distributor to the employer to the user, from the manufacturer to the employer to the user, or directly from the manufacturer to the user (as on a product label). The warnings may take different forms. One example includes safety rules that an employer sets to govern the behavior of employees. Thus, warnings or warning systems may be much more complex than just a sign or label. The concepts of warning systems and indirect warnings are discussed in more detail later in the chapter.

***Human Information Processing*** Cognition is a core area of psychology that is concerned with mental processes such as attention, memory, and decision making. Since the 1960s, much of the theoretical work has been described in terms of stages of processing. Numerous models have been developed and tested. In the next section, C-HIP is described as a model that incorporates some basic stages of mental processing.

***C-HIP Model*** The C-HIP model (Wogalter, 2006a) depicted in Figure 2 is a framework for showing stages
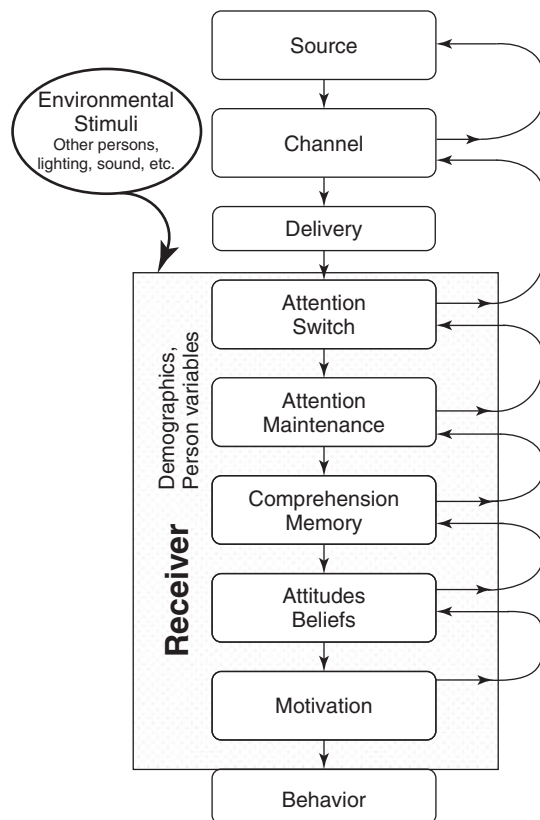


**Figure 2** The C-HIP model.

of information flow from a source to a receiver who in turn may cognitively process the information to subsequently produce compliance behavior. One of the main benefits of the C-HIP model is that it serves as a guiding framework for organizing diverse findings in the warning research literature.

At each stage of the model, warning information is processed and, if successful at that stage, "flows through" to the next stage. If processing at a stage is unsuccessful, it can produce a bottleneck, blocking the flow from getting to the next stage. If all of the stages are successful, the process ends in behavior (compliance). While the processing of the warning might not make it to the last stage, it still may be effective at influencing earlier stages. For example, a warning might positively influence comprehension but not change behavior. Such a warning cannot be said to be totally "ineffective" because it produces better understanding and can potentially lead to better, more informed decisions. However, it is ineffective in the sense that it may not curtail certain unsafe behaviors.

The C-HIP model can be particularly useful in describing the factors that influence warning effectiveness. It also can be helpful in diagnosing and understanding warning failures and inadequacies. If a source (or sender) does not issue a warning, no information will be transmitted and nothing will be communicated to the receiver. Even if a warning is issued by the source, it will not be effective if the channel or transmission medium is poorly matched with the message, the receiver, or the environment. Each of the processing stages within the receiver can also produce a bottleneck preventing further processing. The receiver might not notice the warning and thus not be directly affected. Even if the warning is noticed, the individual may not maintain attention to the warning to encode the information. If the receiver encodes the details of the warning, it still may not be understood. If understood, it still might not be believed; and so on.

Although the processing described above is linear, there are feedback loops from later stages to earlier stages as illustrated in Figure 2. For example, when a warning stimulus becomes habituated from repeated exposures over time, attention is less likely to be allocated to the warning on subsequent occasions. Here, memory (as part of the comprehension stage) affects an earlier attention stage of processing. Another example is that some people might not believe that a product or situation is hazardous, and as a consequence not look for a warning. A third example is that the person may not understand the warning and therefore might switch attention to read it again. These nonlinear effects between the stages resulting from feedback show how later stages influence earlier stages in ongoing cognitive processing.

In the sections that follow, we describe each stage of the C-HIP model and some of the factors that influence it. The purpose is to assist in analyzing how or why warnings may fail or, conversely, what they have to accomplish to succeed. In many respects the model is similar to the information processing models employed by others (Lehto and Miller, 1986; Lehto and

Papastavrou, 1993; Rogers et al., 2000). The model presented here is somewhat different than those presented in Wogalter et al. (1999b) and Wogalter and Laughery (2005). Over the years, the body of research has grown to the extent that it now requires fairly substantial books to describe and summarize the literature (e.g., Wogalter et al. 1999b; Wogalter 2006b). This chapter gives an overview of research findings relevant to each stage of C-HIP. In both Wogalter et al. (1999b) and Wogalter (2006b) there are individual detailed chapters on most of the model's stages. The model has evolved over time. The model that predated the C-HIP (Wogalter and Laughery 1996) simply presented some of the main human information processing stages (i.e., in the receiver section), in other words, only the second section of stages of the eventual C-HIP model. The Wogalter et al. (1999b) version of C-HIP added the first section from communication theory (source and channel). The most recent model from Wogalter (2006a) (shown in Figure 2) is different in four ways from Wogalter et al.'s (1999b) C-HIP model. First, in the current model the attention stage is split into two separate stages, attention switch and attention maintenance. The reason for the split is that these two stages are different and are affected by different variables. The second major difference in the models is that there is now the stage of delivery (Williamson 2006). Delivery refers to the point of warning reception where information is provided to the receiver via one or more channels. The third change in the current model is an explicit reference to the influence of other environmental stimuli. Environmental influences are aspects other than the warning itself that could affect how the warning is processed. They are extrinsic to the warning. Environmental influences can include other information on the product label, the product itself, other people's involvement, other warnings, and other aspects in the environment, including illumination and background noise (Vredenburgh and Helmick-Rich, 2006). The fourth major change from the Wogalter et al. (1999b) C-HIP model to the current model is greater emphasis on the receiver's personal characteristics (e.g., demographics) and task involvement (Smith-Jackson, 2006b; Smith-Jackson and Wogalter, 2007; Wogalter and Usher, 1999). Both the third and the fourth changes serve to emphasize how context (outside the person and warning and internal aspects of the target person) can influence the processing of warning content.

Table 1 shows a summary of some of the primary considerations associated with successful processing at each stage.

### 4.1   Source

The source is the originator or initial transmitter of the warning information. The source can be a person(s) or an organization (e.g., company or government). Research shows that differences in the perceived characteristics of the source can influence people's beliefs about the credibility and relevance of the warning (Wogalter et al., 1999c). Information from a reliable, expert source [e.g., the Surgeon General, the U.S. Food and Drug Administration (FDA)] is given greater credibility, particularly when the expertise is relevant

(e.g., the American Medical Association and the FDA for a health-related warning) (Wogalter et al., 1999c). Indeed, Internet users are more likely to believe facts they encounter on websites that have domain suffixes such as .edu and .gov than .com because they are from educational- or governmental-related sources as opposed to for-profit companies (Wogalter and Mayhorn, 2008). An important aspect that will be discussed in more detail later is that a warning attributed to an expert source may aid in changing erroneous beliefs and attitudes that the receiver may have.

A critical role of the source is to determine if there is a need for a warning and, if so, what should be warned. This decision typically hinges on the outcomes of hazard analyses that determine foreseeable ways injuries could occur. Assuming that the product or environment has been determined to need a warning, one or more communications channels must be used to reach the receiver.

### 4.2   Channel

The channel is the medium in which information is transmitted from the source to one or more receivers. In the past, most warnings have been presented on product labels, on posters, or in brochures. These traditional methods of "static" display will be enhanced through the use of technology-based dynamic displays in the future. Future warning systems will likely have properties that are different and better than those inherent in traditional static warnings [see Wogalter and Mayhorn (2005a) for a review]. For example, computers and sensors can be used to process information to enable warnings to be appropriately tailored to the situation and characteristics of the target user (Wogalter and Mayhorn, 2005a). Whether communicated via traditional static or technology-based dynamic media, warnings are often sent via the visual (printed text warnings and pictorial symbols) and auditory (alarm tones, live voice, and voice recordings) modalities as opposed to the other senses. There are exceptions: An odor added to flammable gases such as propane (LP) or natural gas can make use of the olfactory sense, and a pilot's control stick that is designed to vibrate when the aircraft begins a potentially dangerous stall makes use of the tactile, haptic, and kinesthetic senses.

***Media and Modality***   There are two basic dimensions of the channel. The first concerns the media in which the information is embedded. The second dimension of the channel is the sensory modality used to capture the information by the receiver. Media and modalities are closely tied. Some studies have examined whether presentation of a language-based warning is more effective when presented in the visual (text) versus the auditory (speech) modality. The results are conflicting (although generally either one is better than no presentation whatsoever). Some cognitive research (Penney, 1989) suggests that longer, more complex messages may be better presented visually and shorter messages auditorily. The auditory modality is usually better for attracting attention (a stage described below). However, auditory presentation can be less effective than

**Table 1 Methods and Influences of C-HIP Stages**

| C-Hip Stage | Methods and Influences |
|---|---|
| **Source** | • Determines that hazard is not designed out or guarded<br>• Credible, expert |
| **Channel** | Sensory modality<br>Visual (signs, labels, tags, inserts, product manuals, video, etc.)<br>Auditory (simple and complex nonverbal; voice; live or recorded)<br>Other senses: vibration, smell, pain<br>Generally, transmission in more than one modality is better.<br>Media<br>Print (label, manual, brochure, magazine advertisement sign)<br>Voice (radio, live), Video (TV), Internet |
| **Delivery** | Make sure message gets to target audience(s).<br>Did it arrive to one or more of the receiver's sensory modalities? |
| **Receiver** | Consider demographics of target audiences (e.g., older adults, illiterates, cultural and language differences, persons with sensory impairments). |
| **Attention switch** | Should be high salience (conspicuous/prominent) in cluttered and noisy environments (e.g., using distinctive color, motion/movement)<br>Visual: high contrast, large<br>Presence of pictorial symbols and other graphics can aid noticeability.<br>Auditory: louder and distinguishable from surround<br>Present when and where needed (placed proximal in time and space)<br>Avoid habituation by changing stimulus.<br>Measurement: recording eye and head movements |
| **Attention maintenance** | Enables message encoding by examining/reading or listening<br>Visual: legible font and symbols, high-contrast aesthetic formatting, brevity<br>Auditory: intelligible voice, distinguishable from other sounds<br>Measurement: duration of looking/listening and subsequent recall and recognition |
| **Comprehension and memory** | Enables informed judgment<br>Understandable message that provides necessary and complete information to avoid hazard<br>Try to relate information to knowledge already in users' heads.<br>Explicitness enables elaborative rehearsal and storage of information.<br>Pictorials can benefit understanding and substitute for some wording; may be useful for certain demographic groups (low literates or unskilled in language).<br>At subsequent exposures, warning can cue or remind user of information.<br>Comprehension testing needed to determine whether warning communicates intended/needed information<br>Measurement: Testing understanding of intended message after exposure: Does it communicate all of the intended necessary information? |
| **Beliefs/attitudes** | Perceived hazard and familiarity are beliefs that affect warning processing.<br>Persuasive argument and prominent warning design are needed when beliefs are discrepant with truth so as to appropriately alter those beliefs.<br>Can have influence receiver's earlier stages<br>Measurement: Determine beliefs (pre- and post-). |
| **Motivation** | Energizes person to carry out next stage (behavior)<br>Perceived low cost (time, effort, money) facilitates compliance.<br>Perceived high cost of compliance increases likelihood of noncompliance.<br>Motivation benefited by explicitness and perceived injury severity.<br>Affected by social influence, time stress, mental workload<br>Measurement: Ratings of willingness to carry out the directed behavior |
| **Behavior** | Carrying out safe behavior that does not result in injury or property damage<br>Measurement: Behavioral compliance |

visual presentation, particularly for processing lengthy, complex messages because (a) of its primary temporal/sequential nature, (b) its processing speed is slower, and (c) the ability to review previously presented material is often not possible. These characteristics tend to overload working memory (or maintenance attention, to be discussed later).

***Multiple Methods and Redundancy***   Research has generally found that presenting warnings in two modalities is better than one modality. Thus, a warning is better if the words are shown on a visual display while at the same time the same information is given orally. This provides redundancy. Together they can be beneficial as it provides a way for persons who may be occupied on a task involving attention to one or the other modality to be alerted by the warning. If an individual is not watching the display, people can still hear it. Or, if an individual is listening to something else (or is wearing hearing protection), they could potentially see the message on the visual display. Also, if the individual is blind or deaf, the information is available in the other modality. A similar concept for media is described below.

***Warning System***   The idea that a warning is only a sign or a portion of a label is too narrow a view of how safety information gets transmitted. Warning systems for a particular environment or product may consist of a number of components. In the context of the communication model presented in Figure 1, the components may include a variety of media and messages.

A warning system for a pharmaceutical product such as a prescription allergy medication may consist of several components: a verbal warning from a physician, a printed statement on the box, a printed statement on the bottle, and a printed package insert. In addition, there may be text and/or speech warnings in television and radio advertisements that specifically target consumers. In the United States, direct-to-consumer (DTC) advertisements about prescription pharmaceuticals usually include warnings about side effects and contraindications. Due to the brevity of most broadcast commercials, these DTC ads frequently direct people to other sources of information such as manufacturer websites or a toll-free telephone number (Goldsworthy and Mayhorn, 2010; Kim et al., 2010; Vigilante et al., 2007). Likewise, a warning system for pneumatic tools regarding the hazard of long-term vibration exposure causing damage to the nervous and vascular systems of the hand (vibration-induced white finger) might consist of a number of components. Examples include warnings embossed on the tool, a removable tag attached to the product when new, accompanying sheets or a stapled manual, and printing on the box. In addition, manufacturers might provide employers with supplemental materials such as videos and posters to assist in employee training sessions. Organizations including government agencies and consumer and trade groups could provide additional materials via mail or the Internet. Yet another example would be warnings for a solvent used in a work environment

for cleaning parts. Here the components might include warnings printed on labels of the container, printed flyers that accompany the product, and material safety data sheets (MSDSs) provided to employers. They might also include statements in advertisements about the product and verbal statements from the salesperson to a purchasing agent.

The components of a warning system may not be identical in terms of content or purpose. For example, some components may be intended to capture attention and direct the person to another component where more information is presented. Similarly, different components may be intended for different target audiences. In the above solvent example, the label on the product container may be intended for everyone associated with the use of the product, including the end user, while the information in the MSDS may be directed more to fire personnel or to an industrial toxicologist or safety engineer working for the employer (Smith-Jackson and Wogalter, 2007).

***Direct and Indirect Communications***   The distinction between direct and indirect effects of warnings concerns the routes by which information gets to the target person. A direct effect occurs as a result of the person being directly exposed to the warning. That is, he or she directly reads or hears the warning. But warnings can also accomplish their purposes when delivered indirectly (Wogalter and Feng, 2010). One example gleaned from research by Tam and Greenfield (2010) suggests that the indirect effects associated with alcoholic-beverage warnings may explain gender differences in the likelihood to intervene to prevent others from driving while intoxicated. The employer or physician who reads warnings and then verbally communicates the information to employees or patients is also an example. Moreover, the print and broadcast news media may present information that is given in warning labels. The point is that a warning put out by a manufacturer may have utility even if the consumer or user is not directly exposed to the warning.

An example of where an indirect effect was considered in the design of a product warning concerned a herbicide used in agricultural settings. Given that significant numbers of farm workers in parts of the United States read Spanish but not English, there was reason to put the warning in both languages. However, there are sometimes space constraints on product containers. One suggested strategy was to include a short statement on the label in Spanish indicating that the product was hazardous and that the user should get someone to translate the rest of the label before using the product. There are also other ways to increase surface area to print additional warning material, some of which are described later.

There are situations where we rely on indirect communications to transmit warning information. Employers and physicians are examples already noted; however, adults who have responsibility for the safety of children are another important category (Mayhorn et al., 2006). In the design of warning systems, empowering indirect warnings could enhance the spread of warning information to relevant targets.

## 4.3 Delivery

While the source may try to disseminate warnings in one or more channels, the warnings might not reach some of the targets at risk. For example, a safety brochure that is developed and produced by a governmental agency that is never distributed is not very helpful. Purchasers of used products are at risk because the manufacturer's product manual is frequently not available or is not transferred to new owners at resale (Rhoades et al., 1991; Wogalter et al., 1998b). For example, without the manual, the user may not know what the correct and incorrect uses of the product are or what the maintenance schedule is, which could impact safety. Williamson (2006) describes issues associated with communicating warnings on the flash-fire hazard associated with burning plastic-based insulation. Although there are some warnings accompanying bulk lots of the insulation when shipped from the manufacturer/distributor to job sites and some technical warnings that may be seen by architects and high-level supervisors, the warnings infrequently make it downstream to construction workers who may be working with or around the product. Likewise, prescription medications that are shared with others may not be seen in the original containers that include warnings regarding side effects (Goldsworthy et al., 2008b). The point here is that while a warning may be put out by a source (through some channel) it may have limited utility if it does not reach the targets at risk either directly or indirectly.

## 4.4 Receiver

In this section the focus is on the receiver, that is, the person(s) or target audience to whom the warning is directed. As noted earlier, the primary theoretical context for presenting this analysis is an information processing model. This model with respect to the receiver, shown in Figure 2, defines a sequence of processing stages through which warning information flows. By examining each of the stages and the factors that influence success or failure at each stage, a better understanding of how warnings should be designed and whether they are likely to be effective can be attained.

For a warning to effectively communicate information and influence behavior, attention must be switched to it and then maintained long enough for the receiver to extract the necessary information. Next, the warning must be understood and must concur with the receiver's existing beliefs and attitudes. If there is disagreement, the warning must be sufficiently persuasive to evoke an attitude change toward agreement. Finally, the warning must motivate the receiver to perform proper compliance behavior. The next several sections are organized around these stages of information processing.

### 4.4.1 Attention

One of the goals of a warning is to capture attention and then hold it long enough for the contents to be processed. The following sections address these two attention issues.

*Attention Switch* The first stage in the human information processing portion of the C-HIP model concerns the switch of attention. An effective warning must initially attract attention. Often this attraction must occur in environments where other stimuli are competing for attention.

For a warning to capture attention it must first be available to the recipient. As noted earlier, warning messages will not have direct effects if they are not received by the end user. Assuming the warning is present, it needs to be sufficiently salient (conspicuous or prominent) to capture attention. Warnings typically have to compete for attention, and several design factors influence how well they compete.

*Size and Contrast* Bigger is generally better. Increased print size and contrast against the background have been shown to benefit subsequent recall (Barlow and Wogalter, 1993). Young and Wogalter (1990) found that print warnings with highlighting and bigger, bolder print led to higher comprehension of and memory for owner's manual warnings.

Context plays an important role with regard to size effects on salience. What is important is not just the size of the warning but also its size relative to other information in the display. A bold warning on a product label where there are other informational items in larger print is less likely to be viewed than those larger items.

For some products, the available surface area on which warnings can be printed is limited. This is particularly true for small product containers such as pharmaceuticals. Methods available to increase the surface area for print warnings include adding tags or peel-off labels (Barlow and Wogalter, 1991; Wogalter et al., 1999d). Another method is to put some minimum critical information on a primary label and direct the user to additional warning information in a secondary source, such as available in a well-designed owner's manual or website. Wogalter et al. (1995) have shown such a procedure can sometimes be effective.

*Color* While there are some problems with the use of color such as color blindness, fading, and lack of contrast with certain other colors, good use of color can benefit warnings. Coloration can help a warning attract attention more effectively than a warning that is the same color as its surroundings, including other text around it (e.g., Laughery et al., 1993b). The ANSI(2006) Z535.2 and Z535.4 standard for signs and labels uses color in the signal word panel.

*Pictorial Symbols* Pictorial symbols and icons can be useful for attracting attention (Bzostek and Wogalter, 1999; Jaynes and Boles, 1990; Kalsher et al., 1996; Mayhorn et al., 2004b; Mayhorn and Goldsworthy, 2009; Young and Wogalter, 1990). A common icon used in warnings that can help attract attention is the alert icon (triangle enclosing an exclamation point) (Laughery et al., 1993a) that is found in the signal word panel in the ANSI Z535 style warnings.

*Placement* A general principle is that warnings located close to the hazard both physically and in time

will increase the likelihood of a proper attention switch (Frantz and Rhoades, 1993; Wogalter et al., 1995). A warning on the battery of a car regarding a hydrogen gas explosion hazard is much more likely to be effective than a similar warning embedded somewhere in the middle of a vehicle owner's manual. A verbal warning given two days ago before a farm worker uses a hazardous pesticide is less likely to be remembered and effective than one given immediately prior to using the product.

A warning, even a good one that is located in an out-of-view location, drastically reduces its likely effectiveness. In general, placement of warnings directly on a hazardous product is preferred (Wogalter et al., 1987). However, this cannot always be done given the product and the circumstances of use. There are several factors to be considered in warning placement. One is visibility; a warning should be placed so that users are likely to see it (Frantz and Rhoades, 1993). For example, a warning on a hard drive installed inside a computer will not be seen if the user does not open the interior panel of the computer. People generally do not read owner's manuals of cars they rent; thus, unless warned some other way, such as on a dashboard placard or in a quick-tip chart, drivers will not be made aware of certain safety information. Manufacturers need to consider how their product may be used, so they can select proper locations for warnings. In general, warnings should be located near where they are needed both in proximal location and in time. Task analyses are likely to be beneficial here.

Warnings should preferably be placed before or above the instructions for use. Warnings should not be buried in the middle of other text or on a later page. Wogalter et al. (1987) showed warnings in a set of instructions for mixing chemicals were more likely to be noticed and complied with if placed before the task instructions than following them.

Sometimes practical considerations limit the available options. A small container for some over-the-counter medications may simply not have the space for all of the necessary warning information. Some options for addressing this problem are discussed later.

**Formatting**    Another factor that can influence attention is formatting. Aesthetically pleasing warning text, with plenty of white space and coherent information groupings (Hartley, 1994), are more likely to attract and hold attention (Wogalter and Vigilante, 2003). If a warning contains a large amount of dense text, individuals may decide too much effort is required to read it and thus may decide to direct their attention to something else.

**Repeated Exposure**    A related issue is that repeated and long-term exposure to a warning may result in a loss of attention capturing ability (Wogalter and Laughery, 1996). This habituation can occur over time, even with well-designed warnings. Where feasible, changing a warning's format or content can slow the habituation process (Wogalter and Brelsford, 1994). Such efforts to combat habituation may be accomplished through the use of technology-based dynamic warnings where warning content and format can be changed as needed (Wogalter and Mayhorn, 2005a). For example, electronic highway safety signs that change to dynamically report on actual specific information about real-time traffic flow and the presence of construction, vehicular crash, or flooding ahead are probably much more effective in eliciting more informed and better decisions than a general static sign saying "Traffic Congestion Ahead." More about habituation will be described in a later section.

**Other Environmental Stimuli**    Other stimuli in the environment may compete with the warning for attention capture. These stimuli may include the presence of other persons, various objects that comprise the context, and the tasks being performed. Thus, the warning must stand out from the background (i.e., be salient or conspicuous) to be more likely noticed. This factor is particularly important because people typically do not actively seek hazard and warning information. Usually people are focused on the tasks they are trying to accomplish. Because safety considerations are not always on one's mind, warnings need to be prominent.

**Auditory Warnings**    Auditory warnings are frequently used to attract attention. Auditory signals are omnidirectional, so the receiver does not have to be looking at a particular location to be alerted. Like print warnings, their success in capturing attention is largely a matter of salience. Auditory warnings should be louder and distinctively different from expected background noise. Auditory warnings are sometimes used in conjunction with visual warnings, with the auditory warning serving to call attention to the need to examine a visual warning with more specific information.

**Sensor Technology**    In some instances, hazards or indications of hazards are outside the range of human sensory perception, leaving persons at risk unaware of the danger without some additional means of detection. One example is detecting carbon monoxide gas; it its pure form, it has no odor. Technology has enabled sensors capable of detecting the presence of carbon monoxide gas as well as other gases such as propane and natural gas. There are numerous other kinds of detection systems available that can "sense" a variety of indicators such as motion, temperature, and weight. These sensors can provide input into systems that could, in turn, provide a perceptible and informative warning.

**Attention Maintenance**    Individuals may notice the presence of a warning but not stop to examine it. A warning that is noticed but fails to maintain attention long enough for its content to be encoded is of little direct value. Attention must be maintained on the message for some length of time to extract meaning from the material (Wogalter and Leonard, 1999). During this process, the information is encoded or assimilated with existing knowledge in memory.

With brief warnings the message information may be acquired very quickly, sometimes at a glance. For longer warnings to maintain attention, they need to

have qualities that generate interest and do not require considerable effort. Some of the same design features that facilitate the switch of attention also help to maintain attention. For example, large print not only attracts attention but also increases legibility, thus making reading less effortful and more likely.

*Legibility*   If the warning has very small print, it may not be legible, making it difficult to read. Some persons may not be able to read it even with visual correction and some who might be able to read it with some effort will not. Older adults with age-related vision problems are a particular concern (Wogalter and Vigilante, 2003). Distance and environmental conditions such as fog, smoke, and glare can negatively affect legibility.

Sanders and McCormick (1993) give data on legibility of fonts developed for military applications. Legibility of type can be affected by numerous factors, including choice of font, stroke width, letter compression and distance between them, case, resolution, and justification. There is not much research to support a clear preference for certain fonts over others; the general recommendation is to use relatively plain, familiar fonts. It is sometimes recommended that a serif font, with embellishments in the lettering, such as Times Roman be used for small point sizes containing message text and sans serif font (plain fonts without embellishments) such as Helvetica be used in applications requiring larger point size headline-type text. The American National Standards Institute's (ANSI, 2006) Z535.2 and Z535.4 warning sign and label standard include a chart of print size and expected reading distances in good and degraded conditions.

Contrast and color are other considerations. Black on white or the reverse has the highest contrast, but legibility can be adequate with other combinations such as black print on yellow and white print on red. The selection of color should also be governed by the context in which the warning is presented (Young, 1991). One would not want to use a red warning on a largely red background.

*Formatting*   Visual warnings formatted to be aesthetically pleasing are more likely to hold attention (and thus examined and the information extracted) than a single chunk of dense text (Vigilante and Wogalter, 2003). Formatting can show the organization of the warning material, making it easier to assimilate or accommodate into memory. In general, the use of generous white space and bold bulleted lists are preferred to long, dense prose text (e.g., Desaulniers, 1987; Wogalter and Post, 1989). While aesthetically pleasing at a distance, full justification (straight alignment at both margins) is more difficult to read than "ragged right" justification (straight alignment only at the left margin) because the spacing between letters and words is consistent, thus aiding saccadic movement during reading.

*Pictorial Symbols*   Interest is also facilitated by the presence of well-designed pictorial symbols. Further, research indicates people prefer warnings that have a pictorial symbol to warnings without one (Kalsher et al.,

1996; Mayhorn and Goldsworthy, 2009; Young et al., 1995).

*Auditory*   Simple nonverbal auditory warnings are often used as alert (attention-getting) signals. Frequently, these signals carry very little information other than an attention-switch cue. After the alert is given, the visual modality is usually used to access further information (Sanders and McCormick, 1993; Sorkin, 1987).

### 4.4.2   Comprehension

Warning comprehension concerns understanding its meaning. Some comprehension may derive from subjective understanding such as its hazard connotation given it appearance and presentation and some from the specific language and the symbols used. The processes involve people's existing memory and knowledge together with the warning and contextual stimulation.

*Hazard Connotation*   The idea of hazard connotation is that certain aspects of the warning may convey some level or degree of hazard. It is an overall perception of risk, a subjective understanding of the danger conveyed by the warning components. A similar type of connoted hazard was shown in research by Wogalter et al. (1997) for various container types.

In the United States, current standards such as ANSI (2006) Z535 and guidelines (e.g., FMC Corporation, 1985; Westinghouse Electric Corporation, 1981) recommend that warning signs and labels contain a signal word panel that includes one of the terms DANGER, WARNING, or CAUTION. According to ANSI Z535, these terms are intended to denote decreasing levels of hazard, respectively. Figure 3 shows two ANSI-type warning signal word panels. According to ANSI Z535, the DANGER panel should be used for hazards where serious injury or death *will* occur if warning compliance behavior is not followed, such as around high-voltage electrical circuits. The WARNING panel (not pictured) is used when serious injury *might* occur, such as severe chemical burns or exposure to highly flammable gases. The CAUTION panel is used when less severe personal injuries or damage to property might occur, such as getting hands caught in operating



**Figure 3**   Examples of two signal word panels including alert symbol and color. Note that the DANGER panel is white print on red background and the CAUTION is black print on yellow background. Not shown is the WARNING panel, which is black print on orange background.

equipment. Research shows that lay persons often fail to differentiate between CAUTION and WARNING, although both are interpreted as connoting lower levels of hazard than DANGER (e.g., Wogalter and Silver, 1995). The term NOTICE is intended for messages that are important but do not relate to injuries. The term DEADLY, which has been shown in several research studies to connote hazard significantly above DANGER, has not been adopted by the ANSI, yet it might be considered for hazards that are significantly above those connoted by the term DANGER.

Different characteristics of sounds can lead to different hazard connotations. Higher frequency (higher pitch) and greater amplitude (louder), which have faster repetitions, are perceived as more urgent (Edworthy et al., 1991). Similar effects have been shown with verbal speech (Barzegar and Wogalter, 1998; Hellier et al., 2002; Hollander and Wogalter, 2000; Weedon et al., 2000).

In the ANSI warning's top panel, the signal words DANGER, WARNING, AND CAUTION are assigned to a paired color (red, orange, and yellow, respectively). This assignment is a method of redundancy, which is useful if one cannot read or cannot perceive the color. However, the colors for WARNING (with its color pair orange) and CAUTION (with its color pair yellow) are not readily distinguished with regard to hazard connotation. Nevertheless, DANGER (with its color pair red) is consistently judged as having a higher hazard connotation (as measured by ratings) than the other two signal word–color combinations (e.g., Chapanis, 1994; Mayhorn et al., 2004c).

*Competence*   There are many dimensions of receiver competence that may be relevant to the design of warnings. For example, sensory deficits might be a factor in the ability of some special target audiences to be directly influenced by a warning. A blind person would not be able to receive a written warning, nor would a deaf person receive an auditory warning. A person who is illiterate would not be able to read the warning text.

At the opposite end of the sequence of events is behavior. If special equipment is required to comply with the warning, it must be available or at least easily obtainable. If special skills are required, they must be present in the receiver population. It is not difficult to find examples of warnings that violate considerations of people's limitations. One example is the common warning instruction found on containers of solvents: "Avoid breathing fumes." This might be difficult to carry out for several reasons. One reason is difficulty in detecting fumes, particularly if one cannot see or smell them (e.g., if one has nasal congestion). A second reason pertains to behavior with respect to personal protection equipment. If a respirator with an independent air supply is not available, then avoidance may be difficult.

Three characteristics of receivers related to cognitive competence are important in warning design: technical knowledge, language knowledge, and reading skill. The communication of hazards associated with medications, chemicals, and mechanical devices is complex and technical in nature. If the target audience does not have the relevant technical competence needed to interpret the information, a warning concerning hazards in these domain is likely to be unsuccessful. The level of knowledge and understanding of the audience must be taken into account. This point will be discussed further in a later section.

The issue of language is straightforward, and it is increasingly important. Subgroups in the United States speak and read languages other than English, such as Spanish. As trade becomes increasingly international, requirements for warnings to be directed to users of different languages will increase. Potential ways to deal with this problem include use of multiple languages and pictorials (Lim and Wogalter, 2003).

Reading skills and capabilities in the population vary from illiteracy to graduate-level skills. Yet, high reading levels such as a grade 12 (high school graduation level) are common in warnings that are also intended for individuals who have low-level reading skill. In general, the reading level of at least the most important parts of the warning should be as low as feasible. For general target audiences, the reading level might need to be in the fourth- to sixth-grade levels (education of children 10–12 years old). Clearly, some warnings may be directed at professionals such as licensed health care professionals who have some expected level of training and can therefore be more technical. The reading levels should be matched with the intended target audience. There are readability formulas based on word frequency of use, length of words, number of words in statements, and so on, that are used to estimate reading grade level (Duffy, 1985). These formulas have limitations and are notorious for giving inaccurate estimates on comprehensibility. However, they could be useful in analyzing the text while trying to achieve a comprehensible warning. A discussion of reading level measures and their application to the design of instructions and warnings can be found in Duffy (1985).

An additional point on reading ability concerns illiteracy. Even in the richest countries of the world there are a substantial number of functional illiterates. There are estimates that over 16 million functionally illiterate adults exist in the U.S. population. Therefore, successfully communicating warnings may require more than simply keeping reading levels to a minimum. While simple solutions to this problem do not exist, well-designed pictorials, speech warnings, special training programs, and so on, may be important components of warning systems to accommodate these groups.

*Message Content*   The content of the warning message should include information about the hazard, the consequences of the hazard, and instructions on how to avoid the hazard.

*Hazard Information*   The point of giving hazard information is to tell the target audience about potential safety problems. Example hazard statements are:

Toxic vapors
Slippery floor
High voltage (7200 volts)

A general principle is that the hazard should be spelled out clearly in a warning. The exceptions pertain to when the hazard is (a) generally known by the population, (b) known from previous experience, or (c) "open and obvious." (The latter two concepts will be described in more detail in a subsequent section). Other than these exceptions, hazard information is an important component of most warnings (Wogalter et al., 1987).

**Consequences**   Consequences information concerns the nature of the injury, illness, or property damage that could result from the hazard. Hazard and consequence information is usually closely linked in the sense that one leads to the other; or, stating it in the reverse, one is the outcome of the other. Statements regarding these two elements are sometimes purposely sequenced in this way such as in "Toxic Vapor, Severe Lung Damage."

Sometimes, however, it is desirable to put consequences information near the beginning of the warning for the purposes of getting and holding the receiver's attention (Young et al., 1995). This is particularly true for severe consequences such as death, paralysis, and severe lung damage. So the appropriate sequence of statements is the opposite of that mentioned above, as in "Severe Lung Damage, Toxic Vapor."

There are also situations when the hazard information in a warning is presented and understood, where it may not be necessary to state the consequences in the warning. This point is related to the open and obvious aspects of hazards. For example, a sign indicating "Wet Floor" probably does not need to include a consequence statement "You Could Fall." It is reasonable to assume that people will correctly infer the appropriate consequence. Nevertheless, the hazard statement could be improved with including "Slippery" instead of "Wet" so as to include consequences in with the statement. Although this is a simple example, it shows how consequence information can be included together with a hazard statement relatively easily without appearing superfluous.

An important reason why consequences information is needed is that warning recipients may not make the correct inference regarding injury, illness, or property damage outcomes with more complex hazards than a wet floor. Previous research with older adults indicates that people aged 65+ years often have difficulty comprehending warning content when inferences are required (Hancock et al., 2005). Thus, it is important in designing warnings to assess, if necessary, whether people correctly infer the consequences and, if not, then to reword or redesign the warning so it is more specific and informative.

The lack of specificity is a shortcoming in many warnings. They often fail to provide important details. The statement "May be hazardous to your health" in the context of a toxic vapor hazard does not tell the receiver whether he or she may develop a minor cough or suffer severe lung damage (or some other outcome). Also giving only general information frequently fails one of the main purposes of warnings—to give "informed consent" about risks. As will be discussed later, knowledge about severe consequences can motivate attention to and compliance with the warning message (see section on motivation).

Pictorials can also be used to communicate consequence information. Some pictorials (e.g., for a slippery floor hazard) convey both hazard and consequence information without it being stated separately. Figure 4 contains some example industrial safety symbols that convey hazard and consequence information. Pictorial warnings that illustrate both hazard and consequence information are preferred (Goldsworthy et al., 2008a; Mayhorn and Goldsworthy, 2007, 2009).

**Instructions**   In addition to getting people's attention and telling them about the hazard and potential consequences, warnings should also instruct people about what to do or not do in order to stay safe and/or prevent property damage. Typically, but not always, instructions in a warning follow the hazard and consequence information. An example of an instructional statement is "Must Use Respirator Type 1234," which could be included in the context of hazard and consequence statements, as in "Severe Lung Damage, Toxic Vapors, Must Use Respirator Type 1234." The instruction assumes, of course, that the receiver will know what a type 1234 respirator is and have access to one.

Pictorials can be used to communicate instructions. Figure 5 shows examples of instructional information used in warnings. Note that some pictorials use a prohibition symbol, a circle containing the pictorial with a slash through it. Both the circle and slash are usually red, although sometimes they are black.

Sometimes a distinction is made between warnings and instructions. Warnings are communications about



(a)                    (b)

(c)                    (d)
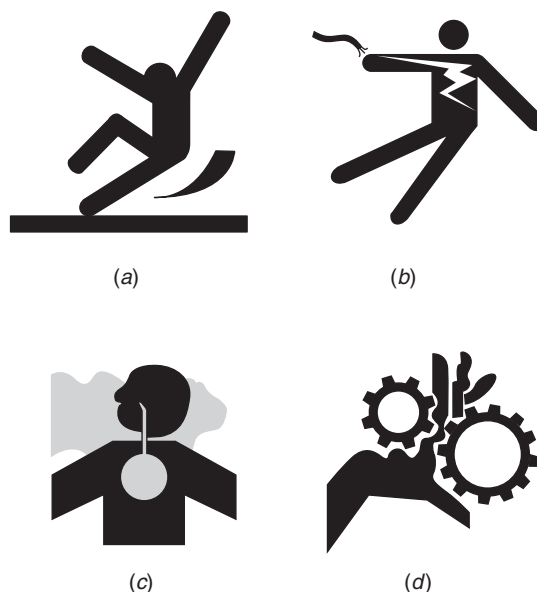
**Figure 4**   Examples of pictorials conveying hazard information: (a) slippery floor; (b) electrical shock; (c) toxic gas; (d) pinch point.
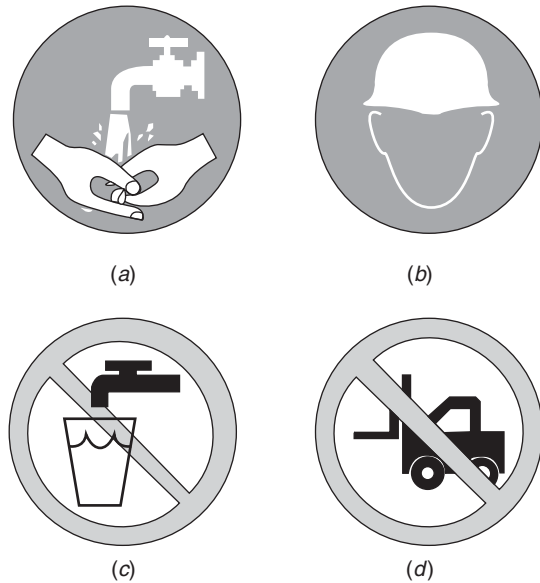
**Figure 5** Examples of pictorials conveying instructions/directions information: (*a*) wash hands; (*b*) wear hard hat; (*c*) do not drink water; (*d*) no forklifts in area.

safety, while instructions may or may not concern safety. "Keep off the grass" is an instruction that generally has nothing to do with safety (unless the grass is infested with fire ants, in which case the statement alone clearly would not be an adequate warning). When instructions are concerned with safety information or safe behavior, then they can be viewed as part of a warning. In short, warnings include instructions, but not all instructions are parts of a warning.

***Explicitness*** Previously, it was mentioned that specificity is generally preferred over generalities. An important design principle relevant to warning comprehension is explicitness (Laughery et al., 1993a; Laughery and Paige-Smith, 2006). Explicit messages contain information that is sufficiently clear and detailed to permit the receiver to understand at an appropriate level the nature of the hazard, the consequences, and the instructions. The key here is the word "appropriate." A classic example is "Use with adequate ventilation." Does this statement mean open a window, use a fan, or something much more technical in terms of volume of air flow per unit time? Obviously the instruction is not clear. Warnings are frequently not detailed or specific enough. However, sometimes, as stated earlier, technical details are not necessary and could be detrimental in certain instances. The following two examples of warnings, each with hazard, consequence, and instructional statements, are inadequate with regard to explicitness: (a) "Dangerous Environment, Health Hazard, Use Precautions" and (b) "Mechanical Hazard, Injury Possible, Exercise Care." Explicit alternatives might be (a) "Severe Lung Damage, Toxic Chlorine

Vapor, Must Use Respirator-Type 123" and (b) "Pinch Point Hazard—Moving Rollers, Your Hand/Arm May Be Severely Crushed or Amputated, Do Not Operate without Guard X89 in Place."

***Pictorial Symbols*** Pictorial symbols are used to communicate hazard-related information, often in conjunction with a printed text message. Guidelines such as ANSI (2006) Z535.3 and FMC Corporation (1985) place considerable emphasis on the use of safety symbols. Pictorials are particularly useful in helping to increase comprehension (Boersema and Zwaga, 1989; Collins, 1983; Dewar, 1999; Lerner and Collins, 1980; Laux et al., 1989; Wolff and Wogalter, 1993, 1998; Zwaga and Easterby, 1984). Well-designed symbols can be useful to low literates or to persons who do not use the regional language (Mayhorn and Goldsworthy, 2007, 2009). Well-designed pictorials can potentially cue large amounts of knowledge at a glance.

Clearly comprehension is a primary concern for pictorials. In some pictorials, the depiction directly represents the information or object being communicated and will be understood if the person recognizes the intended depiction. Figure 6 shows two examples of direct representation. One shows both a hazard and consequences by depicting a raging fire, and the other shows both the hazard and the instructions, depicting the need for an eye shield. In other pictorials, the symbol may be recognized, but its meaning has to be learned. People may recognize a skull and crossbones, but the fact that it represents a poison hazard would have to be learned. Nowhere is this more apparent than the instance cited by Casey (1998) where hundreds of Kurdish farmers in Northern Iraq died when they consumed grain treated
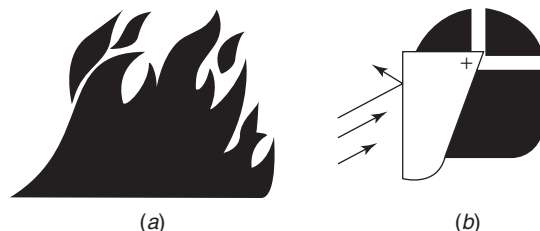


**Figure 6** Examples of pictorials showing a direct representation: (*a*) raging fire and (*b*) wear eye shield.
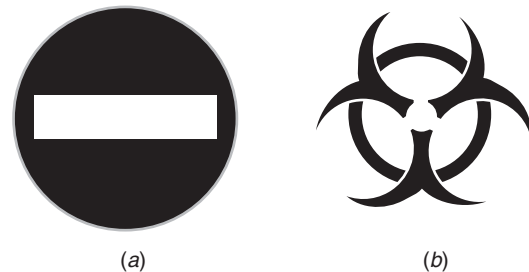


**Figure 7** Examples of pictorials that can be recognized only after learning: (*a*) do not enter and (*b*) biohazard.

with alkyl mercury fungicide because they did not recognize the skull and crossbones symbol as meaning "poison." Reports following the incident suggest that the Kurd farmers believed the skull and crossbones symbol to be a piece of artwork associated with a corporate logo. This example clearly illustrates that cultural differences can also affect warning comprehension (Smith-Jackson and Wogalter, 2000). Other pictorials are completely abstract, such as the symbols for the "do not enter" (shown in Figure 7) and biohazard concepts. Symbols such as these also must be learned to be understood. As a general principle, pictorials that directly represent the information, such as a the "wash hands" symbol showing two hands under a faucet, are recognized at a higher rate than pictorials representing abstract concepts.

What is an acceptable level of comprehension for pictorials? This question has been addressed in the ANSI (2006) Z535.3 standard, which suggests a goal of 85% comprehension by the target audience. There are two criteria that seem relevant here. The first is simply that pictorial symbols should be designed to accomplish the highest level of comprehension attainable. If 85% cannot be achieved, the symbol may still be useful if it is better than alternative designs. A second criterion is that the pictorial not be misinterpreted or communicate incorrect information. According to the ANSI (2006) Z535.3 standard, an acceptable symbol must have less than 5% critical confusions (opposite meaning or a meaning that would produce unsafe behavior). Research by Mayhorn and Goldsworthy (2007) illustrates an example of a misinterpretation of a pictorial that was part of a warning for the drug Accutane. This drug is used for severe acne but causes birth defects in babies of women taking the drug during pregnancy. The pictorial shows a side-view outline shape of a pregnant woman within a circle-slash prohibition symbol. The intended meaning of the pictorial is that women should not take the drug if they are pregnant or plan to become pregnant. However, some women incorrectly interpreted the symbol to mean that the drug might help in preventing pregnancy.

**Habituation**    Repeated exposure to a warning over time may result in its being less effective in attracting attention. Even a well-designed warning will eventually become habituated if repeatedly encountered. Sometimes the warning may become habituated with only partial knowledge. While there are no easy solutions to the habituation problem, one approach is to use attention-related features described in this chapter to slow the progress of habituation or to cause dishabituation compared to warnings without the features (Kim and Wogalter, 2009). However, there may be some utility in varying the warnings from time to time. Rotational warnings such as on cigarette packages in the United States were intended to serve this purpose. However, these warnings have not changed in content or appearance in several decades and regular smokers have likely habituated to them. Cigarette warnings in countries like Australia and Canada also have rotating warnings but also have large, highly explicit color pictured ones depicting severe consequences that are more likely to capture attention and reduce warning habituation relative to U.S. cigarette packages. Legislation regarding a U.S. Food and Drug Administration proposal is being considered to update cigarette package warnings to be similar in type to Australia's and Canada's.

**Memory and Experience**    There are several ways to enhance safety knowledge. Employer training, mentioned earlier, is one method. Experience is another way that people acquire safety knowledge. "Learning the hard way" by having experienced an incident (or knowing someone who did) can certainly result in knowledge. Older adults commonly cite personal experience as a source of knowledge regarding hazards associated with household products such as cleaners and appliances (Mayhorn et al., 2004a). However, such experiences are not good experiences to have (!), and they do not necessarily produce accurate perceptions of risk. More on this topic will be given later in the section on beliefs and attitudes.

**Warnings as Reminders**    Although individuals may have knowledge about a hazard, they may not be aware of it at the time they are at risk. In short, there is a distinction between awareness and knowledge. This distinction is analogous to the short-term and long-term memory distinction in cognitive psychology. Short-term, or working, memory is sometimes thought of as conscious awareness, which is known to have limitations. Long-term memory is the vast contents of one's knowledge of the world. The point is that people may have information or experience in their overall knowledge base, but at a given time, it is not in their current awareness—or what they are thinking about. It is not enough to say that people know something. Rather, it is important that people be aware of the relevant information at the critical time. No one knew better than the three-fingered punch press operators of the 1920s that their hand should not be under the piston when it stroked, but such incidents continued to occur. Warnings are insufficient solution in this case. A better solution was a procedural guard requiring the two hands to simultaneously activate separate controls for the press to punch. A similar example comes from hazards associated with farm equipment. Experienced farm workers are quite knowledgeable when asked about the dangers of power take-off (PTO) machinery on tractors, yet a large number of farmers interviewed in a recent study reported knowing someone that had gotten hurt or killed while using this device (McLaughlin and Mayhorn, 2011). Thus, the distinction between knowledge and awareness has implications for the role of warnings as reminders. Potentially warnings could serve to cue information in long-term memory to bring forth related and previously dormant knowledge into conscious awareness (Smith and Wogalter, 2010).

There are several circumstances in which warning reminders are useful and/or needed. Some of the more noteworthy are:

1.   A hazardous situation or product (that is not open and obvious) is encountered infrequently where forgetting may be a factor.

2. Distractions occur during the performance of a task or the use of a product (e.g., environmental stimuli) that will compete for attention.

3. High task loads which exceed attentional capacity, limiting access to related knowledge (high mental workload and task involvement).

When warnings are intended to function only as reminders, it is not always necessary to provide the same information usually required as a full warning. With reminders, getting the person's attention is emphasized. The automobile driver who forgets to fasten the seat belt might be reminded by the buzzer and light warning. (Persons already habituated to cues may need the cues changed.) Another example is the personal digital assistant that can assist users in adhering to medication regimens by sounding an auditory signal when it is time to take a particular medication (Mayhorn et al., 2005). Technology provides the cues to prompt memory.

***"Open and Obvious"***   A source of information about dangers is the situation or product itself. In U.S. law there is a concept of "open and obvious." This concept means that the appearance of a situation or product or the manner in which it functions may communicate the nature of the safety problem. That a knife can cut is apparent to all people except young children. The hazard and consequence of a fall from a height in a construction setting is considered open and obvious unless there are special circumstances. Many hazardous situations are not open and obvious. Some are associated with chemical hazards where labeling and warnings are necessary because the chemical itself might not make the hazard known. Another issue is an attentional one, in which one hazard attracts more attention than another. Hidden hazards have been documented in the agricultural context. Farmers working to repair tractors may actively work to avoid the dangers of moving parts but in doing that succumb to another hazard such as carbon monoxide in an enclosed space (McLaughlin and Mayhorn, 2011).

***Technical Information***   Many warnings require an appreciation of technical information for full and complete understanding of the material. Examples include the chemical content of a toxic material, the maximum safe level of a substance in the atmosphere in parts per million (ppm), and the biological reaction to exposure to a substance. While there are circumstances where it is appropriate to communicate such information (e.g., to the toxicologist on the staff of a chemical plant or the physician prescribing medicine), as a general rule it is neither necessary nor useful to communicate such information to a general target audience. Indeed, it may be counterproductive in the sense that encountering such information may result in the receiver not attending to the remainder of the message. The end user of the toxic material typically does not need to know technical chemical information such as its density in the atmosphere. Rather, he or she needs to be informed that the substance is toxic, what it can do in the way of injury or illness, and how to use it safely. Different

components of the warning system can and often should be used to communicate to the different groups in the target audience.

***Auditory***   Besides simple auditory alerts described earlier in the section on attention, auditory warnings may be used for the specific purpose of conveying particular meanings. These auditory warnings may be nonverbal (distinguishable sounds to cue different things) or verbal (voice).

***Nonverbal Warnings***   Nonverbal auditory warnings can be further divided into simple and complex. Such simple warnings were mentioned in the context of the attention switch stage. Complex nonverbal signals are composed of sounds differing (sometimes dynamically) in amplitude, frequency, and temporal pattern. Their purpose is to communicate different levels or types of hazards. They can transmit more information than simple auditory warnings, but the listener must know what the signal means. Some form of education and training is necessary. Only a limited number of different nonverbal auditory signals should be used to avoid problems in discriminating and cuing their associated meaning (Banks and Boone, 1981; Cooper, 1977).

***Voice Warnings***   Auditory warnings are also transmitted via voice (speech) as in a child being warned from afar by a caretaker. In recent years, voice chips and digitized sound processors have been developed, making voice warnings feasible for a wide range of applications. Under certain circumstances, voice warnings can be more effective in transmitting information than printed signs (Wogalter et al., 1993b; Wogalter and Young, 1991). Additionally voice modifications and manipulations can produce different levels of perceived urgency (Edworthy and Hellier, 2000; Hollander and Wogalter, 2000). Thus there is great promise for voice warnings as they will be increasingly incorporated into daily life. There are, however, some problems inherently associated with voice warnings. Transmitting speech messages requires longer durations than simple auditory warnings or reading an equivalent message. Comprehension can also be a problem with complex voice messages. To be effective, voice messages should be intelligible and brief.

One example of previous research that has successfully demonstrated the utility of voice warnings is Conzola and Wogalter's (1999) "talking box" study. When participants opened the box, a miniaturized voice system delivered a sequence of precautionary steps to be performed before installing a computer disk drive in the box. With safety instructions that require numerous complex steps, working memory could be overloaded if the sequence is provided in one continuous presentation. A system that provides cognitive support by giving carefully timed or user-prompted instructions might be effective in reducing the likelihood of overloading the cognitive system.

### 4.4.3   Beliefs and Attitudes

If a warning successfully captures and maintains attention and is understood, it still might fail to elicit safety

behavior due to discrepant beliefs and attitudes held by the receiver. Beliefs refer to an individual's knowledge of a topic that is accepted as true. Attitudes are similar to beliefs but have greater emotional involvement (DeJoy, 1999). According to the C-HIP model, a warning will be successfully processed at the beliefs-and-attitudes stage if the information concurs with the receiver's current beliefs and attitudes. The warning message is easily processed as this stage if it matches up (and concurrently reinforces) what the receiver already knows. In the process, it will tend to make those beliefs and attitudes stronger and more resistant to change. If, however, the warning information does not concur with the receiver's existing beliefs and attitudes, the beliefs and attitudes must be altered by the warning for it to be effective. The warning must be salient and the message must be strong and persuasive to override preexisting beliefs and motivate compliance.

People's experiences with a situation or product can result in their believing it is safer than it is. It can also be a problem when people believe that their own abilities or competence will enable them to overcome the hazard, such as the drivers who believe their skills with driving will not suffer when they divide their attention by using cellular telephones (Strayer et al., 2003; Wogalter and Mayhorn, 2005b).

*Risk Perception*  One of the important factors in whether people will read and comply with warnings is their perception of the level of hazard and consequences associated with the situation or product. The greater the perceived level of hazard and consequences, the more responsive people will be to warnings (Wogalter et al., 1991, 1993a). Persons who do not perceive products as being hazardous are less likely to notice or read an associated warning (Wogalter et al., 1991; Wogalter et al., 1993a). Perceived hazard is also closely related to the expected injury's severity level. The greater the potential injury, the more hazardous the product is perceived (Wogalter et al., 1991). Even if the warning is read and understood, compliance may be low if the consequence is believed to be low.

*Familiarity*  Familiarity beliefs are formed from past similar experience where at least some relevant information has been acquired and stored in memory. Familiarity may produce a belief that everything that needs to be known about a product or situation is already known (Wogalter et al., 1991, 1993a). A person who is familiar with a piece of equipment might assume that a new, similar piece of equipment operates in the same way as their previous equipment. This may not actually be true, but due to their belief, the person does not read the product manual and as a result could be seriously injured. Numerous studies have explored the effects of people's familiarity/experience with a product on how they respond to warnings associated with the product. Results indicate that the more familiar people are with a product, the less likely they are to look for, notice, or read a warning (Godfrey et al., 1983; Godfrey and Laughery, 1984; LaRue and Cohen, 1987; Otsubo, 1988; Wogalter et al., 1991). Some research has also examined the effects of familiarity on compliance (Goldhaber and deTurck, 1988; Otsubo, 1988). The results have shown that greater familiarity is associated with a lower likelihood to comply with warnings.

This notion of "familiarity breeds contempt," however, should not be overemphasized for at least two reasons. First, people more familiar with a situation or product may have more knowledge about the hazards and consequences as well as an understanding about how to avoid them. Second, with increased use of the product, people are exposed more frequently to the warnings, which can increase the opportunity to be influenced by them. Of course, warnings in tiny dense print may never be read even over many cycles of use. When there is a potential for the negative effects of familiarity to be a factor, stronger warnings may be needed or other efforts required. Clearly, hazardous products that are used repetitively pose special challenges.

Prior experience can be influential in other ways. Having experienced some form of injury or having personal knowledge of someone else being injured has been shown to lead to overestimation of the degree of danger. Similarly, the lack of such experiences may lead to underestimation of danger or not thinking about them at all (Wogalter et al., 1991, 1993a).

A related point concerns the problem of overestimating what people know. Experts in a domain may be so facile with that knowledge that they fail to realize that nonexperts do not have similar skills and knowledge. To the extent it is incorrectly assumed that people have information and knowledge, there may be a tendency to provide inadequate warnings. Fewer cues are necessary for experts to enlist large stores of knowledge relative to the general public. Thus, an important part of the job, environment, and product design is to take into account the target audience's understanding and knowledge of hazards and their consequences [see Laughery (1993) for a discussion of this topic].

### 4.4.4 Motivation

Even if people see, understand, and believe a warning, they may not comply with it. Motivation is very closely tied to behavior because it can serve to energize individuals to carry out activities that they might not otherwise do. Among the most influential factors for motivation with respect to warnings are the cost of compliance and the cost of noncompliance (severity of the potential injury, illness, or property damage). If the warning calls for actions that are inconvenient, time consuming, or costly, there is an increased likelihood that it will not be effective unless the consequences of noncompliance are perceived as highly undesirable.

*Cost of Compliance*  The cost associated with compliance can be a strong motivator. Generally, compliance with a warning requires that people take some action. Usually there are costs associated with taking action. Cost of complying may include time, effort, or even money to carry out the behavior instructed by the warning. When people perceive the costs of compliance to be greater than the benefits, they are less likely to perform the safety behavior. This problem

is commonly encountered in warning analyses, when the instruction statement requires an inconvenient, difficult, or occasionally impossible behavior to carry out. "Always have two or more persons to lift [box or object]" cannot be done if no one else is around. "Wear rubber gloves when handling this product" is inconvenient to do if the user does not have easy access to appropriate gloves and a hardware store is not nearby.

Thus, the requirement to expend extra time or effort can reduce motivation to comply with a warning (Dingus et al., 1991; Wogalter et al., 1987, 1989). A primary way of reducing the cost of compliance is to make the directed behavior easier to perform. For example, if hand protection is required when using a product, gloves might accompany the product. The general rule is that safe use of a product should be as simple, easy, and convenient as possible.

Also, the costs of noncompliance can affect compliance motivation and behavior. This effect is particularly true when the possible consequences of the hazards are severe. Injury associated with noncompliance should be explicitly stated in the warning (Laughery et al., 1993a). Explicit injury–outcome statements such as "Can cause liver disease—a condition that almost always leads to death" provide reasons for complying and are preferred to general, nonexplicit statements such as "Can lead to serious illness." In a sense, compliance decisions can be viewed in part as a trade-off between the perceived costs of compliance and noncompliance.

***Severity of Consequences*** A related issue to costs of noncompliance is severity of consequences. Perceived severity of injury is intimately tied to risk perception, as discussed in the the section on beliefs and attitudes. Severity of injury is a major factor in people's reported willingness to comply with warnings. People's notions of hazardousness are almost entirely based on the seriousness of the potential outcome (Wogalter et al., 1991, 1993a). The likelihood of such events, however, is considered less readily in people's hazard-related judgments (Wogalter and Barlow, 1990; Young et al., 1990, 1992). These findings emphasize the importance of clear, explicit consequence information in warnings. Such information can be critical to people's risk perception and their evaluation of trade-offs between cost of compliance and cost of noncompliance.

***Social Influence and Stress*** Another motivator of warning compliance is social influence. Research (Wogalter et al., 1989) has shown that if people see others comply with a warning they are more likely to comply themselves. Similarly, seeing that others do not comply lessens the likelihood of compliance. Social influence is an external factor with respect to warnings in that it is not part of the warning design. An example of a risky behavior that is strongly influenced by social interaction is the "sharing" of prescription medications by teenagers (Goldsworthy and Mayhorn, 2009; Goldsworthy et al., 2008b). Explicit warnings are needed to counteract misconceptions exacerbated by social factors.

Other factors that influence motivation to comply with a warning are time stress (Wogalter et al., 1998a)

and mental workload (Wogalter and Usher, 1999). In high-stress and high-workload situations, competing activities distribute away some of the cognitive resources available for processing warning information and carrying out compliance behavior.

### 4.4.5 Behavior

The last stage of the sequential process is to carry out the warning-directed safe behavior. Determining what people will do in the context of a warning is a very desirable measure of its effectiveness. Behavioral compliance research shows that warnings can change behavior (e.g., Laughery et al., 1994; Cox et al., 1997; Wogalter et al., 2001). The main issue in contemporary research is to determine the factors and conditions that underlie whether a warning will be effective in producing compliance or not. Silver and Braun (1999) and Kalsher and Williams (2006) have reviewed published research that has measured compliance with warnings under various conditions. Wogalter and Dingus (1999) showed indirect measures may also be useful where a residual outcome of the behavior is examined (e.g., whether a pair of protective gloves have been used according to its stretch marks). Due to the ethical concerns associated with exposing research participants to real hazards, many researchers have measured intentions to comply as a proxy for compliance behavior. Recently, Duarte et al. (2010) described the potential for virtual reality technology to enable the exploration of behavioral compliance without placing users at risk from physical harm, which is one of the main difficulties in doing research that measures actual behavioral compliance.

### 4.4.6 Demographic Factors

The above sections have provided a review of major concepts and findings organized on the basis of the C-HIP model. Newer versions of C-HIP (Wogalter, 2006a) give greater emphasis on demographics differences of receivers. There are also relevant demographic characteristics of receivers. Receivers differ and such differences must be considered in warning design. Laughery and Brelsford (1991) discussed a number of relevant dimensions along which intended receivers may differ. Several such factors have already been discussed, including experience and competence. A number of studies have shown that gender and age may be related to how people respond to warnings. With regard to gender, results suggest a slightly greater tendency for women to be more likely than men to look for and read warnings (Godfrey et al., 1983; LaRue and Cohen, 1987; Young et al, 1989). Similarly, there are research results that show women are more likely to comply with warnings (Goldhaber and deTurck, 1988; Viscussi et al., 1986). However, many other studies either do not report or do not find a gender difference.

Regarding age, the results are mixed. There are results suggesting that people older than 40 are more likely to take precautions in response to warnings (Hancock et al., 2005; Mayhorn et al., 2004a; Mayhorn and Podany, 2006). However, some research (Wogalter and Vigilante, 2003; Wogalter et al., 1999d) has shown that

older adults have more difficulties reading small print on product labels than younger adults. Other research (Collins and Lerner, 1982; Easterby and Hakiel, 1981; Ringseis and Caird, 1995; Schroeder et al., 2001; Shorr et al., 2009) has shown that older subjects had lower levels of comprehension for safety-related symbols than younger adults. Results such as these suggest that older adults may be more influenced by warnings, but legibility and comprehension need to be considered in their design.

Other potentially important demographics include locus of control (Laux and Brelsford, 1989; Donner, 1991) and self-efficacy (Lust et al., 1993). Persons who believe that they can control their destiny and/or who are less confident in a situation or task are more likely to read available warnings than persons who believe that fate controls their lives and/or who are more confident in a situation or task. When designing warnings for the general population, it may not be possible to address all of the needs of different people with a single warning; thus, a multimethod systems approach may be needed to meet the needs of the varying target audience.

### 4.4.7 Summary and Benefit of C-HIP

The above review of factors influencing warning effectiveness was organized around the C-HIP model. This model divides the processing of warning information into separate stages that must be completed successfully for compliance behavior to occur. A bottleneck at any given stage can inhibit processing at subsequent stages. Table 1 summarizes some of factors that influence the processing at each stage.

The basic C-HIP model can be a valuable tool in developing and evaluating warnings. Identifying potential processing bottlenecks can be useful in determining why a warning may or may not be successful. The model, in conjunction with empirical data obtained in various types of testing, can identify specific deficiencies in the warning system. Suppose a manufacturer finds that a critical warning on their product label is not working to prevent injury. The first reaction to solving the compliance problem might be to increase the size of the font so more people are likely to see it. But noticing the warning label (the attention switch stage) might not be the problem. Product testing might instead reveal that virtually all users report having seen the warning (attention switch stage), having read the warning (attention maintenance stage), having understood the warning (comprehension and memory stage), and having believed the message (the beliefs and attitudes stage). Thus, the problem with the manufacturer's warning in this case is likely to be at the motivation stage—users may not be complying because they believe the cost of complying with the warning (e.g., wearing uncomfortable personal protection equipment) did not outweigh the small perceived risk about getting injured. The point here is that one could use the model to pinpoint the causes of the warning not working and try to remedy it by targeted means. By using the model as an investigative tool, one can determine the specific causes of a warning's failure and not waste resources trying to fix a wrong aspect of the warning's design.

For the practitioner, the model has utility in determining the adequacy and potential effectiveness of a warning. To the extent that a warning fails to meet various design criteria, the model can be a basis for judging adequacy. The lack of signal words, color, and pictorials or a poor location can be a basis for judging its adequacy regarding attention. A high reading level, the use of technical terminology, or the omission of critical information may be the basis of a warning's comprehension inadequacy. The failure to give a persuasive statements and a conspicuous presentation could result in low effectiveness. The failure to provide explicit consequences information when the outcome of noncompliance is catastrophic is inconsistent with warning adequacy criteria regarding motivation. Considerations such as these can be useful in formulating opinions and addressing issues on why a warning was not successful.

## 5 DESIGNING FOR APPLICATION

It is important to design warning systems that will maximize their effectiveness. This section considers basic guidelines and principles to assist in the design and production of warnings.

### 5.1 Standards

A starting point in designing warnings is to consider existing guidelines such as the ANSI (2006) Z535, FMC Corporation (1985), or Westinghouse Electric Corporation (1981). ANSI Z535 is currently a six-part standard which includes descriptions of safety colors, signs, symbols, labels, tags, and ancillary materials. ANSI standards are voluntary standards; that is, they are only recommendations and are generally considered "minimums." We believe that blindly following the ANSI standard will not lead to great warnings. There is a need for some human factors judgment and testing to fine tune the warning for the particular product or situation. In the ANSI Z535 standard, there is an emphasis on a standardized way to format signs (Z535.2) and product labels (Z535.4). According to these standards, warning signs and labels should possess the following components: (1) a signal word panel such as DANGER, WARNING, or CAUTION (with corresponding red, orange, or yellow color) and an alert symbol (triangle enclosing an exclamation mark) to attract attention to the warning and connote levels of hazard, (2) a hazard statement that briefly describes the nature of the hazard, (3) a description of the possible consequences associated with noncompliance, and (4) instructions for how to avoid the hazard. Research indicates that each of these four components can provide benefit to warning efficacy. There may be exceptions when one (or more) of the message components are clear or redundant from the other statements (Wogalter et al., 1987; Young et al., 1995) or from the presence of a pictorial symbol. Pictorial symbols can provide information on the hazard, consequences, or appropriate (or inappropriate) behavior and so can be used in lieu of some of the component text, assuming understandable symbols are used. Safety symbols should meet certain comprehension criteria to be acceptable for use by itself (without words). Both

the ANSI (2006) Z535.3 and the International Organization for Standardization (ISO, 2001) 9186 symbol standards provide guidelines and methods to assess symbol comprehension.

## 5.2  Checklist of Potential Warning Components

Use of only standards and guidelines may not always produce an effective warning. Table 2 presents a checklist of factors that should be considered in designing warnings. These factors are based not only on standards and guidelines but also on empirical research. Examples of measurement methods are also provided in the table. While not an exhaustive list, the table contains a set of factors that the warning literature indicates should be considered in warning design. Thus, one method of assessing warning quality is simply to determine the extent to which the design meets appropriate criteria such as those given in Table 2. With respect to attention, the effectiveness of the warning might be questioned if

**Table 2  Warning Design Guidelines**

| Warning Components | Design Guidelines |
|---|---|
| **Signal words** | DANGER — Indicates immediately hazardous situation that will result in death or serious injury if not avoided; use only in extreme situations. Use white print on a red background (ANSI Z535).<br>WARNING — Indicates a potentially hazardous situation that may result in death or serious injury if not avoided. Use black print on an orange background.<br>CAUTION — Indicates a potentially hazardous situation that may result in minor or moderate injury. Use black print on a yellow background.<br>NOTICE — Indicates important nonhazard information. Use white print on a blue background.<br>Although not in ANSI Z535, the term DEADLY connotes higher-level hazard than DANGER.<br>On the left side of the signal word is the alert symbol (triangle surrounding an exclamation mark). |
| **Format** | Text should be high contrast, e.g., black print on white or yellow background, or vice versa.<br>Left justify message text although headings can be centered.<br>Orient messages to read from left to right.<br>Each statement starts on its own line (list or outline format).<br>Use white space or bullet points to separate statements or sets of statements.<br>Give priority most important warning statements, e.g., position at the top. |
| **Wording** | Use as little text as necessary to clearly convey the message.<br>Give information about the hazard, instructions on how to avoid hazard, and consequences of failing to comply.<br>Be explicit — tell exactly what the hazard is, what the consequences are, and what to do or not do.<br>Use short statements rather than long, complicated sentences.<br>Use concrete rather than abstract wording.<br>Use short, familiar words.<br>Use active voice rather than passive voice.<br>Remove unnecessary connector words, e.g., prepositions, articles.<br>Avoid using words or statements that might have multiple interpretations.<br>Avoid using abbreviations unless they have been tested on the user population.<br>Use multiple languages when necessary. |
| **Pictorial symbols** | When used alone, acceptable symbols should have at least 85% comprehension scores, with no more than 5% critical confusions (opposite or very wrong answers).<br>Comprehension test — use open ended with relevant context.<br>Pictorials not passing a comprehension test should be accompanied by words, but critical confusions should still be avoided.<br>Use bold shapes. Avoid including irrelevant details.<br>Prohibition (circle slash should not obscure critical elements of symbol).<br>Should be legible under degraded conditions, e.g., distance, size, abrasion. |
| **Font** | Text should be legible enough to be seen by the intended audience and expected viewing distance and angle.<br>Use mixed-case letters. Avoid using all caps except for signal words or for specific emphasis.<br>Use san serif fonts (Arial, Helvetica, etc.) for signal words and larger size text.<br>Use serif (Times, Times New Roman, etc.) fonts for smaller size text.<br>Use plain, familiar, nonfancy font.<br>Do not have letters too close to or touch each other. |
| **Other** | Located/positioned so presentation is where it will be seen or heard.<br>Test to assure message fulfills C-HIP stages in Table 1. |

no signal word is used, no color is employed, the print is small, the message is embedded in other types of information, and so on. With respect to comprehension, if the reading level is high, technical language is used, or the statements are vague and not explicit, then the warning may not be interpreted as intended. Similar considerations can be applied with respect to the criteria for the other stages.

Implementation of specific factors may also depend on situational-specific considerations such as target audience knowledge and/or characteristics of the product. For example, some warning components may not be necessary if the target audience consists of trained experts or if the information is apparent from other aspects of the situation.

## 5.3 Principles

In addition to the factors specified in Table 2, there are several other important principles or general guidelines that should be considered when designing warning systems. These principles are described in the following sections.

### 5.3.1 Principle 1: Brief and Complete

As a general rule, warnings should be as brief as possible. Two separate statements should not be included if one will do, such as in the slippery floor example cited earlier. Longer warnings or those with nonessential information are less likely to be read, and they may be more difficult to understand. Thus, the brevity criterion conflicts to some extent with the explicitness criterion. Being explicit about every hazard could result in very long warnings. Obviously, the brevity criterion should not be interpreted as a license to omit important information. A "happy medium" between brevity and completeness is discussed in the next section on prioritization.

A concept related to completeness is overwarning. The term overwarning is sometimes used to label the extent to which our world is filled with warnings. The negative cited from overwarning is that people may not attend to them or may become highly selective, attending only to some warnings. The notion is that if warnings were to be placed on everything, people would simply ignore them. While this notion has face validity, there has been little empirical data assessing the limits implied. Nevertheless, overwarning may be a valid concern, and unnecessary warnings should be avoided.

An important issue related to overwarning that frequently arises in litigation is the absence of certain information. An argument that is sometimes made is that information being left off was somehow a benefit to consumers because its inclusion would hurt the likelihood of other important information being read. However, this is often just a post hoc defense and it does not comport with "right to know." The notion of informed consent says that warnings should provide to people the opportunity to know about hazards. Indeed, research indicates that people want to know about hazards even if is difficult to give definitive risk information (Freeman and Wogalter, 2002; Cheatham and Wogalter, 2003). Prioritization, discussed in the next

section, is a useful approach in dealing with warnings for products and equipment that have multiple hazards.

### 5.3.2 Principle 2: Prioritization

Prioritization concerns what hazards to warn about and to emphasize when multiple hazards exist. How are priorities defined in deciding what to include/delete, how to sequence items, or how much relative emphasis to give them? The criteria overlap the rules about what and when to warn. According to Vigilante and Wogalter (1997a, 1997b), considerations include:

1. *Likelihood*. The more frequently an undesirable event occurs, the greater the priority it should be given as a warning.

2. *Severity*. The more severe the potential consequences of a hazard, the greater priority it should be warned. If a chemical product poses a skin contact hazard, a higher priority would be given to a severe chemical burn consequence than if it were a minor rash.

3. *Known (or Not Known) to Target Population*. If the hazard is already known and understood or if it is open and obvious, warnings may not be needed (except as a possible reminder).

4. *Importance*. Is it important for individuals to know? In most cases, people want the opportunity to know about risks. Some hazards may be more important to people than others.

5. *Practicality*. There are occasions when limited space (a small label) or limited time (a television commercial) does not permit all hazards to be addressed in a single component of the warning system.

As a general rule, unknown and important hazards leading to more severe consequences and/or those more likely to occur should have higher priority than less severe or less likely hazards. Higher priority warnings should be placed on the product label. If not practical to place them all on the label, those with lower priority might go on other warning system components such as package inserts, manuals, websites, or other media.

### 5.3.3 Principle 3: Know the Receiver

Gather information and data about relevant receiver characteristics. To illustrate such an effort, Goldsworthy et al. (2010) describe an analytic technique known as latent class analysis (LCA) to facilitate the tailoring of warning content designed to prevent the sharing and exchanging of prescription medications. Receiver-centered testing of the target audience was particularly important because of the complex risk-related scenarios involved.

A related way to meet the needs of receivers is to purposely tailor for the warning as appropriate to the person, product, and/or situation. One approach to tailoring warnings can be accomplished through the use of technology, such as using sensors, computers, software, and displays (Wogalter and Mayhorn, 2005a). To

provide such customization, data must be collected and quickly processed to anticipate and present the needed warning information at the appropriate time. Users could carry relevant data with him or her. Currently, there are "smart" credit cards that contain user information and wireless electronic tags that can transmit information within a short proximity (e.g., ExxonMobil's Smart Pass, which identifies credit customers by passing an electronic key near the face of the gas pump). Advanced warning systems would be able to supply information tailored to meet people's particular needs.

### 5.3.4 Principle 4: Design for Low-End Receiver

When there is variability in the target population, which is almost always the case with the general public, design for the low-end extreme. Safety communications should not be written at the level of the average or median percentile person in the target audience. Such warnings will present comprehension problems for people at lower competence, experience, and knowledge levels. Likewise, formatting and presentation should take into consideration those who are older, perceptually disabled, and otherwise unable to access the warning information. An added benefit of designing warning systems for the low-end user is the realization that these solutions typically result in more user-friendly products and environments that benefit all consumers regardless of ability and demographic differences (Vanderheiden, 1997).

### 5.3.5 Principle 5: Warning System

When the target audience consists of subgroups that differ on relevant dimensions or when they may be involved under different conditions, consider employing a warning system that includes different components for the different subgroups. Do not assume that everything will be accomplished with a single warning or warning method.

### 5.3.6 Principle 6: Durability

Warnings should be designed to last as long as needed. There are circumstances where durability is typically not a problem. A product purchased off the shelf of a drug store that will be completely and immediately consumed is an example. On the other hand, products with a long lifespan, such as cars and lawn mowers, may present a challenge (Glasscock and Dorris, 2006). Similarly, in situations where warnings are exposed to weather such as on construction sites or extensive handling such as on some containers, durability problems can influence comprehension (Shorr et al., 2009). Some products have manuals that list warning labels with part numbers, presumably to enable ordering label replacements when needed. Undoubtedly replacement labels are not frequently ordered, a factor that suggests the original labels should be as durable as possible so as to last to the high-end range of the expected life of the product.

Related to durability is ancillary material that accompanies the product when originally purchased as new. Warnings may be printed on a an outer container box or packaging and on an insert or in an owner's manual. These ancillary materials may not be available at later uses of the product. The box or packaging may be discarded (Cheatham and Wogalter, 2003) or the owner's manual may be discarded or misplaced (Wogalter et al., 1998b) or never transferred to subsequent owners or users of a product (Mehlenbacher et al., 2002). This is why consideration of what warnings to place directly on a product (or on a container) is critical because they may be the only ones available to users at later points in time.

### 5.3.7 Principle 7: Test the Warning

In addition to considering design criteria, it is frequently necessary to carry out some sort of testing to evaluate a particular warning or several prototype warnings. This approach may entail using small groups of people to give ideas for improvement and/or formal assessments involving larger numbers of people giving independent evaluations. Of course, the sample should be representative of the target audience while also considering practicality and feasibility.

To assess attention, a warning could be placed on a product and have people carry out a relevant task using the product to determine if they look at or notice it. Regarding comprehension, conducting studies to assess the extent to which a warning is understood probably has one of the best cost–benefit ratios of any procedure in the warnings design process. Relative to behavioral studies, comprehension can be assessed easily and quickly and at low cost. Well-established methodologies involving memory tests, open-ended response tests, interviews, and so on, are applicable. While the qualitative data that result from open-ended and interview methodologies can be problematic, such studies can be exceptionally valuable in determining what information in the warning was or was not understood as well as what might be done in the way of redesign to increase the level of comprehension.

Studies can also be carried out to determine the extent to which members of the target audience accept the warning information as true and to be applicable to them (beliefs and attitudes). Negative results on these dimensions would indicate the warning lacks sufficient persuasiveness. Motivation can be assessed by obtaining measures of compliance intentions. While such intention measures will generally reflect higher levels than actual compliance, they can be useful for determining whether or not the warning is likely to be effective as well as for comparing warnings to determine which would likely be more effective.

While behavioral compliance studies are generally difficult to execute, in situations where negative consequences of an ineffective warning are high, the effort may be warranted. As mentioned earlier, a possible alternative is to utilize virtual reality methodology to avoid such ethical issues (Duarte et al., 2010). If such technology is not available, behavioral intentions can be measured as a proxy for behavioral data. Poor warnings tend to result from no testing whatsoever.

Studies carried out to evaluate the potential effectiveness of a warning must, of course, incorporate appropriate principles of research design. The selection of

subjects to be representative of the target population, avoiding confounding by extraneous variables, guarding against contamination by expected outcomes, and determining the best coding rubric to assess qualitative comprehension data from open-ended assessments are a few of the more salient factors that must be considered. For a more complete discussion of approaches to evaluating warning effectiveness, see Frantz et al. (1999), Kalsher and Williams (2006), Mayhorn and Goldsworthy (2009), Wogalter and Dingus (1999), Young and Lovvoll (1999), and Wogalter et al. (1999a).

## 6 SUMMARY AND CONCLUSIONS

Warning design and effectiveness are comprised of many factors and considerations. In this chapter we have presented an overview of the current status of research, guidelines, and criteria for designing warnings.

Approaches to dealing with environmental or product hazards are generally prioritized such that the first one tries to solve the problem by design, then by guarding, then by warning. Thus, in the domain of safety, warnings are viewed as a third but important line of defense.

Warnings can be properly viewed as communications whose purposes include informing and influencing the behavior of people. Warnings are not simply signs or labels. They can include a variety of media through which various kinds of information get communicated to a broad spectrum of people. The use of various media or channels and an understanding of the characteristics of the receivers or target audiences to whom warnings are directed are important in the design of effective warnings. The concept of a warning system with multiple components or channels for communication to a variety of receivers is central in this regard.

The design of warnings can and should be viewed as an integral part of systems design. Too often it is carried out after the environment or product design is essentially completed, a kind of afterthought phenomenon. Importantly, warnings cannot and should not be expected to serve as a cure for bad design.

In this chapter, the C-HIP model was described. It involves processing stages based on communication theory and human information processing theory. As part of this discussion, relevant factors influential at each stage were presented. In addition, guidelines and principles for warning design in application were presented. Its potential use as an investigative tool was also discussed.

Determining whether or not a warning will influence behavior is often a difficult assignment. In addition to ethical problems of exposing people to hazards, actual field studies testing warnings are likely to be time consuming and costly. Certainly, where feasible, such studies are desirable. Also, while laboratory or other controlled simulations of warning situations can be useful in assessing behavioral effects, such approaches leave open questions of generalizability. Studies that examine the effects of warnings on attention, comprehension, beliefs and attitudes, and motivation to comply can be valuable as part of the

process of designing and assessing warnings. Such studies can help in isolating why a warning is not effective. A behavioral study that shows people do not comply with a warning may not tell us if it failed because it was not noticed, because it was not understood, because it was not believed, or because it was unable to motivate. Studies employing attention, comprehension, risk perception, or behavioral intention measures can provide information that, in turn, can be useful in developing improved warning designs.

The issue of warning effectiveness has received a great deal of attention in recent years, especially the means by which effectiveness is assessed. Several criteria can be employed in assessing warnings, including whether they capture and maintain attention, are understood, are consistent with or capable of modifying beliefs and attitudes, motivate people to comply, and result in people behaving safely. The assessment of warning effectiveness employing approaches provides useful input toward the goal of providing effective warnings.

## REFERENCES

American National Standards Institute (ANSI) (2006), *Accredited Standards Committee on Safety Signs and Colors. Z535.1-6*, National Electrical Manufacturers Association, Arlington, VA.

Banks, W. W., and Boone, M. P. (1981), *Nuclear Control Room Enunciators: Problems and Recommendations*, NUREG/CR-2147, National Technical Information Service, Springfield, VA.

Barlow, T., and Wogalter, M. S. (1991), "Increasing the Surface Area on Small Product Containers to Facilitate Communication of Label Information and Warnings," in *Proceedings of Interface 91*, Human Factors Society, Santa Monica. CA, pp. 88–93.

Barlow, T., and Wogalter, M. S. (1993), "Alcoholic Beverage Warnings in Magazine and Television Advertisements," *Journal of Consumer Research*, Vol. 20, pp. 147–155.

Barzegar, R. S., and Wogalter, M. S. (1998), "Intended Carefulness for Voiced Warning Signal Words," *Proceedings of the Human Factors and Ergonomics Society*, Vol. 42, pp. 1068–1072.

Boersema, T., and Zwaga, H. J. G. (1989), "Selecting Comprehensible Warning Symbols for Swimming Pool Slides," *Proceedings of the Human Factors Society*, Vol. 33, pp. 994–998.

Bzostek, J. A., and Wogalter, M. S. (1999), "Measuring Visual Search Time for a Product Warning Label as a Function of Icon, Color, Column, and Vertical Placement," *Proceedings of the Human Factors and Ergonomics Society*, Vol. 43, pp. 888–892.

Casey, S. (1998), *Set Phasers* On *Stun: And Other True Tales* Of *Design, Technology,* And *Human Error*, 2nd ed., Aegean Publishing, Santa Barbara, CA.

Chapanis, A. (1994), "Hazards Associated with Three Signal Words and Four Colours on Warning Signs," *Ergonomics*, Vol. 37, pp. 265–275.

Cheatham, D. B., and Wogalter, M. S. (2003), "Comprehension of Over-the-Counter Drug Label Warnings Regarding Consumption of Acetaminophen and Alcohol,"

*Proceedings of the Human Factors and Ergonomics Society*, Vol. 47, pp. 1540–1544.

Collins, B. L. (1983), "Evaluation of Mine-Safety Symbols," *Proceedings of the Human Factors Society*, Vol. 27, pp. 947–949.

Collins, B. L., and Lerner, N. D. (1982), "Assessment of Fire-Safety Symbols," *Human Factors*, Vol. 24, pp. 75–84.

Conzola, V. C., and Wogalter, M. S. (1999), "Using Voice and Print Directives and Warnings to Supplement Product Manual Instructions," *International Journal of Industrial Ergonomics*, Vol. 23, pp. 549–556.

Cooper, G. E. (1977), *A Survey of the Status and Philosophies Relating to Cockpit Warning Systems*, NASA-CR-152071, NASA Ames Research Center, Moffett Field, CA.

Cox, E. P. III, Wogalter, M. S., Stokes, S. L., and Murff, E. J. T. (1997), "Do Product Warnings Increase Safe Behavior?: A Meta-Analysis," *Journal of Public Policy and Marketing*, Vol. 16, pp. 195–204.

Desaulniers, D. R. (1987), "Layout, Organization, and the Effectiveness of Consumer Product Warnings," *Proceedings of the Human Factors Society*, Vol. 31, pp. 56–60.

DeJoy, D. M. (1999), "Beliefs and Attitudes," in *Warnings and Risk Communication*, M. S. Wogalter, D. M. DeJoy, and K. R. Laughery, Eds., Taylor and Francis, London, pp. 183–219.

Dewar, R., (1999), "Design and Evaluation of Graphic Symbols," in *Visual Information for Everyday Use: Design and Research Perspectives*, H. J. G. Zwaga, T. Boersema, and H. C. M. Hoonhout, Eds., Taylor and Francis, London, pp. 285–303.

Dingus, T. A., Hathaway, J. A., and Hunn, B. P. (1991), "A Most Critical Warning Variable: Two Demonstrations of the Powerful Effects of Cost on Warning Compliance," *Proceedings of the Human Factors Society*, Vol. 35, pp. 1034–1038.

Donner, K. A. (1991), "Prediction of Safety Behaviors from Locus of Control Statements," in *Interface '91', Proceedings of the 7th Symposium on Human Factors and Industrial Design in Consumer Products*, Human Factors Society, Santa Monica, CA, pp. 94–98.

Duarte, M. E. C., Rebelo, F., and Wogalter, M. S. (2010), "The Potential of Virtual Reality (VR) for Evaluating Warning Compliance," *Human Factors and Ergonomics in Manufacturing and Service Industries*, Vol. 20 pp. 526–537.

Duffy, T. M. (1985), "Readability Formulas: What's The Use?," in *Designing Usable Texts*, T.M. Duffy and R. Waller, Eds., Academic Press, Orlando, FL.

Easterby, R. S., and Hakiel, S. R. (1981), "The Comprehension of Pictorially Presented Messages," *Applied Ergonomics*, Vol. 12, pp. 143–152.

Edworthy, J., and Hellier, E. J. (2000), "Auditory Warnings in Noisy Environments," *Noise and Health*, Vol. 6, pp. 27–39.

Edworthy, J. Loxley, S., and Dennis, I. (1991), "Improving Auditory Warning Design: Relationship between Warning Sound Parameters and Perceived Urgency," *Human Factors*, Vol. 33, pp. 205–231

FMC Corporation (1985), *Product Safety Sign and Label System*, FMC Corporation, Santa Clara, CA.

Frantz, J. P., and Rhoades, T. P. (1993), "A Task Analytic Approach to the Temporal Placement of Product Warnings," *Human Factors*, Vol. 35, pp. 719–730.

Frantz, J. P., Rhoades, T. P., and Lehto, M. R. (1999), "Practical Considerations Regarding the Design and Evaluation of Product Warnings," in *Warnings and Risk Communication*, M. S. Wogalter, D. M. DeJoy, and K. R. Laughery, Eds., Taylor and Francis, London, pp. 291–311.

Freeman, K., and Wogalter, M. S. (2002), "On Informing Women of Child Bearing Age about Seat Belt Risk during Pregnancy," *Proceedings of the Human Factors and Ergonomics Society*, Vol. 46, pp. 943–946.

Glasscock, N. F. and Dorris, N. T. (2006), "Warning Degradation and Durability," in *Handbook of Warnings*, M. S. Wogalter, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 499–511.

Godfrey, S. S., and Laughery, K. R. (1984), "The Biasing Effect of Familiarity on Consumer's Awareness of Hazard," *Proceedings of the Human Factors Society*, Vol. 28, pp. 483–486.

Godfrey, S. S., Allender, L., Laughery, K. R., and Smith, V. L. (1983), "Warning Messages: Will the Consumer Bother To Look?" *Proceedings of the Human Factors* Society, Vol. 27, pp. 950–954.

Goldhaber, G. M., and deTurck, M. A. (1988), "Effects of Consumer's Familiarity with a Product on Attention and Compliance with Warnings," *Journal of Products Liability*, Vol. 11, pp. 29–37.

Goldsworthy, R. C. and Mayhorn, C. B. (2009), "Prescription Medication Sharing among Adolescents: Prevalence, Risks, and Outcomes," *Journal of Adolescent Health*, Vol. 45, No. 6, pp. 634–637.

Goldsworthy, R. C., and Mayhorn, C. B. (2010), "Direct to Consumer (DTC) Prescription Drug Advertising: Exploring Self Reports of Media Exposure and Associated Behaviors," in *Proceedings of the Third Applied Human Factors and Ergonomics International Conference*, Miami, FL.

Goldsworthy, R. C., Schwartz, N., and Mayhorn, C. B. (2008a), "Interpretation of Pharmaceutical Warnings among Adolescents," *Journal of Adolescent Health*, Vol. 42, No. 6, pp. 617–625.

Goldsworthy, R. C., Schwartz, N., and Mayhorn, C. B. (2008b), "Beyond Abuse and Exposure: Framing the Impact of Prescription Medication Sharing," *American Journal of Public Health*, Vol. 98, No. 6, pp. 1115–1121.

Goldsworthy, R. C., Mayhorn, C. B., and Meade, A. W. (2010), "Warnings in Manufacturing: Improving Hazard Mitigation Messaging through Audience Analysis," *Human Factors and Ergonomics in Manufacturing and Service Industries*, Vol. 20, No. 6, pp. 484–499.

Hancock, H. E., Fisk, A. D., and Rogers, W. A. (2005), "Comprehending Product Warning Information: Age-Related Effects and the Roles of Memory, Inferencing, and Knowledge," *Human Factors*, Vol. 47, No. 2, pp. 219–234.

Hartley, J. (1994), *Designing Instructional Text*, 3rd ed., Kogan Page, London/ Nichols, East Brunswick, NJ.

Hellier, E., Edworthy, J., Weedon, B, Walters, K, and Adams, A. (2002), "The Perceived Urgency of Speech Warnings: Semantics versus Acoustics," *Human Factors*, Vol. 44, pp. 1–17.

Hollander, T. D., and Wogalter, M. S. (2000), "Connoted Hazard of Voice Warning Signal Words: An Examination of Auditory Components," *Proceedings of the International Ergonomics Association and the Human Factors and Ergonomics Society Congress*, Vol. 44, No. 3, pp. 702–705.

International Organization for Standardization (ISO) (2001), *Graphical Symbols–Test Methods For Judged Comprehensibility And For Comprehension*, ISO 9186, ISO, Geneva.

Jaynes, L. S., and Boles, D. B. (1990), "The Effects of Symbols on Warning Compliance," *Proceedings of the Human Factors* Society, Vo. 34, pp. 984–987.

Kalsher, M. J., and Williams, K. J. (2006), "Behavioral Compliance: Theory, Methodology, and Results," in *Handbook of Warnings*, M. S. Wogalter, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, Chapter 23, pp. 313–331.

Kalsher, M. J., Wogalter, M. S., and Racicot, B. M. (1996), "Pharmaceutical Container Labels and Warnings: Preference and Perceived Readability of Alternative Designs and Pictorials," *International Journal of Industrial Ergonomics*, Vol. 18, pp. 83–90.

Kim, S., and Wogalter, M. S. (2009). "Habituation, Dishabituation, and Recovery Effects in Visual Warnings," *Proceedings of the Human Factors and Ergonomics Society*, Vol. 53, pp. 1612–1616.

Kim, S., Mayhorn, C. B., and Wogalter, M. S. (2010), "Use of Information Sources Referenced in and Attitudes about Televised DTC Prescription Drug Advertisements," in *Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1880–1884.

LaRue, C., and Cohen, H. (1987), "Factors Influencing Consumer's Perceptions of Warning: An Examination of the Differences between Male and Female Consumers," *Proceedings of the Human Factors Society*, Vol. 31, pp. 610–614.

Lasswell, H. D. (1948), "The Structure and Function of Communication in Society," in *The Communication of Ideas*, L. Bryson, Ed., Wiley, New York.

Laughery, K. R. (1993), "Everybody Knows: Or Do They?" *Ergonomics in Design*, July, pp. 8–13.

Laughery, K. R., and Brelsford, J. W. (1991), "Receiver Characteristics in Safety Communications," *Proceedings of the Human Factors Society*, Vol. 35, pp. 1068–1072.

Laughery, K. R., and Paige-Smith, D. (2006), "Explicit Information in Warnings," in *Handbook of Warnings*, M. S. Wogalter, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, Chapter 31, pp. 419–428.

Laughery, K. R., and Wogalter, M. S. (2011), "The Hazard-Control Hierarchy and Its Utility in Safety Decisions about Consumer Products," in *Handbook of Human Factors and Ergonomics in Consumer Product Design: Uses and Applications*, Vol. 2, W. Karwowski, M. Soares, and N. Stanton (Eds), CRC Press, Boca Raton, FL, pp. 33–40.

Laughery, K. R., Vaubel, K. P., Young, S. L., Brelsford, J. W., and Rowe, A. L. (1993a), "Explicitness of Consequence Information In Warning," *Safety Science*, Vol. 16, pp. 597–613.

Laughery, K. R., Young, S. L., Vaubel, K. P., and Brelsford, J. W. (1993b), "The Noticeability of Warnings on Alcoholic Beverage Containers," *Journal of Public Policy and Marketing*, Vol. 12, pp. 38–56.

Laughery, K. R., Wogalter, M. S., and Young, S. L., Eds. (1994), *Human Factors Perspectives on Warnings: Selections from Human Factors and Ergonomics Society Annual Meetings 1980–1993*, Human Factors and Ergonomics Society, Santa Monica, CA.

Laux, L., and Brelsford, J. W. (1989) "Locus of Control, Risk Perception, and Precautionary Behavior," *Proceedings of Interface 89—Sixth Symposium on Human Factors and Industrial Design in Consumer Products*, Human Factors Society, Santa Monica, CA, pp. 121–124.

Laux, L. F., Mayer, D. L., and Thompson, N. B. (1989), "Usefulness of Symbols and Pictorials to Communicate Hazard Information," in *Proceedings of the Interface 89—Sixth Symposium on Human Factors and Industrial Design in Consumer Products*, Human Factors Society, Santa Monica, CA, pp. 79–93.

Lehto, M. R., and Miller, J. M. (1986), *Warnings*, Vol. 1, *Fundamentals, Design and Evaluation Methodologies*, Fuller Technical Publications, Ann Arbor, MI.

Lehto, M. R., and Papastavrou, J. D. (1993), "Models of the Warning Process: Important Implications towards Effectiveness," *Safety Science*, Vol. 16, pp. 569–595.

Lehto, M. R., and Salvendy, G. (1995), "Warnings: A Supplement not a Substitute for Other Approaches to Safety," *Ergonomics*, Vol. 38, pp. 2155–2163.

Lerner, N. D., and Collins, B. L. (1980), "The Assessment of Safety Symbol Understandability by Different Testing Methods," PB81-185647, National Bureau of Standards, Washington, DC.

Lim, R. W., and Wogalter, M. S. (2003), "Beliefs about Bilingual Labels on Consumer Products," *Proceedings of the Human Factors and Ergonomics Society*, Vol. 47, pp. 839–843.

Lust, J. A., Celuch, K. G., and Showers, L. S. (1993), "A Note on Issues Concerning the Measurement of Self-Efficacy," *Journal of Applied Social Psychology*, Vol. 23, pp. 1426–1434.

Madden, M. S. (1999), "The Law Related to Warnings," in *Warnings and Risk Communication*, M. S. Wogalter, D. M. DeJoy, and K. R. Laughery, Eds., Taylor and Francis, London, pp. 315–329.

Mayhorn, C. B., and Goldsworthy, R. C. (2007), "Refining Teratogen Warning Symbols for Diverse Populations," *Birth Defects Research Part A: Clinical and Molecular Teratology*, Vol. 79, No. 6, pp. 494–506.

Mayhorn, C. B., and Goldsworthy, R. C. (2009). "'New And Improved': The Role Text Augmentation and the Application of Responses Interpretation Standards (Coding Schemes) in a Final Iteration of Birth Defects Warnings Development," *Birth Defects Research Part A: Clinical and Molecular Teratology*, Vol. 85, No. 10, pp. 864–871.

Mayhorn, C. B., and Podany, K. I. (2006), "Older Adults and Warnings," in *Handbook of Warnings*, M. S. Wogalter, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 355–361.

Mayhorn, C. B., Nichols, T. A., Rogers, W. A., and Fisk, A. D. (2004a), "Hazards in the Home: Using Older Adults' Perceptions to Inform Warning Design," *Journal of Injury Control and Safety Promotion*, Vol. 11, No. 4, pp. 211–218.

Mayhorn, C. B., Wogalter, M. S., and Bell, J. L. (2004b), "Are We Ready? Misunderstanding Homeland Security Safety Symbols," *Ergonomics in Design*, Vol. 12, No. 4, pp. 6–14.

Mayhorn, C. B., Wogalter, M. S., and Shaver, E. F. (2004c), "What Does Code Red Mean?" *Ergonomics in Design*, Vol. 2, No. 4, p. 12.

Mayhorn, C. B., Lanzolla, V. R., Wogalter, M. S., and Watson, A. M. (2005), "Personal Digital Assistants (PDAs) as Medication Reminding Tools: Exploring Age Differences in Usability," *Gerontechnology*, Vol. 4, No. 3, pp. 128–140.

Mayhorn, C. B., Wogalter, M. S., and Mendat, C. C. (2006), "The Matching Game: Educating Children about House-hold Hazards and Warning Symbols," in *Proceedings of the 16th World Congress of the International Ergonomics Association*, Maastricht, The Netherlands.

McLaughlin, A. C., and Mayhorn, C. B. (2011), "Avoiding Harm on the Farm: The Need for Human Factors in Understanding Safety in a Global Agricultural Context", *Gerontechnology*, Vol. 10, No. 1, pp. 26–37.

Mehlenbacher, B., Wogalter, M. S., and Laughery, K. R. (2002), "On the Reading of Product Owner's Manuals: Perceptions and Product Complexity," *Proceedings of the Human Factors and Ergonomics Society*, Vol. 46, pp. 730–734.

Otsubo, S. M. (1988), "A Behavioral Study of Warning Labels for Consumer Products: Perceived Danger and Use of Pictographs," *Proceedings of the Human Factors Society*, Vol. 32, pp. 536–540.

Penney, C. G. (1989), "Modality Effects and the Structure of Short-Term Verbal Memory," *Memory and Cognition*, Vol. 17, pp. 398–422.

Rhoades, T. P., Frantz, J. P., and Hopp, K. M. (1991), "Product Information: Is It Transferred to the Second Owner of a Product?" in *Proceedings of Interface '91'*, Human Factors Society, Santa Monica, CA, pp. 100–104.

Ringseis, E. L., and Caird, J. K. (1995), "The Comprehensibility and Legibility of Twenty Pharmaceutical Warning Pictograms," *Proceedings of the Human Factors and Ergonomics Society*, Vol. 39, pp. pp. 974–978.

Rogers, W. A., Lamson, N., and Rouseau, G.K. (2000), "Warning Research: An Integrative Perspective," *Human Factors*, Vol. 42, pp. 102–139.

Sanders, M. S., and McCormick, E. J. (1993), *Human Factors in Engineering and Design*, 7th ed., McGraw-Hill, New York.

Schroeder, D. G., Hancock, H. E., Rogers, W. A., and Fisk, A. D. (2001), "Phrase Generation and Symbol Comprehension for 40 Safety Symbols," in *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1470–1480.

Shannon, C. E., and Weaver, W. (1949), *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL.

Shorr, D. J., Ezer, N., Fisk, A. D., and Rogers, W.A. (2009), "Comprehension of Warning Symbols by Younger and Older Adults: Effects of Visual Degradation," in *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1598–1602.

Silver, N. C., and Braun, C. C. (1999), "Behavior," in *Warnings and Risk Communication*, M. S. Wogalter, D. M. DeJoy, and K. R. Laughery, Eds., Taylor and Francis, London, pp. 245–262.

Smith, J. J., and Wogalter, M. S. (2010), "Behavioral Compliance to In-Manual and On-Product Warnings," in *Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1846–1850.

Smith-Jackson, T. (2006a), "Culture and Warnings," in *Handbook of Warnings*, M. S. Wogalter, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 363–372.

Smith-Jackson, T. L. (2006b), "Receiver Characteristics," in *Handbook of Warnings*. M. S. Wogalter, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 335–344.

Smith-Jackson, T. L., and Wogalter, M. S. (2000), "Applying Cultural Ergonomics/Human Factors to Safety Information Research," *Proceedings of the International Ergonomics Association and the Human Factors and Ergonomics Society Congress*, Vol. 44, No. 6, pp. 150–153.

Smith-Jackson, T. L., and Wogalter, M. S. (2007), "Application of Mental Models Approach to MSDS Design," *Theoretical Issues in Ergonomics Science*, Vol. 8, pp. 303–319.

Sorkin, R. D. (1987), "Design of Auditory and Tactile Displays," in *Handbook of Human Factors*, G. Salvendy, Ed., Wiley, New York.

Strayer, D. L., Drews, F. A., and Johnston, W. A. (2003), "Cell Phone-Induced Failures of Visual Attention during Simulated Driving," *Journal of Experimental Psychology: Applied*, Vol. 9, No. 1, pp. 23–32.

Tam, T. W., and Greenfield, T. K. (2010), "Do Alcohol Warnings Influence Men's and Women's Attempts to Deter Others from Driving While Intoxicated?" *Human Factors and Ergonomics in Manufacturing Service Industries*, Vol. 29, No. 6, pp. 538–546.

Vanderheiden, G. C. (1997), "Designing for People with Functional Limitations Resulting from Disability, Aging, or Circumstance," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York, pp. 2010–2052.

Vigilante Jr., W. J., and Wogalter, M. S. (1997a), "On the Prioritization of Safety Warnings in Product Manuals," *International Journal of Industrial Ergonomics*, Vol. 20, pp. 277–285.

Vigilante, Jr., W. J.., and Wogalter, M. S. (1997b), "The Preferred Order of Over-the-Counter (OTC) Pharmaceutical Label Components," *Drug Information Journal*, Vol. 31, pp. 973–988.

Vigilante, Jr., W. J., Mayhorn, C. B., and Wogalter, M. S. (2007), "Direct-to-Consumer (DTC) Drug Advertising on Television and Online Purchases of Medications," in *Proceedings of the Human Factors and Ergonomics Society 51st Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1272–1276.

Viscussi, W. K., Magat, W. A., and Huber, J. (1986), "Informational Regulation of Consumer Health Risks: An Empirical Evaluation of Hazard Warnings," *Rand Journal of Economics*, Vol. 17, pp. 351–365.

Vredenburgh, A. G., and Helmick-Rich, J. (2006), "Extrinsic Nonwarning Factors," in *Handbook of Warnings*, M. S. Wogalter, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 373–382.

Weedon, B., Hellier, E. Edworthy, J., and Walters, K. (2000), "Perceived Urgency in Speech Warnings," *Proceedings of the International Ergonomics Association and the Human Factors and Ergonomics Society Congress*, Vol. 44, No. 3, pp. 690–693

Westinghouse Electric Corporation (1981), *Product Safety Label Handbook*, Westinghouse Printing Division, Trafford, PA.

Williamson, R. B. (2006), "Fire Warnings," in *Handbook of Warnings*, M. S. Wogalter, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 701–710.

Wogalter, M. S. (1999), "Factors Influencing the Effectiveness of Warnings," in *Visual Information for Everyday Use: Design and Research Perspectives*, H. J. G. Zwaga, T. Boersema, and H. C. M. Hoonhout, Eds., Taylor and Francis, London, pp. 93–110.

Wogalter, M. S. (2006a), "Communication-Human Information Processing (C-HIP) Model," in *Handbook of Warnings*, M. S. Wogalter, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, Chapter 5, pp. 51–61.

Wogalter, M. S., Ed. (2006b), *Handbook of Warnings*, Lawrence Erlbaum Associates, Mahwah, NJ ( CRC Press, Boca Raton, FL).

Wogalter, M.S., and Barlow, T. (1990), "Injury Likelihood and Severity in Warnings," *Proceedings of the Human Factors Society*, Vol. 34, pp. 580–583.

Wogalter, M. S., and Brelsford, J. W. (1994), "Incidental Exposure to Rotating Warnings on Alcoholic Beverage Labels," *Proceedings of the Human Factors and Ergonomics Society*, Vol. 38, pp. 374–378.

Wogalter, M. S., and Dingus, T. A. (1999), "Methodological Techniques for Evaluating Behavioral Intentions and Compliance," in *Warnings and Risk Communication*, M. S. Wogalter, D. M. DeJoy, and K. R. Laughery, Eds., Taylor and Francis, London, pp. 53–82.

Wogalter, M. S., and Feng, E. (2010), "Indirect Warnings/Instructions Produce Behavioral Compliance," *Human Factors and Ergonomics in Manufacturing and Service Industries*, Vol. 20, pp. 500–510.

Wogalter, M. S., and Laughery, K. R. (1996), "WARNING: Sign and Label Effectiveness," *Current Directions in Psychology*, Vol. 5, pp. 33–37.

Wogalter, M. S., and Laughery, K. R. (2005), "Effectiveness of Consumer Product Warnings: Design and Forensic Considerations," in *Handbook of Human Factors in Litigation*, I. Noy and W. Karwowski, Eds., Taylor and Francis, London.

Wogalter, M. S., and Leonard, S. D. (1999), "Attention Capture and Maintenance," in *Warnings and Risk Communication*, M. S. Wogalter, D. M. DeJoy, and K. R. Laughery, Eds., Taylor and Francis, London, pp. 123–148.

Wogalter, M. S., and Mayhorn, C. B. (2005a), "Perceptions of Driver Distraction Due to Cellular Phones by Cellular Phone Owners and Non-Owners," *Human Factors*, Vol. 47, No. 2, pp. 455–467.

Wogalter, M. S., and Mayhorn, C. B. (2005b), "Providing Cognitive Support with Technology-Based Warning Systems," *Ergonomics*, Vol. 48, No. 5, pp. 522–533.

Wogalter, M. S., and Mayhorn, C. B. (2008), "Trusting the Internet: Cues Affecting Perceived Credibility," *International Journal of Technology and Human Interaction*, Vol. 4, No. 1, pp. 76–94.

Wogalter, M. S., and Post, M. P. (1989), "Printed Computer Instructions: The Effects of Screen Pictographs and Text Format on Task Performance," in *Proceedings of Interface 89*, Human Factors Society, Santa Monica, CA pp. 133–138.

Wogalter, M. S., and Silver, N. C. (1995), "Warning Signal Words: Connoted Strength and Understandability by Children, Elders, and Non-Native English Speakers," *Ergonomics*, Vol. 38, pp. 2188–2206.

Wogalter, M. S., and Usher, M. (1999), "Effects of Concurrent Cognitive Task Loading on Warning Compliance Behavior," *Proceedings of the Human Factors and Ergonomics Society*, Vol. 43, pp. 106–110.

Wogalter, M. S., and Vigilante, W. J. Jr. (2003), "Effects of Label Format on Knowledge Acquisition and Perceived Readability by Younger and Older Adults," *Ergonomics*, pp. 327–344.

Wogalter, M. S., and Young, S. L. (1991), "Behavioural Compliance to Voice and Print Warnings," *Ergonomics*, Vol. 34, pp. 79–89.

Wogalter, M. S., and Young, S. L. (1994), "Enhancing Warning Compliance through Alternative Product Label Designs," *Applied Ergonomics*, Vol. 25, pp. 53–57.

Wogalter, M. S., Godfrey, S. S., Fontenelle, G. A., Desaulniers, D. R., Rothstein, P. R., and Laughery, K. R. (1987), "Effectiveness of Warnings," *Human Factors*, Vol. 29, pp. 599–612.

Wogalter, M. S., Allison, S. T., and McKenna, N. (1989), "Effects of Cost and Social Influence on Warning Compliance," *Human Factors*, Vol. 31, pp. 133–140.

Wogalter, M. S., Brelsford, J. W., Desaulniers, D. R., and Laughery, K. R. (1991), "Consumer Product Warnings: The Role of Hazard Perception," *Journal of Safety Research*, Vol. 22, pp. 71–82.

Wogalter, M. S., Brems, D. J., and Martin, E. G. (1993a), "Risk Perception of Common Consumer Products: Judgments of Accident Frequency and Precautionary Intent," *Journal of Safety Research*, Vol. 24, pp. 97–106.

Wogalter, M. S., Kalsher, J. J., and Racicot, B. (1993b), "Behavioral Compliance with Warnings: Effects of Voice, Context and Location," *Safety Science*, Vol. 16, pp. 637–654.

Wogalter, M. S., Barlow, T., and Murphy, S. (1995), "Compliance to Owner's Manual Warnings: Influence of Familiarity and the Task-Relevant Placement of a Supplemental Directive," *Ergonomics*, Vol. 38, pp. 1081–1091.

Wogalter, M. S., Laughery, K. R., and Barfield, D. A. (1997), "Effect of Container Shape on Hazard Perceptions," *Proceedings of the Human Factors and Ergonomics Society*, Vol. 41, pp. 390–394.

Wogalter, M. S., Magurno, A. B., Rashid, R., and Klein, K. W. (1998a), "The Influence of Time Stress and Location on Behavioral Compliance," *Safety Science*, Vol. 29, pp. 143–158.

Wogalter, M. S., Vigilante, W. J., and Baneth, R. C. (1998b), "Availability of Operator Manuals for Used Consumer Products," *Applied Ergonomics*, Vol. 29, pp. 193–200.

Wogalter, M. S., Conzola, V. C., and Vigilante, W. J. (1999a), "Applying Usability Engineering Principles to the Design and Testing of Warning Messages," *Proceedings of the Human Factors and Ergonomics Society*, Vol. 43, pp. 921–925.

Wogalter, M. S., DeJoy, D. M., and Laughery, K. R. (1999b), "Organizing Framework: A Consolidated Communication-Human Information Processing (C-HIP) Model," in *Warnings and Risk Communication*, M. S. Wogalter, D. M. DeJoy, and K. R. Laughery, Eds., Taylor and Francis, London, pp. 15–24.

Wogalter, M. S., Kalsher, M. J., and Rashid. R. (1999c), "Effect of Signal Word and Source Attribution on Judgments of Warning Credibility and Compliance Likelihood," *International Journal of Industrial Ergonomics*, Vol. 24, pp. 185–192.

Wogalter, M. S., Magurno, A. B., Dietrich, D., and Scott, K. (1999d), "Enhancing Information Acquisition for Over-the-Counter Medications by Making Better Use of Container Surface Space," *Experimental Aging Research*, Vol. 25, pp. 27–48.

Wogalter, M. S., Young, S. L., and Laughery, K. R., Eds. (2001), *Human Factors Perspectives on Warnings*, Vol. 2, *Selections from Human Factors and Ergonomics Society Annual Meetings* 1993–2000, Human Factors and Ergonomics Society, Santa Monica, CA.

Wolff, J. S., and Wogalter, M. S., (1993), "Test and Development of Pharmaceutical Pictorials," in *Proceedings of Interface 93*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 187–192.

Wolff, J. S., and Wogalter, M. S. (1998), "Comprehension of Pictorial Symbols: Effects of Context and Test Method," *Human Factors*, Vol. 40, pp. 173–186.

Young, S. L. (1991), "Increasing the Noticeability of Warnings: Effects of Pictorial, Color, Signal Icon and Border," *Proceedings of the Human Factors Society*, Vol. 34, pp. 580–584.

Young, S. L., and Lovvoll, D. R. (1999), "Intermediate Processing: Assessment of Eye Movement, Subjective Response and Memory," in *Warnings and Risk Communication*, M. S. Wogalter, D. M. DeJoy, and K. R. Laughery, Eds., Taylor and Francis, London, pp. 27–51.

Young, S. L., and Wogalter, M. S. (1990), "Comprehension and Memory of Instruction Manual Warnings: Conspicuous Print and Pictorial Icons," *Human Factors*, Vol. 32, pp. 637–649.

Young, S. L., Martin, E. G., and Wogalter, M.S. (1989), "Gender Differences in Consumer Product Hazard Perceptions," in *Proceedings of Interface 89*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 73–78.

Young, S. L., Brelsford, J. W., and Wogalter, M. S. (1990), "Judgments of Hazard, Risk and Danger: Do They Differ?" *Proceedings of the Human Factors Society*, Vol. 34, pp. 503–507.

Young, S. L., Wogalter, J. S., and Brelsford, J. D. (1992), "Relative Contribution of Likelihood and Severity of Injury to Risk Perceptions," *Proceedings of the Human Factors Society*, Vol. 36, pp. 1014–1018.

Young, S. L., Wogalter, M. S., Laughery, K. R., Magurno, A., and Lovvoll, D. (1995), "Relative Order and Space Allocation of Message Components in Hazard Warning Signs," *Proceedings of the Human Factors and Ergonomics Society*, Vol. 39, pp. 969–973.

Zwaga, H. J. G., and Easterby, R. S. (1984), "Developing Effective Symbols or Public Information," in *Information Design: The Design and Evaluation of Signs and Printed Material*, R.S. Easterby and H.J.G. Zwaga, Eds., Wiley, New York.

# CHAPTER 30

# USE OF PERSONAL PROTECTIVE EQUIPMENT IN THE WORKPLACE

**Grażyna Bartkowiak, Krzysztof Baszczyński, Anna Bogdan, Agnieszka Brochocka, Rafał Hrynyk, Emilia Irzmańska, Danuta Koradecka, Katarzyna Majchrzycka, Krzysztof Makowski, Anna Marszałek, Grzegorz Owczarek, and Jan Żera**
Central Institute for Labour Protection—National Research Institute
Warsaw Poland

## 1 INTRODUCTION

Personal protective equipment (PPE) belongs to a group of equipment that protects employees against dangers in the workplace. The decision to use such equipment must be preceded by all possible actions, both technical and organizational, aimed at eliminating or reducing the hazard to an admissible level. Despite the above rule, the use of PPE is still very common in many workplaces. This includes, in particular, mining, construction, transportation, and working in rooms with small capacity, for example, containers, manholes, and canals, and refers to all types of emergency actions.

A frequent cause of accidents in the workplace is failure to use PPE by the workers or their wrong selection for the level of risk connected with the occurring hazards. Workers' reluctance to use PPE may result from the equipment not being well fitted to the needs of a user and additional conditions connected with work organization in a specific workplace.

In this chapter, issues concerning the safety of specific protection equipment will be discussed with special attention on their types, considering their correct selection for the type of hazard, ergonomics, and rules of application in workplaces.

Awareness of the discussed issues will make it possible to prepare a management system for protection equipment in companies which should include the following elements:

- Risk evaluation that enables selection of equipment and protection class (i.e., hazard identification, the way it influences the body, the set hygienic norms exceeded)
- Workplace characteristics, including the type of occupational activities of a worker, microclimate, space limitation, necessity to move and communicate, speed of potential evacuation from a dangerous zone, additional hazards, for example, fire or explosion
- Participation of workers using this equipment in the process of selecting construction solutions
- Permanent training for the users, with particular stress placed on workers' motivation, increasing their awareness concerning not using protection, understanding the manufacturer's instructions, practical fitting and application, time limitations, and problems that may occur during use
- Selection of workers on the basis of psychomotor predispositions with particular emphasis on their current health
- Marking the zones where it is necessary to use respiratory protection equipment
- Ensuring correct maintenance and repairs
- Permanent monitoring through a system of audits in the field of appropriate selection, use, maintenance, and updating of training courses

## 2 SELECTION OF RESPIRATORY PROTECTIVE DEVICES FOR DIFFERENT TYPES OF WORKPLACES

Respiratory protective devices have a complex structure which depends on their purpose and scope of use. They protect workers against risks to life or dangers that can cause serious and irreversible damage to health and whose effects cannot be determined by users quickly enough. The need to ensure such high-performance protection imposes specific requirements for the protection parameters and on conducting proper training for users. Selecting which employees should use the equipment while working and during rescue operations is very important as well.

Because of the way they function, respiratory protective devices are divided into two main groups: purifying and isolating (breathing apparatus).

Filtering equipment cleans the air in the worker's breathing zone of the harmful substances present in the working environment in the form of aerosols (including bioaerosols), vapor, and gases.

This equipment can be divided into three main groups: particle-filtering, gas-filtering, and combined devices.

Particle-filtering respiratory devices are used in the working environment if there is pollution in the form of dust, smoke, or mist in excess of designated values corresponding to the maximum allowable concentration (MAC). They are available as filtering half masks, filters with facepiece (mask, half mask, quarter mask, mouthpiece), power-assisted devices with masks or half masks, and powered devices with loose-fitting facepieces (helmets, hoods).

Gas-filtering respiratory protective devices are used if in the working environment pollution occurs in the form of vapor and/orr gas.

The equipment is in the form of gas-filtering half masks, gas filters with facepiece (mask, half mask, quarter mask, mouthpiece), power-assisted devices with masks or half masks, and powered devices with loose-fitting facepieces (helmets, hoods).

Often the working environment may contain impurities in the form of vapor and/or toxic gases and aerosols (dust, fumes, mists). In such situations it is necessary to use the combined respiratory protective devices, which can filter both particles and gases.

Because of the way they operate and the way the air is supplied, breathing apparatus are of two types: air line and self-contained. The first group consists of compressed line and fresh-air line breathing apparatus. These apparatus may contain various types of facepieces depending on user needs and requirements: masks, hoods, face shields, helmets, and so on. The apparatus in the second group, the self-contained apparatus, operate in an open circuit (air-breathing apparatus) and a closed circuit (oxygen apparatus). Self-contained apparatus are completed only with the masks or mouthpieces. Such equipment is used in conditions of oxygen deficiency (below 19%) and in cases of unidentified air pollutants in unknown concentrations.

Efficiency of energetic processes of human being organism is only 20%. It means that 80% of energy turns into the heat and cause extra load to organism. This fact is especially important in an environment with increased humidity and temperature, where cooling of an organism is very difficult. Such ambient conditions appear in many workplaces where usage of respiratory protective devices (RPDs) is necessary. Therefore, it was necessary to conduct complex tests in the laboratory and at the workplace with selected types of respiratory protective equipment to make a complete and objective assessment of ergonomic parameters.

Practical performance tests are performed by participants wearing RPDs in accordance with instructions given by the manufacturer. Test subjects perform several exercises that simulate practical use of the device. Then participants are asked to assess the equipment and give their comments. Before and after tests carried out with and without the RPD, a psychological test is carried out to measure the mental burden resulting from the use of RPDs. Practical performance tests are performed at an ambient temperature of $16-32°C$ and relative humidity of $30-80\%$ accompanied by additional sound and light effects (simulating the real conditions of rescue operations) in order to assess the ability to communicate and the effects of light.

In both the workplace and the laboratory changes in temperature under the facepiece were found to be highest for completed half masks with combined filters, which are characterized by highest breathing resistance, and lowest for filtering half masks, which have the lowest resistance. At the same time, considerable differences in temperature were recorded in laboratory conditions as well as at workstations, which suggests that the simulation under laboratory conditions does not correspond to actual conditions. Tests on the energy expenditure of workers were also carried out at workstations and in the laboratory. In laboratory conditions, where the level of activities done was balanced, the energy expenditure reached 0.59 kcal/min and was 3.5 times less than at workplaces. In the workplace, where the energy expenditure changed depending on work methods, work intensity, work experience, and the kind of tools used, the energy expenditure value is different. The highest expenditure value was achieved for such work as pulling down walls and frameworks and the lowest by workers chrome plating ring pistons.

To comply with requirements of the quality of air inhaled by a wearer of a RPD (concentration of harmful substances below the level of the MAC, minimum 19% oxygen content in ambient air), it is necessary to evaluate the efficiency of the equipment and then estimate the multiplication factor that will reduce the contaminant concentration in the workplace environment, achieved as a result of using a any RPD device. The above task can be realized by calculating boundary values of total inward leakage for all kinds of respiratory protective equipment and then evaluating the nominal protection factor.

The protection factor can be described as the level of efficiency of a given respiratory device. It is expressed by a multiplication factor of reduction of contaminant concentration in the workplace when an adequate protective device is used. Selection of the device is based

on the assumption that the device should guarantee the wearer sufficient quantity and purity of air for breathing. The nominal protection factor (NPF) is expressed as the maximum percentage of inward leakage permitted for a given European Union (EU) standard class of RPD and is presented as a ratio:

$$\text{NPF} = \frac{100}{\text{Percent of allowable inward leakage}}$$

The NPF values for different types of respiratory protective equipment differ from actual values fixed in the workplace. Studies have shown that terms of use of RPD, such as air temperature, ambient humidity, and energy expenditure of the user, have an impact on the values of protection factors.

The Central Institute for Labour Protection— National Research Institute (CIOP-PIB) studies to determine the NPF value performed by Makowski and Majchrzycka (2004) and by Brochocka and Makowski (2005) were carried out in a climatic chamber where exercises were performed to assess the impact of ambient usage conditions on the health of workers using RPDs. The decision to keep the total inward leakage testing under conditions simulating the use of respiratory protection equipment was dictated by the need to ensure reproducible test conditions for all types of equipment, which due to the significant differences in the application (sprays, vapors, gases, oxygen deficiency) would not be possible in the workplace.

## 3 PERSONAL EYE PROTECTORS

There are four types of personal protectors for the eyes: protective glasses, protective goggles, face shields, and welding shields (the latter category includes shields, visors, goggles, and hoods).

Eye and face protectors should be used where the following dangers occur:

- Impact (e.g., splinters of solid bodies)— mechanical hazards
- Optical radiation (e.g., radiation arising from welding, solar glare, laser radiation)—a risk from physical factor; radiation

- Dust and gases (e.g., coal dust or aerosols of harmful chemicals)—risks from chemical and/or mechanical factors
- Drops and splashes of liquids (e.g., splashes resulting from flowing liquid substances)—risks from chemical and/or mechanical factors
- Molten metals and hot solids (e.g., molten metal splashes, resulting from metallurgical processes)—mechanical hazards
- Electrical arc (e.g., an electric arc formed during work under high voltage)—risk of physical factor, an electric shock and/or harmful UV rays arising from the emergence of the electric arc

Viewers, oculars, mesh, and filters are mounted in the above-mentioned categories of eye protectors (the filter category includes welding filters, ultraviolet protection filters, infrared radiation protection filters, glare protection filters, and laser radiation protection filters). Eye protectors may also be part of the RPD system (viewers in the air-cylinder apparatus) or head protector (shields mounted onto industrial safety helmets). All eye protectors consist of the viewing part (viewers, oculars, mesh, and filters) and a frame (for glasses and goggles) or a body together with the harness (for the guards).

Because many professions benefit from measures that protect the eyes, the demand for these products has resulted in upgrades to eye protector designs.

The components for the frame and temples of personal eye protectors are made of high-quality plastic, which does not cause allergic reaction when in direct contact with the user's skin. Frames and temples have numerous parts that allow adjustment to get optimal matching to the user's head.

Protective glasses that show significant improvement in ergonomics of use are systems that allowadjustment in the length and depression angle of the tamples. Such a system is shown in Figure 1.

Another feature which has had a significant impact on the ergonomics of eye protector use is resistance to fogging. The problem of creating a layer of moisture on solid surfaces, glass and plastics in particular, relates to glasses, goggles, and face shields used under conditions of high air humidity or as a result of perspiration of



(a)                    (b)
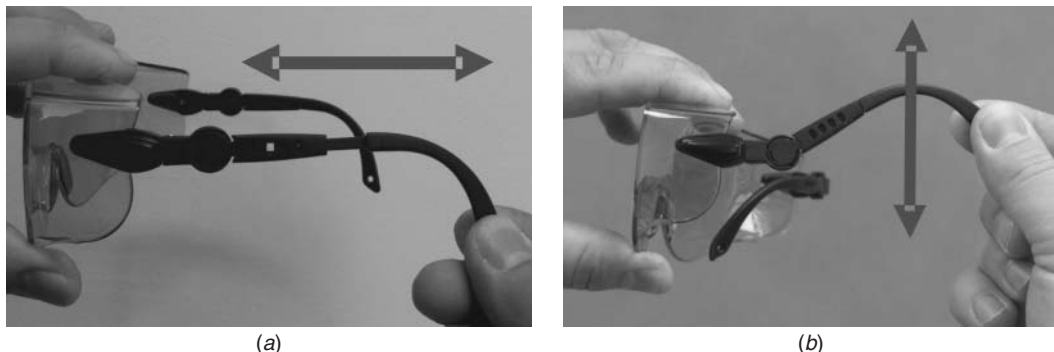
**Figure 1** Adjustment of (a) length and (b) depression angle of tample in protective glasses.

the user (the problem applies to protective goggles which tightly adhere to the user's face and ventilation systems built in goggles are not able to remove excess perspiration) and moving between areas with different temperature and humidity. Fogged viewers and oculars in the eye protectors make the work less comfortable and often impossible to do regardless of whether the moisture is deposited on the surface of glasses as liquid drops or mist.

For eye protectors where the body is mounted onto the head strip, it is very important to have the head strip well matched so that it can firmly support the eye protector and not cause too high compression. Currently widely used control systems allow a seamless fit between the head strips and the user's head.

When designing the components of eye protectors mounted onto the user's head, one must consider convenience, in terms of fitting and adjustment and ventilation systems, how perspiration is removed through materials adjacent to the user's forehead, and so on.

Another advanced technology that has improved ergonomics and safety of use is the automatic welding filter (Kubacki et al., 2001). Automatic welding filters are being increasingly used for eye protection during arc welding . They are auto darkening (transition from light to dark state) due to the initiation of the welding arc. In the light state of the filter the welder can preview welding elements. In the dark state complete protection from the harmful rays of the welding arc is assured. Automatic welding filters are mostly installed in welding masks.

Oculars and inorganic glass filters are replaced with high-quality plastic components. This substantially reduces the weight of the eye protectors. Currently one of the materials commonly used for the construction of oculars and filters is polycarbonate. This material features very high mechanical resistance and natural ability of UV absorption. Polycarbonate can also be easily colored, thus giving it good filtration properties.

To protect the eyes from the harmful optical radiation occurring in many workplaces, it is necessary to use PPE with appropriate protective filters. Modification of the spectral transmission characteristics of filters is designed to ensure an adequate level of protection while maintaining visible-light transmission properties to an extent which will allow the visual activity as defined in the work process.

Protective filters used in eye protectors are divided into welding filters (to protect against radiation emitted during welding and allied techniques), UV protection filters (to protect eyes when working with sources of ultraviolet radiation used for medical, technical, and scientific purposes), infrared radiation protection filters (to protect against glare and short-wavelength near-infrared radiation), sun glare protection filters (for protection against solar radiation, intended primarily to protect the human eye from glare; in addition to protection from required absorption of visible radiation, may also be intended to protect the eye from ultraviolet and infrared radiation), and filters for protection against laser radiation (to protect eyes from laser radiation at wavelengths from 180 to 1000 μm).

## 4   PROTECTIVE HELMETS: SELECTION

Workers are exposed to a number of hazards depending on the specific characteristics of the work site and the operations performed. The most serious and common hazards include those associated with mechanical factors, for example, a fall from a high work site; impacts caused by falling objects; impacts caused by sharp, protruding elements, which may result in punctures or cuts; impacts caused by mobile parts of machines, posing the risk of dragging into a machine and crushing; and abrasions caused by coarse, rough elements and falls on slippery and uneven surfaces.

The aforementioned factors can cause injuries to various parts of the human body. However, considering the consequences of potential accidents, the head should be treated with particular care. Hazards due to mechanical factors are present primarily in such industries as mining, civil engineering, transport, warehousing, communications, trade, energy, and gas and water supply. Taking into account the statistical data concerning accidents at work, the most frequent causes of head injuries include impacts caused by falling elements and sharp, hard objects. The consequences of such events depend mainly on the kinetic energy of the impact as well as the shape and hardness of the material making up the object which comes in contact with the head. The injuries caused by mechanical factors may involve the skin of the scalp, the cranial bones, the brain, and the cervical vertebrae. In extreme cases, such injuries may lead to permanent disability or even death.

Considering the potential consequences of head injuries, it is the duty of both the employer and the employee to eliminate the hazard. Hazards should be eliminated by appropriate safety measures taken at the work sites or by appropriate organization of work. If such solutions of the problem are impossible, the workers should be equipped with suitable PPE. However, it should be remembered that the use of such equipment, in this case protective helmets of various types, does not eliminate hazardous factors but rather only reduces the severity of their effects.

To ensure appropriate protection of the user's head against mechanical factors, a protective helmet must be selected correctly from among various types. The primary factors to be considered include:

- Specific factors against which protection is needed (e.g., central impacts, side impacts, lateral forces)

- Temperature range characterizing the usage conditions

- Regulation range as related to the user's head dimensions

- Presence of other hazards (e.g., electric shock, high temperature)

- Other personal protective devices and additional equipment to be used (e.g., eye and face shields, hearing protectors, lamps mounted on the helmet)
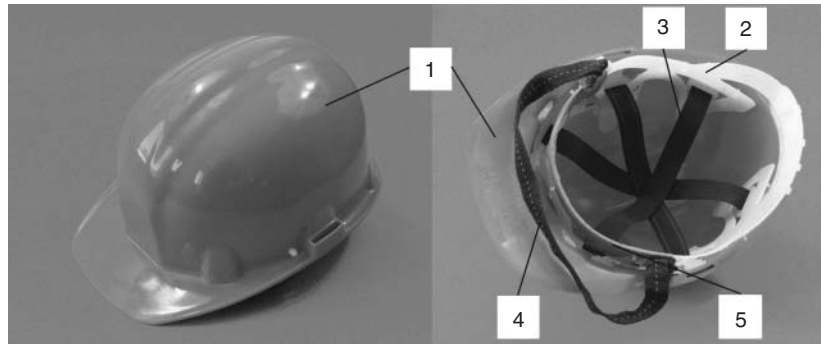
**Figure 2** Industrial safety helmet: 1, shell; 2, headband with nape strap; 3, cradle; 4, chin strap; 5, sweatband.

- Activities which may cause the helmet to fall off the head (e.g., use of the helmet together with fall-arresting equipment)

Industrial safety helmets (Figure 2) provide a basic and most common means of protection of workers' heads in the working environment. Irrespective of their construction, the following two elements are found in such helmets: the shell, and the harness.

The helmet shell, most often made of polyethylene or acrylonitrile butadiene styrene (ABS), constitutes its external part, which gives the helmet its characteristic shape. It is designed primarily to receive the impact, absorb part of the energy, and transfer the rest of it to the harness. The shell also prevents direct contact of a hazardous object with the user's head. Depending on the specific structural solution, the shell can be equipped with a visor, a brim, a rain gutter, vents, attachment devices for eye and face shields as well as hearing protectors.

The harness constitutes the internal part of the helmet in the form of a system of straps made of textile tapes or polyethylene which rest on the user's head. It is connected with the helmet shell by appropriate attachment devices. The main task of the harness is to absorb the energy of an impact acting on the shell and distribute the forces over as much of the head as possible. It is noteworthy that a helmet with harness connected to the shell close to its rim and not equipped with any additional shock-absorbing lining (protective padding) provides no protection against side impacts. This phenomenon was explained by Baszczyński (2002) and Korycki (2002). Helmets with appropriately stiff shells protect the head to some extent against lateral forces.

The headband, encircling the head at the level of the forehead in front and skull base at the back, together with the harness, ensures stable positioning of the helmet on the user's head. The headband is equipped with two mechanisms that allow the wearer to adjust the head circumference as well as enhance its stability on the head. In most industrial protective helmets the headband is equipped with a sweatband.

Industrial protective helmets can be equipped with additional elements, for example, a chin strap to prevent the helmet from falling off the head and attachment devices for eye and face protectors.

Numerous work sites do not pose the risk associated with impacts caused by falling objects but do pose the risk of superficial head injuries resulting from bumps against construction elements. In such a situation, industrial protective helmets should not be used. Rather what should be used are industrial bump caps, in which case the worker is not exposed to discomfort due to pressure exerted by the harness and the headband, load exerted on the muscles of the neck by additional weight on the head, and impaired ventilation of the upper part of the head.

The most important features of industrial bump caps, in comparison with industrial protective helmet structures, are significantly lower weight and smaller size.

Industrial conditions include work sites where the risk of mechanical injuries is so high that the industrial protective helmets fail to ensure an appropriate level of protection. Such work sites can be found, for example, in mining and construction sites. In such a situation, the workers should be equipped with high-performance industrial safety helmets. As far as protection against mechanical factors is concerned, these helmets have the following characteristics in comparison with industrial protective helmets:

- Provide the same level of shock absorption (i.e., force transmited to the user's head) for impacts with twice the energy
- Protect the head against both central impacts (at the highest point of the shell) and lateral impacts—from the front, back, and sides
- Provide higher level of protection against impacts exerted by sharp-tipped objects

The shell and the harness are also the main elements of high-performance industrial safety helmets. The method most commonly used to improve the shock-absorbing properties and protection against side impact involves introduction of appropriate lining materials absorbing the impact energy and thus reducing the forces acting on the user's head. Such a lining is usually made of foam possessing appropriate force deformation

characteristics, for example, high-density polyurethane and foamed polystyrene.

To ensure appropriate protection of the user's head against mechanical factors, a protective helmet must be selected, fitted, and used correctly. Selection means appropriate choice from among the three types available (industrial bump caps, industrial safety helmets, and high-performance industrial safety helmets) as well as various structural solutions. Protection against other hazardous factors, such as electric shock, molten metal splashes, and high temperatures, should also be taken into consideration. Fitting involves appropriate regulation, for example, circumference of the head band, wearing height, length of the chin strap, and adjusting the helmet to the dimensions of the user's head. Correct usage involves following the manufacturer's instructions on the conditions and modes of use, recommended maintenance and storage methods, and terms for phasing the helmets out of service.

## 5  HEARING PROTECTION DEVICES

Hearing protection devices (HPDs) should be applied in conditions in which there are no other technical means to reduce noise levels or when it is not economically feasible to reduce noise at the source according to European Directive (2003) 2003/10/EC. The HPDs are divided into two broad categories: earmuffs and earplugs. Earmuffs are placed around the ears and provide an acoustic barrier to sound. Earplugs fit into the ear canal to block its entrance and keep noise from entering. Some other head protecting devices such as military helmets and sand-blasting helmets usually include in their design parts to protect hearing.

Earplugs include foam, premolded, formable, custom molded and semi-insert earplugs (Figure 3). Foam earplugs (Figure 3a) made from slow-recovery material, usually polyurethane, have to be rolled and compressed to the diameter which would allow for easy insertion into the ear canal. Expansion of the earplug leads to a very good sealing of the ear canal against external sound.

Premolded earplugs (Figure 3b) are formed from a flexible material such as thermoplastic elastomer (TPE) in different shapes with flanges, rings, and cups on a stem that is held with the fingers while inserting the earplug into the ear canal. Formable earplugs (Figure 3c) are made from plastic materials, usually mixtures of wax with cotton and mineral wool. Such earplugs are formed by the user who inserts the material into the ear canal. These earplugs are often not used in the workplace and are more popular in the consumer market (Berger, 2000). Custom-molded earplugs (Figure 3d) are more expensive but are more comfortable. These earplugs are formed for an individual user and are separately fitted to the left and right ear canal shapes. They are often used when comfort is the primary prerequisite. Earmuffs (Figure 4) consist of rigid plastic earcups sealed around the ears by foam or, more rarely, with fluid-filled or partly fluid-filled cushions. Earmuffs are held on the ears by a plastic or a metal headband placed on top of the head (Figure 4a), behind the head (Figure 4b), or under the chin (Figure 4c) to allow for use in combination with other personal protective devices placed on the head, for example, helmets. For helmets, however, helmet mounted earmuffs (Figure 4d) are most often used integrated with the helmet by a spring. Regardless of passive designs, special kinds of HPDs are equipped with electronic circuits to control sound attenuation, provide communication, and apply active noise reduction.

Attenuation of earmuffs is affected by their cup volume, mass, headband force, diameter of opening in the cushion, and construction materials (Berger, 2000). Large volume enclosed by an earmuff improves



**Figure 3**  Earplugs: (a) foam, (b) premolded, (c) formable, and (d) custom molded.
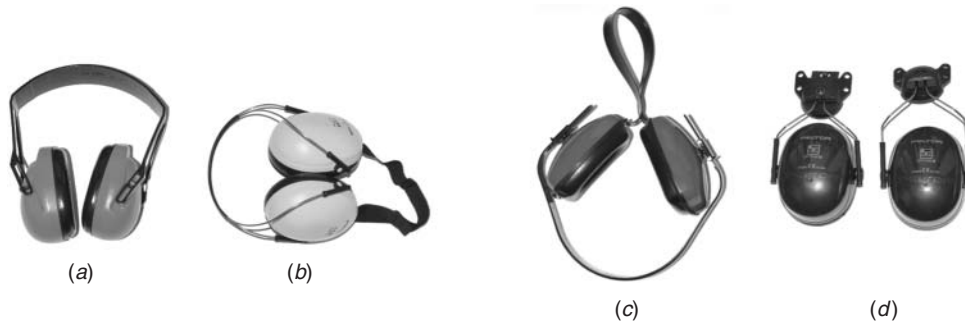


**Figure 4**  Earmuffs: (a) top of head headband, (b) behind the head headband, (c) under the chin headband, and (d) attached to helmet.

attenuation in the low-frequency range, while dumping acoustic material in the earcup absorbs high-frequency sound energy (Berger, 2000).

HPDs currently on the market are labeled with attenuation data obtained by the real ear at threshold (REAT) method, standardized by international and national standards [e.g., International Organization for Standardization (ISO), 1990; American National Standards Institute (ANSI), 2008]. In the REAT method, HPD attenuation is determined for frequencies spanned in octaves from .125 to 8 kHz as the difference between the hearing thresholds in two conditions: When an HPD is not worn and when it is worn. The average attenuation calculated for a group of subjects and several samples of the HPD represents the attenuation. To estimate the attenuation values protecting 84 or 98% of the population one or two standard deviations are subtracted from the mean attenuation. The value obtained by subtracting one standard deviation is called the approved protection value (APV) and is used to label the HPDs according to the ISO (1994) standard.

While calculations in octave-bands are considered the most accurate method to estimate the sound level under the HPD, the ISO (1994) standard introduces a single number rating (SNR) as an index for overall HPD attenuation and H, M, and L parameters for high-, mid- and low-frequency content noise, respectively. In the United States, the noise reduction rating (NRR) is used as a single number index describing the HPD attenuation,. Despite some differences (generally the NRR is 3.5 dB lower than the SNR), the basic idea of using the NRR or SNR is to estimate the A-weighted sound level under the HPD from the C-weighted sound level of noise measured outside the HPD.

Attenuation of HPDs may also be determined with the use of various objective methods involving measurement with a microphone. In the microphone in real ear technique (MIRE; ISO, 2002), the microphone is placed in the subject's ear at the entrance to the ear canal. This method is considered an objective counterpart of the REAT method (Berger, 2000). Other technics involve the use of acoustic couplers such as an artificial test fixture (ATF; ISO, 2007) or an acoustic manikin (ISO, 2004) equipped with an internal microphone and designed to replicate with some simplification the dimensions and shape of the human head and/or the torso. Measurements with the ATFs or manikins are suitable in all conditions in which there is a risk of exposure to high-level sound such as in impulse noise measurements (Żera and Młyński, 2007).

Bone conduction is the factor limiting protection provided by earmuffs or earplugs. Even if the ear canal and air sound transmission are completely blocked by the HPD, sound reaches the inner ear through the skull at a level of 40–55 dB lower than that transmitted through the ear canal (Berger et al., 2003).

In the workplace, attenuation of an HPD may be much lower than measured in the laboratory by the REAT method. To accommodate for this, the ANSI S12.6-1997 standard (see revised version ANSI/ASA, 2008) introduced an alternative *method B* which takes into account fitting of the HPD by an inexperienced user;

also see the ISO (2006) technical specification. Participation by inexperienced subjects yields lower attenuation values, better corresponding to attenuation obtained in real-life application of HPDs.

Improper and interrupted fit of an HPD can reduce effective attenuation to 60% (earmuffs), 40% (foam earplugs), and 25% (other than foam earplugs) of the labeled value (Berger, 2000). Air leaks resulting from improper earmuff fits by users, reducing attenuation by more than 10 dB in a wide frequency range, are the major cause of the difference between attenuation measured in the laboratory and in the real world. An accepted way of estimating real-world attenuation is derating the values determined in the laboratory. For instance, in the United States, NRR values are derated to 0.75, 0.5, and 0.25 of the value measured in the laboratory for earmuffs, foam earplugs, and other earplugs, respectively. In Germany, 5 dB is subtracted from the value of attenuation on the label for earmuffs, 3 dB for foam, and 9 dB for other types of earplugs. In general, the strategy is to accept only a part of the attenuation value estimated in laboratory conditions.

Factors of primary importance in the choice of an HPD are comfort, motivation to use, conditions for speech communication, and the ability to receive auditory signals. For instance, earplugs are comfortable in hot climate owing to their small dimensions. Earmuffs are convenient for intermittent use. Small earmuffs may often provide better comfort than large earmuffs. Thus, unnecessary use of large overprotecting earmuffs [European Committee for Standardization (CEN), 2004a] should be avoided. It has to be stressed that it takes time and requires training to get used to earplugs and earmuffs. Wearing an HPD may cause a boost in low-frequency sounds, a change in the perception of a person's own voice, increased physiological noise, and changes in sound quality of speech and other sounds.

It is often of concern how the use of HPDs influences the ability to understand speech and audibility of warning signals in the work environment. For speech, excessive levels decrease speech intelligibility due to the increase of nonlinear effects in hearing. Therefore, in many cases decreasing very high levels of both speech and noise may improve intelligibility, even if the signal-to-noise ratio remains unchanged. Whether the use of HPDs improves or deteriorates the ability to understand speech and receive the warning signals depends on the level, the sound spectrum of noise, the attenuation characteristics of the HPD, and the hearing threshold of the person who uses the HPD.

## 6 INFLUENCE OF THERMAL ENVIRONMENT AND PROTECTIVE CLOTHING ON THERMAL CONDITION OF THE HUMAN BODY

As a result of metabolic conversion, the human body produces energy that is transferred to a person's activity to maintain vital functions and warm the body (Fanger, 1970; Gagge et al., 1971). Human beings are warm-blooded organisms; therefore, the internal temperature of the body should be $37 \pm 0.3°C$. An increase in
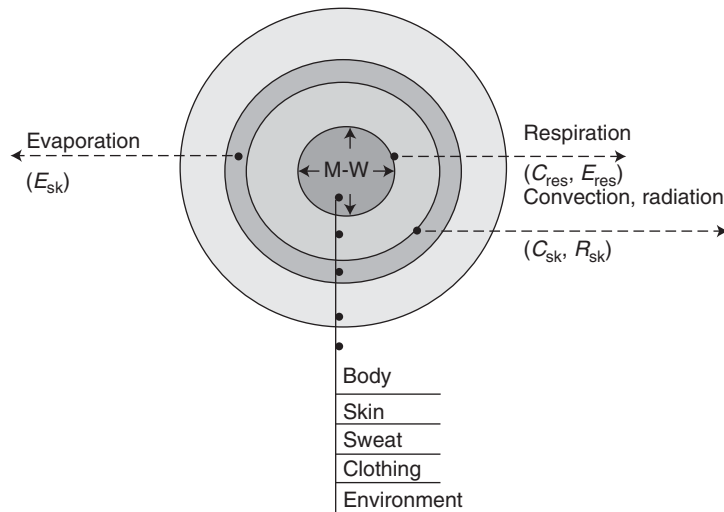
**Figure 5**  Heat transfer processes between human and environment.

internal temperature above $1.0°C$ may cause thermal stress and consequently lead to hyperthermia, while lowering internal temperature below $36.0°C$ may lead to loss of consciousness and hypothermia. In both cases, too high/low internal temperature can lead to death; therefore, the main task for the proper functioning of the human body is to maintain a constant internal temperature. Internal temperature depends on the heat exchange between the body and the environment. A diagram of heat exchange is shown in Figure 5 and presented below in the form of an equation [American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE), 2009]:

$$M - W = C_{sk} + R_{sk} + E_{sk} + C_{res} + E_{res} + S$$

where:

$M$ = metabolic heat production, $W/m^2$
$W$ = external work rate, $W/m^2$
$C_{sk}$ = convective heat transfer from the skin, $W/m^2$
$R_{sk}$ = radiation heat transfer from the skin, $W/m^2$
$E_{sk}$ = evaporative heat transfer from the skin, $W/m^2$
$C_{res}$ = convective heat transfer from respiration, $W/m^2$
$E_{res}$ = evaporation heat transfer from respiration, $W/m^2$
$S$ = heat storage, $W/m^2$

Some of the heat produced by the body is converted to external work, especially for the performance of hard physical work. The rest of the heat is dissipated into the environment (in a quantity which depends on the cooling power of the ambient environment) from the skin surface and through the respiratory tract; the remaining portion of heat is stored in the body, which may lead to a rise or a fall in internal temperature. Heat is transferred from the body to the ambient environment through different physical phenomena: convection, radiation, and evaporation of moisture. The amount of heat exchanged between the human and the environment is affected by the following:

- Environmental conditions, corresponding to the cooling power, for example, temperature, velocity and humidity of the air, radiation temperature of the room surfaces, and water vapor pressure
- Individual conditions depending on the nature of the work and the clothing, for example, the activity (determining metabolic heat production) and the thermal insulation of clothing, which is a barrier constricting heat dissipation

At many workstations protective clothing is required to protect workers against the thermal environment but also against other physical and chemical stress factors. It is a requirement for the safety of industrial workers that protective clothing offer sufficient protection. However, it is also important that PPE meet certain ergonomics requirements so that protection is not compromised by increased physiological or mental strain, impaired performance, or increased discomfort (Holmer, 1995).

Clothing creates a kind of obstruction in the heat exchange between the human body and the environment. In moderate and cold thermal environments, in most cases clothing protects against excessive heat loss from the body and acts as a regulating mechanism that helps maintain optimal body temperature.

In each kind of thermal environment, the thermal balance of the human body depends on three basic parameters: thermal conditions, intensity of activity, and thermal insulation of clothing.

In a cold environment, clothing with required insulation should protect against hypothermia and a decrease in internal temperature of no more than $1.0°C$, which means down to $36.0°C$ (CEN, 2004b). In a cold

environment clothing should be thicker and its evaporation resistance higher. Moisture accumulation inside clothing, because of its absorption and condensation, decreases insulation. For convective heat loss from the body, maintaining a high temperature gradient between body core and skin surface is a natural strategy to keep clothing dry and perspiration to the required minimum.

In most cases protection against physical and chemical factors means using protective clothing made of fabrics that are highly impermeable to moisture. Thus, suitable protection is often achieved by disrupting the body's thermal balance. Body overheating is the result of using this kind of clothing. More impermeable clothing means there is a higher risk of discomfort and a greater thermal load for the users (White and Hodous 1987), and exposure time in this kind of clothing is from several to a few dozen minutes, depending on work intensity and the parameters of the environment (Marszałek and Sawicka, 1997).

An important condition of maintaining comfort during work in protective clothing is to eliminate condensation in underclothing and consequently perspiration on the skin and the internal layers of the protective clothing. Developments in textile production technology, including barrier materials, enabled significant progress in the construction of protective clothing. Some materials that acts as barriers to hazardous factors, are water vapor permeable, and provide removal of excessive heat from the user's body. Materials used in protective clothing include laminates of fabrics and water vapor–permeable membranes as well as microporous coatings.

Currently manufacturers of clothing that provides comfort are focusing on high-performance underwear that transfers heat and sweat from the skin to the environment. Specialized underwear can significantly reduce the discomfort of protective clothing; it is especially recommended for use under tight protective clothing (Bartkowiak, 2000).

Another method of reducing moisture under tight protective clothing and the thermal discomfort is application of nonwoven inserts with high-sorption fibers (Bartkowiak, 2006) which absorb large amounts of liquid in relation to their mass.

Improvement in the physiological parameters of tight protective clothing could result from wearing a vest with a phase change material (PCM). Cooling and ventilation systems are another way of reducing the discomfort of working in protective clothing. Two types of systems are used to cool the body in oppressive working conditions: passive systems (vests with ice) and systems with forced circulation of cooling liquid or air.

During intensive work in tight protective clothing or in protective clothing in a hot environment, proper organization of the work is very important, for example, taking breaks. The working time and duration of the breaks depend on the intensity of the work, the temperature of the working environment, and each worker's characteristics. Workers should be healthy, with correct blood pressure, high physical efficiency, and an efficient sweating system. People working in protective clothing, especially tight protective clothing, and in a hot

environment should regularly replenish the liquids in their bodies.

A new type of clothing is the subject of much research and the focus of users, that is, so-called smart or intelligent clothing. Intelligent textile fabrics may be divided into two groups:

- Those that change their properties under the influence of certain stimuli
- Those thar are integrated with electronics

The products in the first group receive stimuli directly from the human body or the environment and react with significant physical, chemical, and biological changes and in many cases are reversible. Active materials are stimulated by tension, the electromagnetic field, temperature, humidity, UV or IR radiation, or chemical substances.

The other group of intelligent textiles consists of conductive materials designed to transfer electric signals and the so-called e-textiles in which microelectronic devices are integrated with textiles, offering products with additional functions concerning information and communication.

## 6.1 Phase Change Materials

Phase change materials are able to change their phase in the phase change temperature range. They are able to absorb, store, and release large quantities of energy in the form of latent heat.

In the form of capsules, PCMs may be incorporated into a textile in various ways. Currently, fibers with microcapsules of PCM are produced. They are totally surrounded by a polymer and permanently enclosed in the fiber. PCM can also be introduced into textiles through imbuing, printing, coating, or spraying.

Because protective clothing causes heat load, attempts have been made to use PCMs in protective clothing in order to cool the user's body. Research in this field where PCM was added to the polymer coating resulted in improvement in comfort for clothing protecting against chemicals. Due to the higher physiological comfort obtained, the time that protective clothing could be used was extended.

Research on the application of PCM for thermoregulating the microclimate under tight protective clothing is being conducted at the Central Institute for Labour Protection in Poland. Two types of clothing with PCM were prepared to be used under protective clothing: waistcoats and underwear with viscous fibers including PCM and waistcoats with PCM macrocapsules placed in special channels of suitably prepared knitted material. Using waistcoats with PCM under tight protective clothing for protection against chemicals extended the time that such clothing could be worn (Bartkowiak et al., 2010).

## 6.2 Shape Memory Materials

Shape memory materials are materials that, under the influence of certain stimuli, return from their present shape to the original one, that is, the shape that has been "remembered." The effect of shape memory is seen mostly in shape memory alloys and polymers.

Shape memory alloys are a unique class of metal alloys that can change their shape when heated to a certain temperature. There were attempts to use shape memory alloys in clothing that protects against high temperature, flame, heat radiation, and cold. Using materials that at lower temperatures increase their volume enables the construction of clothing that is thinner and more ergonomic and is a better insulator.

Polyurethanes with a shape memory that have a glaze temperature greater than 55°C may be used in protective clothing, for example, in the chemical, metallurgy, and food industries. They were also found useful in intelligent waterproof and water vapor-permeable membranes for clothing protecting against bad weather.

## 6.3 Integration of Electronic Microsystems with Textiles

Integration of electronic microsystems with textiles makes it possible to create products that can be used in protecting health and in medicine, safety and rescue, industry logistics, and sport. Clothing with electronics, which monitors and registers heart rate, number of breaths, and skin temperature has been designed for athletes. Clothing with an installed GPS system and an electronic compass and altimeter has been also constructed.

Of key importance for the integration of microelectronics and textiles are technologies that enable textiles to function as electronic interfaces. Textile interfaces can be implemented into clothing together with mobile electronic equipment. E-textiles are used in medicine to monitor life functions, for example, for infants and bed-ridden patients. The lifeshirt, with sensors that enable constant monitoring of pulse, breath, and temperature while patients sleep and perform everyday activities, is an example of electronic textiles in medicine application.

Electronics finds practical use in protective clothing particularly in extreme conditions. Clothing with electronic microsystems that monitor the user's physiological parameters as well as the level of danger has been developed. Particulary it has been dedicated for rescue teams, for example, firefighters. Such clothing monitors the physiological state of a firefighter and the environment's conditions

In a cold microclimate, when a worker's activity and the amount of heat emitted increase, passive protective clothing will not provide the workers thermal comfort. Suitable protection and comfort can be obtained only through an active shield, which changes its insulation depending on changes in the outdoor environment and the user's metabolic rate. Clothing equipped with integrated heating systems ensures precise undergarment temperature regulation. The protective clothing actively reacts to temperature changes and changes its heat insulation so as to provide the user with adequate warmth (Kurczewska and Leśnikowski, 2008).

## 7 PROTECTIVE GLOVES IN THE WORKPLACE

Protective gloves are used to protect the hands in the workplace. They can protect the entire hand or part of the hand as well as the forearm and arm. Protective gloves also include gloves with no fingers and products that protect the fingers.

They can also protect against mechanical factors, heat and fire, cold, chemicals and microorganisms, electric shock (isolation while working at a voltage), ionizing radiation, radioactive contamination, and mechanical vibrations.

For protection against mechanical factors a wide range of gloves are made of chain mail (designed to protect against cuts and stabs by sharp knives); p-aramid yarns (Kevlar®, Kevlar® KleenTM, Kevlar® PlusTM, Kevlar® Amor, Twaron®, Twaron® Premium Line); core yarns where the core can be made of stainless steel or textile yarns (high resistance to cutting) and the sheath is made of textile yarn; polyethylene (Dyneema®, Spectra®, Spectra® GuardTM, Spectra® GuardTMCX); and glass fiber mixed with yarn and other materials (cotton, polyamide, polyester, polyurethane) (Stefko, 2009).

Specialized gloves should be used to protect against severe mechanical injuries such as chain mail gloves to protect the hands against cuts and punctures by hand knives.

Gloves that protect against heat are usually made of fabric or knitted from yarn or fibers, such as Kevlar®, Nomex®, Twaron®, Preox®, PBI, PBI/Kevlar®, Basofil® or cotton yarn, woolen impregnated nonflammable substances, heat-resistant leather, and fabric from glass fiber yarn. Depending on the type of yarn, weave or knit fabric, and number of layers of materials used in the construction, gloves with various protective properties are received. Examples are knit gloves that protect against burns during brief contact with flame or hot objects with a temperature of 250°C made of yarn: cotton-impregnated nonflammable substances, mixed yarns with polyester yarns, and frotte-type cotton yarns. Another example is gloves of woven fabric made of glass fiber yarn lined with nonflammable cotton that protect workers' hands in the metallurgical industry.

The outer layer of gloves that protect against cold is mostly of cow leather and fabric. Gloves made of Thinsulate® yarns and fibers also provide good thermal isolation. Gloves can also be made entirely of plastic or rubber as well as polymer covered knit and woven fabrics. Nonwoven and knit fabrics contribute to thermal insulation. Some gloves also contain waterproof or vapor-permeable membranes (GORE-TEX®, HYDROTEX®, OSMOSIS®, TEXA-POR®, NO-WET®, SYMPATEX®, POWERTECH®).

For protection against chemicals tight gloves made from various synthetic rubbers (e.g., natural rubber, synthetic rubbers, polychloroprene, polyacrylonitrile, butyl, Viton) or plastics (e.g., polyvinyl chloride, polyvinyl alcohol, polyethylene, Hypalon) should be used (Irzmańska et al., 2010).

Gloves that are not made entirely of plastic or rubber or fabric or knit fabric entirely coated with polymer should not be used for protection against chemical agents. All-polymer or all-rubber gloves may decrease perspiration and cause discomfort. Moreover, allergies

to the components of the rubber mixtures may occur as a result of direct or indirect contact with human skin.

At the same time, depending on the degree of mechanical wear and degradation of the material which is in direct contact with the chemical, the barrier of the material may weaken. In addition, gloves of the same polymer but made by different manufacturers can have—and often do have—different protective properties. The permeation time can also be different for gloves made of the same material but produced by different manufacturers. The level of protection of a material depends on many factors, factors that are sometimes not taken into account when assessing a laboratory.

When selecting the appropriate protective gloves, it is important to use the information supplied by the manufacturer. The protective properties given by the manufacturer are the result of laboratory tests of various parameters on the basis of a set of standards. It should be noted that gloves do not provide protection against all harmful and hazardous factors or for an indefinite period of time. Terms of use and storage and maintenance of gloves impact the protective properties and can reduce the amount of time that they can be used. Changes in the glove's material should be a signal to immediately stop their use.

## 8 FOOTWEAR: COMFORT OF USE

In recent years, lifestyle changes and the fact that more people are using personal protective equipment have resulted in increasing demands on their performance. The popularity of comfortable footwear has increased the demand for comfortable safety footwear. Intensive development of advanced materials that began with sports shoes has resulted in the introduction of many new foot and leg protectors. Modern protective footwear for professional usemust also meet requirements of hygiene and convenience. Used shoes often prevent the dissipation of heat and sweat which is produced in large quantities during work and other activities.

High temperatures and excessive humidity in the shoes lead to discomfort of varying intensity. If the adverse conditions of the microclimate in the shoe are present for a long time, pathogenic bacteria and fungi grow. Therefore, materials intended for construction of protective footwear should not only meet the requirements of the protection parameters but also be hygienic so that they can actively support the thermoregulatory processes of the body.

Cambrelle Extreme by DuPont and DRYZ IntelliTemp by Dicon, materials used in the manufacture of socks and footwear, combine very good thermal insulation with the ability to evaporate moisture from the immediate environment of the foot and provide active and lasting protection against microorganisms. In recent years intensive development of multifunctional membrane materials has been observed.

Breathable fabrics, such as the Coolmax lining used for LaCrosse footwear, have changed comfort in footwear today. For warmer climates where breathable materials are needed in footwear, a lining that wicks away moisture will keep feet more dry and comfortable. On the other hand, waterproof material is a necessity in wet conditions. Many manufacturers are turning to fabrics featuring new technology, such as Gore-Tex or Hyper-dri.

Proper protective toe cap is essential element of the footwear responsible for ensuring safety against mechanical risk, i.e., compression and impact. Although protective footwear does not guarantee protection from a foot injury, it can reduce the severity of an injury. Statistics have shown that three out of four people who receive a foot injury while at work did not have any protective footwear. One reason given was that the shoes were uncomfortable. The fact is it is a lot more uncomfortable to have an injured foot than it is to wear a steel-toe boot. Up-to-date toe caps in protective and safety footwear can be made with steel or polymers, for example, polycarbonate. The nonmetallic toe is lighter and more comfortable as well as electrically nonconductive and the resistance to the transmission of heat or cold can make a big difference on the job site. The other solution, using an alloy, is much lighter than a steel toe and just as strong, if not stronger.

The conclusion is that the comfort of use of footwear corresponds to protective parameters and safety. One such parameter is slip resistance. The physical parameter characterized slip resistance is the coefficient of friction (CoF). The higher the CoF, the better the slip resistance. The safety features of footwear, including slip resistance, are tested according to a set of European test standards written into EN ISO 20344:2004 (A1: 2007) (CEN, 2004c). Footwear which has passed the EN test for slip resistance will be coded SRA (tested on ceramic tile wetted with dilute soap solution), SRB (tested on smooth steel with glycerol), or SRC (tested under both conditions).

The sole tread pattern and sole compound are both important for slip resistance. Generally a softer sole and close-packed tread pattern work well with fluid contaminants and indoor environments. A more open pattern works better outdoors or with solid contaminants. Several soles that meet the requirements of the standards are shown in Figure 6.

Another modern footwear is the Shock Protection System (SPS). The combination of SPS with Blundstone's unique dual-density soling has been designed to increase comfort in safety footwear. SPS reduces workplace fatigue and orthopedic problems in the lower back,



**Figure 6**   Examples of up-to-date soles.

legs, and feet. Dynamic tests indicate an average reduction in shock transmitted to the legs of 33% at a brisk walking pace. SPS-Total includes advanced microcellular shock-absorbing material.

## 9   FALL PROTECTION SYSTEMS: SELECTION OF EQUIPMENT

The data concerning accidents at work published annually in many European countries demonstrate that work on heights still belongs to the most hazardous occupations. The practice of work on heights in such sectors as civil engineering, energy engineering, mining, and telecommunications demonstrates that in many cases it is impossible either to eliminate the risk of falls from a height or to use group protections such as barriers and protective nets. In such situations, the use of personal systems protecting against falls from a height is the only method available. To play their role correctly, such systems must be made up of appropriately selected components. The most important factors determining the selection of components for systems that protect against falls from a height include:

- Topography of the work site, including the available space which can be used for fall arrest
- Presence of construction elements which can be used to anchor the protective equipment
- Presence of other work site factors which may affect the technical efficiency of the protective equipment, for example, high temperature, molten metal splashes, aggressive chemicals
- Typical movements of the user at the work site, for example, in the vertical or horizontal direction
- Single-instance or repeated nature of tasks performed at a particular work site
- Necessity to minimize the free-fall distance
- Necessity of work positioning while performing tasks on a height

The first step in selecting systems that protect against falls from a height is type selection. Three types of systems characterized by function are available: systems designed for fall arrest, systems designed for work positioning, and systems designed for restraint of a fall from a height.

The fall arrest systems described by Sulowski (1991) are designed for work sites whose topography and workers activities required make it impossible to eliminate the risk of a fall. The main functions of such a system include fall arrest, alleviation of fall effects by reduction of forces acting on the human body, prevention of injuries caused by crashing against dangerous objects on the ground or at the work site, and maintaining the position of the user's body during and after fall arrest, allowing for assistance to arrive. An example of a fall arrest system is presented in Figure 7.

A fall arrest system consists of three basic components: an anchor component, a connecting and shock-absorbing component, and full-body harnesses. The
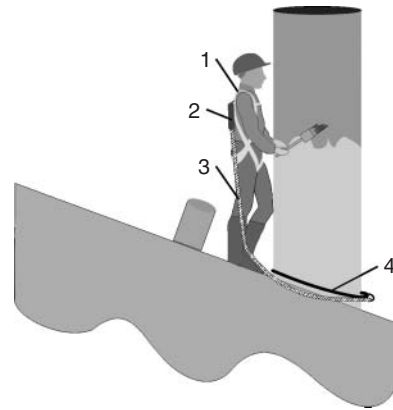


**Figure 7**  Fall arrest system: 1, full-body harnesses; 2, energy absorber; 3, lanyard; 4, anchor lanyard.

anchor component is the first link of the system and is connected directly with the work site, anchoring the connecting and shock-absorbing component to the work site to prevent the user from falling down. The role of universal anchor connectors can be played by various types of connectors attached to safety lanyards, wire grider grips, wire rope slings, anchor slings, and hooks.

Such components can be connected to steel constructions, beams, ties, trusses, and other work site elements of appropriate shape and mechanical strength. Such anchor elements are generally unsuitable for work where the user moves in the horizontal plane because they must be disconnected and then reconnected to the new anchor points. Horizontal flexible anchor lines and horizontal rigid anchor lines presented by Baszczyński and Zrobek (2000) can be used as the anchor components, enabling the worker to move in the horizontal plane.

To protect workers in wells (e.g., the sewage system, shafts), the best solutions for anchoring fall arrest systems include tripods or horizontal anchor beams with an anchor point.

The second component of a system protecting against falls from a height is a connecting and shock-absorbing component located between the anchor component and the full-body harness. Its main functions include fall arrest, reduction to safe values (not exceeding 6 kN) of the forces acting on the user's body during fall arrest, and reduction of falling distance. Owing to such functions, the connecting and shock-absorbing component alleviates the conditions of fall arrest and minimizes the risk of impacts caused by crashes with elements at the work site. The function of this component involves absorption of the kinetic energy of the human body falling down. This energy is converted to deformation and friction forces of the component elements. The most popular connecting and shock-absorbing components used at present include:

- Lanyards and textile energy absorbers, in which the kinetic energy is converted into the work of separation of two layers of shock-absorbing webbing

- Retractable-type fall arresters described by Baszczyński and Zrobek (2003) and Baszczyński (2006), in which the kinetic energy is converted into the work involving friction of brake disks or tearing of a textile element
- Guided-type fall arrester on a rigid or flexible anchor line, in which the kinetic energy is converted into the work involving friction between the anchor line and the self-locking mechanism and deformation of the anchor line or other elements specifically designed for that purpose

The third component of a fall arrest system, which is in direct contact with the user's body, is the full-body harness. The main functions of such equipment include:

- Distribution of the dynamic forces acting during fall arrest on the human body in a manner reducing the risk of injuries
- Appropriate positioning of the user's body during fall arrest to prevent damage to the internal organs and vertebral column
- Appropriate positioning of the user's body after completion of fall arrest to enable the user to wait for help safely as comfortably as possible

The fall arrest attachment element (usually a buckle) is an important element of the full-body harness, whose construction and positioning determine its application. Most frequently, this element is placed at the back and then the body harness can cooperate with energy absorbers with lanyards, retractable-type fall arresters, and guided-type fall arresters on flexible anchor lines. The attachment buckle placed on the chest can also be connected to the above connecting and shock-absorbing elements as well as with guided-type fall arresters on rigid anchor lines.

Work-positioning systems, presented by Baszczyński and Zrobek (2005), are the second type of system protecting against falls from a height. They are designed to position the users so that they are firmly supported and can use both hands for work. A work-positioning belt equipped with side attachment elements connected to a work-positioning lanyard with a length adjuster is an example of such equipment presented in Figure 8.

During exploitation of the equipment user's back is supported by a belt, while the legs rest on the work site construction.

The lanyard, the ends of which are connected to the attachment elements of the work-positioning belt, is tied around a work site construction element, for example, a pole. Its length is adjusted by the user so as to ensure a safe and comfortable position. The work-positioning lanyard must be tied around such an element, which will make its displacement and, consequently, the initiation of the user's fall impossible. If such conditions cannot be provided at the work site, an additional fall arrest system and work-positioning system are necessary. In such cases, full-body harnesses equipped with a work-positioning belt should be used.
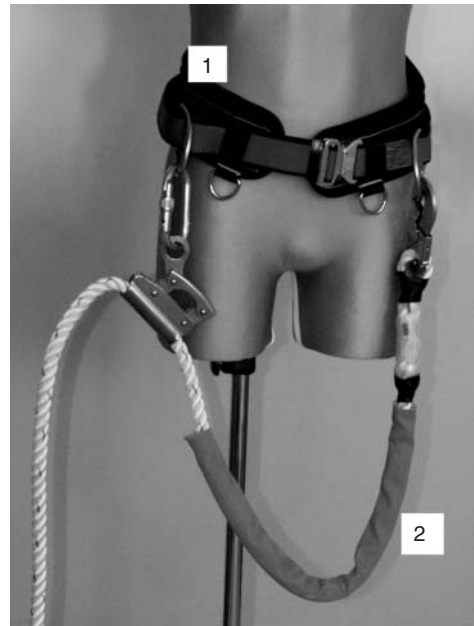


**Figure 8** Work-positioning system: 1, belt for work positioning; 2, work-positioning lanyard.

The third type of system protecting against falls from a height are restraint systems. Their main task is to restrain the users' mobility, keeping them away from the dangerous area associated with the risk of falls. A restraint system consists of an anchor component allowing attachment to work site construction elements, a connecting component with one end attached to the anchor element and the other to the harness (e.g., a lanyard with adjustable length and a guided-type fall arrester on a flexible anchor line equipped with hand-operated blockade), and harnesses, such as full-body harnesses, work-positioning belts, and sit harnesses.

Fall restraint systems are designed primarily for large-space gentle-slope work sites. Such systems can be used only at the work sites where the user does not have to stay in places associated with the risk of a fall.

The selection of components to be used in a system protecting against falls from a height should take into account the factors present at the work site, which can negatively affect the protective parameters of these components. The most important of such factors include thermal factors (e.g., molten metal splashes, open flame, high temperature, aggressive chemical substances), mechanical factors (e.g., sharp and rough objects), and atmospheric factors, including in particular low and high temperatures, humidity, and rainfall/snowfall.

The comfort of use is also an important factor in the selection of elements for personal systems protecting against falls from a height. The equipment should be as lightweight as possible, cause no restraint of movements necessary to carry out the required tasks, exert no pressure causing the sensation of discomfort, and provide a feeling of safety.

## REFERENCES

American National Standards Institute/American Standards Association (ANSI/ASA) (2008), "Methods for Measuring the Real-Ear Attenuation of Hearing Protectors," S12.6-2008, ANSI/ASA.

American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) (2009), *Handbook, Fundamentals*, ASHRAE.

Bartkowiak G. (2000), "Liquid Sorption in Two-Layer Textiles Designed for Use under Hermetic Protective Barrier," *Fibres and Textiles in Eastern Europe*, No. 4, pp. 79–83.

Bartkowiak G. (2006), "Liquid Sorption by Nonwovens Containing Superabsorbent Fibers," *Fibres and Textiles in Eastern Europe*, No. 1, pp. 57–61.

Bartkowiak, G., et al. (2010), *Modelling of Heat Transfer through Barrier Protective Clothing, Innovation in Clothes and Footwear*, Kazimierz Pulaski Technical University of Radom, pp. 258–266.

Baszczyński, K. (2002), "Industrial Safety Helmets as a Protection against Side Impacts," *Bezpieczeństwo Pracy*, No. 5, pp. 10–13.

Baszczyński, K. (2006), "The Influence of Anchor Devices on the Performance of Retractable Type Fall Arresters Protecting against Falls from a Height," *International Journal of Occupational Safety and Ergonomics (JOSE)*, Vol. 12, No 3, pp. 307–318.

Baszczyński, K., and Zrobek, Z. (2000), "Dynamic Performance of Horizontal Flexible Anchor Lines during Fall Arrest—A Numerical Method of Simulation," *International Journal of Occupational Safety and Ergonomics (JOSE)*, Vol. 6, No 4, pp. 521–534.

Baszczyński, K., and Zrobek, Z. (2003), "Test Method for Retractable Type Fall Arresters Designed for Horizontal Use," *International Journal of Occupational Safety and Ergonomics (JOSE)*, Vol. 9, No 3, pp. 313–331.

Baszczyński K., and Zrobek, Z. (2005), "Protection against Falls from Poles—Efficiency Assessment of Work Positioning Equipment," *Bezpieczeństwo Pracy*, No. 4, pp. 18–21.

Berger, E. H. (2000), "Hearing Protection Devices," in *The Noise Manual*, E. H. Berger, L. H. Royster, J. D. Royster, D. P. Driscoll, and M. Layne, Eds., AIHA Press, Fairfax, VA, pp. 379–454.

Berger, E. H., Kieper, R.W., and Gauger, D. (2003), "Hearing Protection: Surpassing the Limits to Attenuation Imposed by the Bone-Conduction Pathways," *Journal of the Acoustical Society of America*, Vol. 114, pp. 1955–1967.

Brochocka, A., and Makowski, K. (2005), "Assessment of Usage Comfort of Respiratory Protective Devices in Practical Performance Tests" *Proceedings of International Conference on Research and Standardization in the Field of Development and Use of Personal Protective Equipment, (CIOP-PIB), Cracow,* September 2005, Cracow, Poland.

European Directive (2003) of the European Parliament and of the Council of 6 February 2003 on the Minimum Health and Safety Requirements Regarding the Exposure of Workers to the Risks Arising from Physical Agents (Noise), 2003/10/EC.

European Committee for Standardization (CEN) (2004a), "Hearing Protectors—Recommendations for Selection, Use, Care and Maintenance—Guidance Document," EN 458:2004, CEN, Brussels, Belgium.

European Committee for Standardization (CEN) (2004b), "Ergonomics—Evaluation Of Thermal Strain By Physiological Measurements," EN ISO 9886: 2004, CEN, Brussels, Belgium.

European Committee for Standardization (CEN) (2004c), "Personal Protective Equipment. Test Methods for Footwear," EN ISO 20344:2004 (A1: 2007), CEN, Brussels, Belgium.

European Committee for Standardization (CEN) (2005), "Protective Clothing against Liquid Chemicals—Performance Requirements for Clothing with Liquid-Tight (Type 3) or Spray-Tight (Type 4) Connections, Including Items Providing Protection to Parts of the Body Only (Types PB [3] and PB [4])," EN 14605:2005, CEN, Brussels, Belgium.

Fanger, P. O. (1970), *Thermal Comfort Analysis and Applications in Environmental Engineering*, McGraw-Hill, New York.

Gagge, A. P., Burton, A. C., and Bazet, H. D. (1971), "A Practical System of Units for Destination of Heat Exchange of Man with His Environment," *Science*, Vol. 94, pp. 428–430.

Holmer, I. (1995), "Protective Clothing and Heat Stress," *Ergonomics*, Vol. 38, No 1, pp. 166–182.

International Organization for Standardization (ISO) (1990), "Acoustics—Hearing Protectors—Part 1: Subjective Method for the Measurement of Sound Attenuation," ISO 4869-1:1990, ISO, Geneva.

International Organization for Standardization (ISO) (1994), "Acoustics—Hearing Protectors—Part 2: Estimation of Effective A-Weighted Sound Pressure Levels When Hearing Protectors Are Worn," ISO 4869-2:1994, ISO, Geneva.

International Organization for Standardization (ISO) (2002), "Acoustics—Determination of Sound Immission from Sound Sources Placed Close to the Ear—Part 1: Technique Using a Microphone in Real Ear (MIRE Technique)," ISO 11904-1:2002, ISO, Geneva.

International Organization for Standardization (ISO) (2004), "Acoustics—Determination of Sound Immission from Sound Sources Placed Close to the Ear—Part 2: Technique Using a Manikin," ISO 11904-2:2004, ISO, Geneva.

International Organization for Standardization (ISO) (2006), "Acoustics—Hearing Protectors—Part 5: Method for Estimation of Noise Reduction Using Fitting by Inexperienced Test Subjects," ISO/TS 4869-5:2006, ISO, Geneva.

International Organization for Standardization (ISO) (2007), "Acoustics—Hearing Protectors—Part 3: Measurement of Insertion Loss of Ear-Muff Type Protectors Using an Acoustic Test Fixture," ISO 4869-3:2007, ISO, Geneva.

Irzmańska, E., et al. (2010), "Some Aspects of the Preparation Position to Determine the Resistance of Protective Gloves to Permetation of Oils Using UV Spectrophotometry Method," *Measurement Automation and Monitoring*, No. 8, pp. 883–886

Korycki, R. (2002), "The Damping of Off-Central Impact for Selected Industrial Safety Helmets Used in Poland," *International Journal of Occupational Safety and Ergonomics (JOSE)*, Vol. 8, No 1, pp. 51–70.

Kubacki, Z., Owczarek, G., and Pościk, A. (2001), "New Generation of Welding Eye and Face Protections," *Occupational Safety*, No. 4, pp. 10–13.

Kurczewska, A., and Leśnikowski, J. (2008), "Variable-Thermoinsulation Garments with a Microprocessor Temperature Controller," *International Journal of Occupational Safety and Ergonomics (JOSE)*, Vol. 14, No. 1, pp. 77–87.

Makowski, K., and Majchrzycka, K. (2004), "Assessment of Ergonomic Aspects of Respiratory Protective Devices," in *Conference Book of The XII Conference of International Society for Respiratory Protection (Yokohama)*, (ISRP), Yokohama Yokohama.

Marszałek, A., and Sawicka, A. (1997), "Impermeable Protective Clothing—Work Time Must Be Limited," in *Fifth Scandinavian Symposium on Protective Clothing—Proceedings*, R. Niesen and C. Carsten Borg, Eds., NOKOBETEF, Copenhagen, Denmark, pp. 100–106.

Marszałek, A., Bartkowiak, G., and Łężak, K. (2009), "Physiological Effects of a Modification of the Construction of Impermeable Protective Clothing," *International Journal of Occupational Safety and Ergonomics (JOSE)*, Vol. 15, No 1, pp. 61–73.

Parsons, K. (2003), *Human Thermal Environment*, Taylor & Francis, London and New York

Pause, B. (2006), "New Cooling Undergarment for Protective Garment System," in *Proceedings of 3rd European Conference on Protective Clothing (ECPC) and NOKOBETEF 8*, Gdynia.

Reinertsen, R. E., et al. (2008), "Optimizing the Performance of Phase-Change Materials in Personal Protective Clothing Systems," *International Journal of Occupational Safety and Ergonomics (JOSE)*, Vol. 14, No. 1, pp. 43–53.

Smolander, J., Louhevaara, V., Tuomi, O., Korhonen, O., and Jaakola, J. (1984), "Cardiorespiratory and Thermal Effects of Wearing Gas Protective Clothing," *International Archives of Environmental Health*, Vol. 54, pp. 261–270.

Stefko, A. (2009), "Cut&Puncture Protective Gloves," *Health and Safety International*, No. 31, pp. 37–51.

Sulowski, A. C. (1991), "Fall Arresting Systems—Classification," in *Fundamentals of Fall Protection*, A. C. Sulowski, Ed., International Society for Fall Protection, Toronto, Canada, pp. 285–301.

White, M. K., and Hodous, T. K. (1987), "Reduced Work Tolerance Associated with Wearing Protective Clothing and Respirators," *American Industrial Hygiene Association Journal*, Vol. 48, pp. 304–310.

Zera, J., and Mlynski, R. (2007), "Attenuation of High-Level Impulses by Earmuffs," *Journal of the Acoustical Society of America*, Vol. 122, pp. 2082–2096.

# CHAPTER 31

# HUMAN SPACE FLIGHT

**Barbara Woolford and Walter E. Sipes**
**NASA Johnson Space Center**
**Houston, Texas**

**Edna R. Fiedler**
**Aerospace Psychology Consultants**
**Tucson, Arizona**

## 1 UNIQUE FACTORS IN SPACE FLIGHT

The first human space flight, in the early 1960s, was aimed primarily at determining whether humans could indeed survive and function in microgravity. Would eating and sleeping be possible? What mental and physical tasks could be performed? Subsequent programs increased the complexity of the tasks the crew performed. Table 1 summarizes the history of U.S. space flight, showing the projects, their dates, crew sizes, and mission durations. With almost 50 years of experience with human space flight, the emphasis now is on how to design space vehicles, habitats, and missions to produce the greatest returns to human knowledge. What are the roles of humans in space flight in low Earth orbit, on the moon, and in exploring Mars?

The National Aeronautics and Space Administration (NASA) has captured the information about physical health and human factors in several standards and handbooks, which form the basis for design of future space missions, vehicles, and habitats. NASA-STD-3001, *Space Flight Human System Standards*, Volume 1, *Crew Health*, was approved in 2009 (NASA, 2009a).

Volume II, *Human Factors, Habitability, and Environmental Health*, was approved in 2011 (NASA, 2011). These documents capture standards and their rationale. Much more human factors information is presented in NASA/SP-2010-3407, *Human Integration Design Handbook* (NASA, 2010c). These are the successors to NASA-STD-3000, *Man-Systems Integration Standards* (NASA, 1995), which captured standards, guidelines, lessons learned, and design concepts in one document.

### 1.1 Gravity

The most obvious factor specific to space flight is gravity. Orbiting Earth, crews experience free-fall, or microgravity. This affects all aspects of life and requires special considerations when designing habitat, equipment, tools, and procedures. During launch and entry, crews experience hypergravity for short periods of time. Extensive research and experience with high-performance aircraft has provided great understanding of these environments, and indeed, the tasks to be performed are similar to aviation tasks. On the surface of the moon and Mars, gravity is substantially lower

**Table 1  U.S.-Crewed Space Programs to Date**

| Program | Dates | U.S. Crew Size | Mission Length |
|---|---|---|---|
| Mercury | 1961−1963 | 1 | Up to 34 h |
| Gemini | 1961−1962 | 2 | Up to 6 days |
| Apollo | 1968−1972 | 3 | Up to 12.5 days |
| Skylab | 1973 | 3 | Up to 84 days |
| Apollo−Soyuz (ASTP) | 1975 | 3 | Up to 9 days |
| Space Transportation System (STS) | 1981−current | 2−10 | 3−17 days |
| Shuttle−Mir | 1995−1998 | 2 Russian, 1 U.S. | Up to 6 months |
| International Space Station (ISS) | 2000−current | 2−6 (including international partners) | approx. 6 months |

than on Earth but is definitely sufficient to allow designing habitats, equipment, and tasks analogously to those on Earth.

### 1.2  Mission Constraints

Accommodations for humans in space are constrained by the three major mission drivers: mass, volume, and power, each of which drives the cost of a mission. Mass and volume determine the size of the launch vehicle directly; they limit consumables such as air, water, and propellant; and they affect crew size and the types of activities the crew performs. Power is a limiting factor for a space vehicle. All environmental features—atmosphere, temperature, lighting—require power to be maintained. Power can be generated from batteries, fuel cells, or solar panels. Each of these sources requires lifting mass and volume from Earth, driving mission cost.

### 1.3  Mission Duration

The habitability and human factors requirements for space flight are driven by mission duration. The Space Transportation System (STS) was designed for missions on the order of 2 weeks—analogous to a camping trip. With Mir and the International Space Station (ISS), mission durations of 6 months became standard, requiring far more concern for habitability and for crew efficiency, training, and sustenance. As NASA begins to plan for a mission to the Mars surface, with travel times on the order of 6 months each way and a possible surface stay of 18 months, it must address providing all support and services to crew members: health maintenance, training, recreation, food, clothing, and so on.

### 1.4  Communications

To date, the model for space exploration has had a very small crew—a maximum of seven or eight on a shuttle flight and nominally six people on the ISS—supported by a very large group of scientific and engineering experts on the ground. The crew and ground personnel are linked through the mission control center (MCC). This model has been essential because such a small crew cannot be expert in all the critical subsystems on board. There are too few people to understand the subsystems in sufficient detail to operate and maintain them under nominal circumstances, let alone when malfunctions occur. But this model depends on rapid two-way communications. Video and audio transmissions allow the MCC to see and hear the crew and to transmit questions and procedures in a short enough time to be responsive to time-critical events. Even between Earth and the lunar surface, communications lags are on the order of seconds. But with a mission to Mars, communications can take up to 20 min each way, and a "black-out" period of up to two weeks may occur when the sun is between Earth and Mars. This requires the roles of the ground and flight crews to be reexamined.

### 1.5  Crew Time

Crew time is becoming recognized as another mission driver. The size of the crew affects mass and volume requirements directly. Designing equipment and procedures to maximize returns from crew time is beginning to be considered in the earliest stages of mission planning. Detailed studies of how crew time was actually used during Skylab (Bond, 1977) showed that approximately one-third of the crew time was spent in sleep and one-third in other forms of self-sustenance such as hygiene, exercise, eating, and recreation and one-third was actually devoted to operating the spacecraft and scientific experiments. This has not changed very much on the ISS.

## 2  ANTHROPOMETRY AND BIOMECHANICS

### 2.1  Changes in Posture and Body Size

In a microgravity environment the body changes. Immediately on reaching free-fall, the body assumes a neutral posture quite different from standing or sitting postures on Earth. The neck, shoulders, elbows, hips, and knees all flex somewhat, and the shoulders also abduct and rotate with a large intersubject variability. The result affects a crew member's line of sight, height, and reach envelope (Mount et al., 2003). The range of postures observed on one Shuttle mission is shown in Figure 1. Table 2 gives the joint angles. Figure 2 illustrates reach envelopes based on a typical posture for a 95th percentile crew member. After a short while, on the order of hours, the body height changes due to spinal elongation. Standing height increases about 3% during the first day or so in microgravity. A current study (Young et al., 2010) is measuring changes in seated height, which preliminary results indicate increases about 6%, affecting seat placement and suit design. The
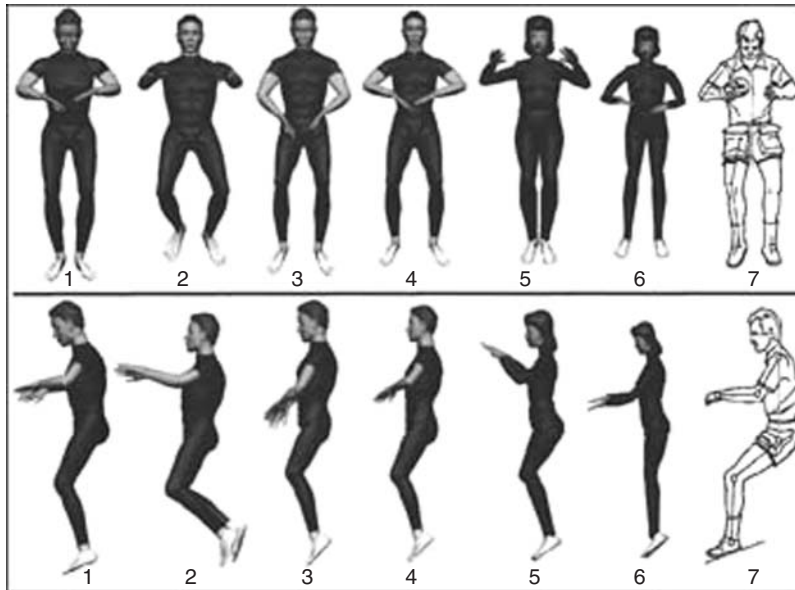
**Figure 1**   Neutral postures in microgravity. Bodies 1–6 are actual crew members. Body 7 is a composite posture based on Skylab data.

**Table 2   Crew Microgravity Posture Measurements (deg)[a]**

| Anthropometric Measurement Joint Angles[b] | Skylab Composite Left–Right | Crew 1 Left–Right | Crew 2 Left–Right | Crew 3 Left–Right | Crew 4 Left–Right | Crew 5 Left–Right | Crew 6 Left–Right |
|---|---|---|---|---|---|---|---|
| Hip flexion | 50 | 33 | 33–29 | 33 | 33 | 29 | 12 |
| Hip abduction | 18.5 | 6.5–5.5 | 20–16 | 13–17.5 | 15.5–16 | 3.5–4.5 | 4–9 |
| Knee flexion | 50 | 50 | 83–87 | 50 | 50 | 44 | 11–12 |
| Ankle plantar extension | 21 | 6–7 | 15–14.5 | 29–30 | 27–24 | 16–14 | 35–41 |
| Waist flexion | 0 | 13 | 0 | 1 | 0 | 0 | 2 |
| Neck flexion | 24 | 16 | 18 | 16 | 5 | 7 | 16 |
| Left neck lateral bend | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| Shoulder flexion | 36 | 49–46 | 67–64 | 29 | 33–35 | 60–57 | 36 |
| Shoulder abduction | 50 | 32–33 | 26–26.5 | 27–29 | 40.5 | 24–45 | 23–36 |
| Medial shoulder rotation | 86.6 | 58–61 | 45.5–41 | 71–77 | 74.5–74 | 25.5–26.5 | 50–48 |
| Elbow flexion | 90 | 78 | 45–53 | 61–57 | 94–91 | 78–80 | 51–64 |
| Wrist extension | 0 | 0 | 3–0 | 0 | 0 | 0 | 0 |
| Wrist ulnar bend | 0 | 0 | 0 | 0 | 0–9 | 0–3 | 0 |
| Forearm pronation | N/A | N/A | 26 | 20–N/A | N/A–2 | 16–N/A | N/A–5 |
| Forearm supination | 30 | 7–10 | N/A | N/A–30 | 15–N/A | N/A–4 | 14–N/A |
| Finger flexion | 0 | 42 | 60 | 30 | 21–57 | 55–47 | 25–35 |

[a]Crews 1–6 correspond to the body positions shown in Figure 1. Skylab composite corresponds to illustration 7.
[b]Angles are based on an upright stature coordinate system.

distribution of body fluids also changes. Fluids move to the head and torso, affecting hand size, facial appearance, the voice, and perhaps the sense of smell.

## 2.2   Changes in Strength

Changes in strength over time in microgravity have been a focus of research because of the direct effect on the ability to perform physical tasks. Jaweed (1994) reports significant (10–20%) decreases between preflight and postflight strength in the antigravity muscles (back and legs) after as few as 5–10 days on orbit. This, taken with the loss of bone mass observed (Schneider et al., 1994), indicates that countermeasures must be taken for long-duration flights and that tasks that can be performed
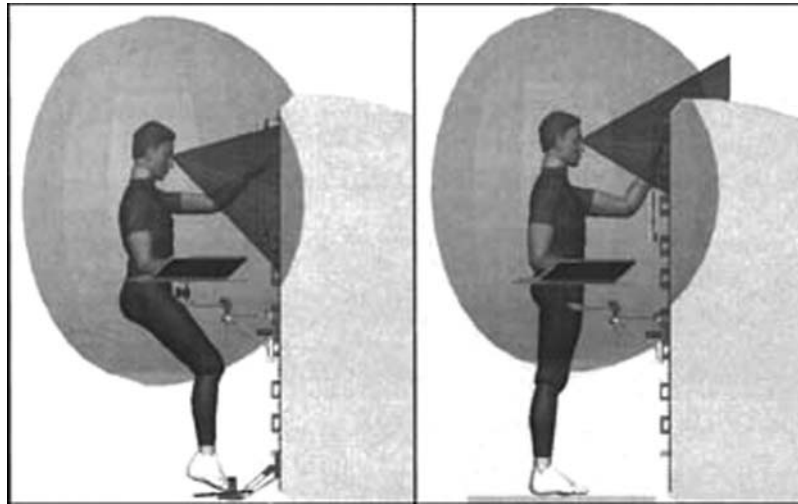
**Figure 2**  A 95th percentile stature crew member is shown in a neutral body posture (left) and standing vertically (right). The elliptical gray shading indicates the reach envelope. The darker gray cone indicates the viewing area.

early in flight might be more difficult or dangerous after an extended time in microgravity. The most common countermeasure for strength loss is exercise, particularly of the legs and back. Typical equipment includes bicycle ergometers and treadmills. When designing spacecraft, volume must be allowed for equipment storage and deployment. Significant periods of crew time, on the order of an hour per day per person, must be reserved for exercise. Design and location of equipment must address isolation of vibration and noise.

## 3  ENVIRONMENTAL FACTORS

### 3.1  Human Factors in a Closed Environment

NASA strives to close the spacecraft environment in the sense that every effort is made to recycle air and water rather than to carry replacement oxygen and water on a mission. This greatly affects design of the habitat and equipment. Materials must not release compounds that are difficult to remove from the atmosphere; this eliminates a variety of plastics and certain types of finishes for other materials. Materials must be compatible with cleaning materials and biocides that are safe for the environment; they must be incompatible with flourishing colonies of bacteria and mold.

### 3.2  Atmosphere

Crew members in a system must be provided with an environment to enable them to survive and function as a system component in space. An artificial atmosphere of suitable composition and pressure is the most immediate need. It supplies the oxygen and the pressure their bodies require. Humans can survive in a wide range of atmospheric compositions and pressures. Atmospheres deemed sufficient for human survival are constrained by the following considerations:

- There must be sufficient total pressure to prevent the vaporization of body fluids.
- There must be free oxygen at sufficient partial pressure for adequate respiration.
- Oxygen partial pressure must not be so great as to induce oxygen toxicity.
- For a long duration (in excess of two weeks), some physiologically inert gas must be provided to prevent atelactasis.
- All other atmospheric constituents must be physiologically inert or of low enough concentration to preclude toxic effects.
- The breathing atmosphere composition should have minimal flame or explosive hazard.

Mission planning must take the foregoing considerations for atmospheric conditions and balance them with the constraints of the mission: length of mission, mission objectives, requirement for prebreathe (for extravehicular activity), research requirements for the mission, and equipment in the vehicle. Carbon dioxide levels become increased with visiting shuttle missions. Because of this increase with an additional seven crew members from the shuttle, the ISS $CO_2$ scrubbers must be adjusted by ground control during visits. In the past high $CO_2$ levels have caused headaches with the crew members. This particular fix helped with the STS-127 shuttle mission (Harwood, 2009).

### 3.3  Water

In addition to the obvious need for drinking water, water is required for a variety of other uses, including personal use, hygiene, and housekeeping. If plants are to be grown during the mission, that is an additional water requirement. Typical water requirements for drinking, hygiene, and washing for each crew member are

2.84–5.16 kg per person per day for standard operational mode (NASA, 1995). A crew depends on water that is clean and safe. The use of water that is reclaimed and stored depends on its quality.

Water management systems changed with the design of the space vehicles and life support requirements of each program. During early Mercury, Gemini, and Skylab missions, water was filled up in tanks, built into the vehicle before launch, and carried into space. However, during the Apollo missions, the water source came from the fuel cells; fuel cells convert hydrogen and oxygen to generate power with water as the by-product. This marked a major breakthrough in the water management technology because water tanks did not have to be prefilled before the launch. The shuttle orbiter uses four 168-lb-capacity steel tanks. The potable water source comes from the fuel cell by-product, water.

The ISS has a water recycling system that reclaims wastewaters from the shuttle's fuel cells, from urine, and from oral hygiene and hand washing and by condensing humidity from the air. This eliminates the need to resupply 40,000 pounds per year of water for the life of the station (NASA, 2010b).

### 3.4 Noise

Noise can affect human physiology and health in a number of ways (Wheelwright et al., 1994). From the perspective of human factors, noise can affect performance by interfering with communications, interfering with sleep, and causing annoyance. In an assessment of the SpaceHab-1 mission (STS-57), Mount et al. (1994) found that, although the measured noise levels did not generally exceed the levels permitted for the shuttle flight deck or middeck, noise levels were substantially above design limits for the SpaceHab. This is probably because of the number and nature of experiments and equipment that were located there. However, most crew members required earplugs during sleep, even though they slept in the shuttle. Crew members principally used the intercom rather than unaided voice to communicate, even when in the same area, and reported difficulty in concentration and noise-induced headaches and fatigue.

Large space vehicles present a significant acoustics challenge because of obvious difficulties with controlling a number of connected, operating modules with payloads and equipment to perform vehicle functions and experiments, sustaining crew, and keeping them in good physical condition. Modules have equipment such as fans, pumps, compressors, avionics, and other noise-producing hardware or systems to serve their functional and life support needs. Payload racks with operating equipment create continuous or intermittent noises or a combination of both. Payload rack contributions to the total on-orbit noise can be and has been shown to be significant. The crew exercises on a treadmill and with other conditioning devices that generate noise. Communications between crew and ground, which are raised to communicate over the background environment, adds to the overall crew noise exposure. The crew members have to work and live in the resulting acoustic environment. The acoustics challenge is further complicated by the fact that there are numerous suppliers of modules, hardware, and payloads from across and outside the United States (Allen and Goodman, 2003).

The ISS is a complicated and sophisticated machine. ISS hardware is divided into categories, including the module (or spacecraft), government-furnished equipment, and payloads (science experiments). These different categories of hardware are governed by different requirements. Acoustic noise emissions verification is performed through actual test measurements of the hardware to the greatest extent possible. However, in some instances a fully integrated end item is not available due to schedule mismatches or physical limitations to the hardware configuration, or the payload may be delivered to ISS and placed in a rack already onboard. An acoustic test-correlated analytical model is used to predict overall noise levels in this case so that crew safety can be ensured. Remedial actions are performed to quiet hardware when necessary (Allen and Goodman, 2003).

The astronauts of the ISS are exposed to an average noise level of 72 dBA for the entire duration of their stay on the ISS, which can last up to six months. The significant noise sources throughout the ISS are the life support system ventilation fans and include the carbon dioxide removal systems (+70 dBA), the refrigerators (70 dBA), and air conditioning and ventilation fans (69–52 dBA). Another source of noise is the treadmill and vibration isolation system with an intermittent noise level of 77 dBA during ground assessment. Countermeasures against spacecraft noise include design engineering controls (like quiet fans and use of advanced composite materials), sound insulation materials, and hearing protection (ear inserts, passive muff headsets, and active noise reduction earpieces or headsets) (Clark and Allen, 2008).

### 3.5 Lighting

Lighting is essential to performing virtually every task in space. When windows are present and unshuttered, the typical 90-min low Earth orbit of the shuttle or station causes problems with time for eyes to adapt to the rapid disappearance of sunlight. In the study by Mount et al. (1994), the most frequent report of lighting problems was that sunlight made electronic displays and video monitors difficult or impossible to read. However, some activities, such as remote manipulator operations, require out-the-window viewing, and Earth watching is a favorite crew activity in any spare time. Wheelwright et al. (1994) and the *Human Integration Design Handbook* (NASA, 2010c) provide tables and guidelines for illumination levels for various intravehicular and extravehicular tasks.

Two critical tasks requiring vision of external targets are docking the shuttle to the ISS and using remote manipulators to position space-suited crew members or large structural components. In low Earth orbit, there is a change from light to dark every 90 min. In vacuum, shadows are much sharper than in an atmosphere, where water vapor, dust particles, and other airborne particles scatter light. To ensure adequate light, tasks may be scheduled to be performed in those parts of the orbit when the combination of sunlight and artificial light are predicted to provide adequate contrast

and visibility. NASA developed software that models realistic images of complex environments. Measured data are used to develop models of shuttle and station artificial light. Natural lighting, such as sun and Earth shine, are also incorporated into the lighting analyses. By incorporating the measured reflectance of each material into the lighting models, an accurate calculation of the amount of light entering a camera can be made. Using this calculated light distribution with the model of the shuttle cameras, camera images can be simulated accurately. Use of these lighting images are essential to predict available lighting during space operations requiring camera viewing, such as the assembly of ISS components. In preparing for a shuttle visit to the ISS, mission planners simulate the lighting environment for critical tasks at 1-min intervals.

### 3.6 Dust and Debris

Debris and dust in the orbiter crew compartment of early shuttle missions created crew health concerns and physiological discomfort and were the cause of some equipment malfunctions. Debris from orbiters during flight and processing was analyzed, quantified, and evaluated to determine its source. Selected ground support equipment and some orbiter hardware were redesigned to preclude or reduce particularization/debris generation. New filters and access ports for cleaning were developed and added to most air-cooled avionics boxes. Most steps to reduce debris were completed before flight STS-26, in 1988. After these improvements were made, there was improved crew compartment habitability and less potential for equipment malfunction (Goodman, 1992).

For future lunar/Mars exploration missions, the problem of dust in these environments is recognized. However, our knowledge at this time is limited as to the specifics of the dust. We have some data from previous lunar missions and are supplementing it with derived data. Derived data from our limited but growing knowledge of Mars is forming a basis of our need for requirements for dust abatement. The dust will cause a serious problem for extravehicular activity (EVA) suits and equipment used external to the vehicle. There is also a concern for dust in the vehicle habitation area. Dust inside the vehicle could increase crew time due to more frequent filter changes and other chores to remove dust from equipment. Basic habitability could also be affected if the dust were to accumulate on display screens and cooking equipment.

## 4  HABITABILITY AND ARCHITECTURE

### 4.1  Architecture

Habitability as a discipline is concerned with providing a space vehicle that within some understandably necessary size restraints provides a comfortable, functionally efficient habitat that will support mixed crews living and working together for the duration of the mission. Attention must be given to the morale, comfort, and health of crews with differing backgrounds, cultures, and physical size. Architectural design of crew interfacing elements should be comfortable for the extremes

of any crew population. The *habitability architecture* design concerns are mainly the fixed architectural elements such as (1) the geometric arrangements of compartments, (2) passageways and traffic paths, (3) windows, (4) color, (5) workstations, (6) off-duty areas, (7) stowage, and (8) lighting (NASA, 1983).

Habitable volume is defined as free, pressurized volume, excluding the space required for equipment, fixtures, furniture, and so on. It does not include "nooks and crannies" (i.e., spaces too small for human access). Total volume requirements depend on the specific program goals of the particular mission. Volume requirements for specific workstations have to be calculated after determination of the tasks required at the workstation and number of crew involved (NASA, 1983).

### 4.1.1  Compartments

The success of an extended mission on a space vehicle depends on the crew being an integral part of the interior design. The focus of any vehicle design should be crew centered. The arrangement and design of any habitable compartment should take into account the possibility of a subsystem failure or damage that could require quick, efficient evacuation. The actual vehicle arrangement depends on the specific program's goals and definition. Based on space flight history, configuration should take into account the following:

- Sleeping and private areas should be separate from traffic paths and noise generators.
- Areas that are to be used by more than one crew member at a time should be arranged to avoid bottlenecks. These are areas such as the galley, workstations, and waste management systems.
- Traffic flow analysis should be done for crew tasks and activities.
- Switches should be located in proximity of associated equipment.
- Adequate electrical outlets should be provided to reduce the use of extension power cords and the resulting "spaghetti all over."
- A dedicated desk/work area should be provided for general paperwork associated with vehicle keeping.

Skylab experience has shown that crew members were able to operate equipment easily from any orientation. Basically, a crew member established a local orientation based on himself or herself and proceeded without difficulty. However, it was also shown that crew could much more easily orient themselves in a room with equipment oriented with consistent up and down directions. An inconsistent zero-*g* orientation of one module caused orientation problems that were time consuming. The conclusion is that a common plane for visual reference should be designated throughout each module.

### 4.1.2  Passageways and Traffic Paths

A *passageway* is defined as a pass-through area between two nonadjacent compartments. Passageways shall be kept free of sharp and protruding objects. Skylab crew

members liked the large "ship-type" doorways. They found round hatches to be much less satisfactory.

Traffic paths consist of three types. *Emergency paths* are those used for crew passage to emergency equipment such as oxygen bottle or mask, firefighting equipment, pressure controls, and escape hatches. *Primary paths* are those used for personnel and equipment transfer between major habitable compartments or between a compartment and a workstation or off-duty area. *Secondary paths* provide access behind equipment, between equipment and structural members, and around workstations. All traffic paths can be superimposed to form a total traffic pattern, which in conjunction with detailed task analysis can be used to determine the most efficient placement of mobility aids. This traffic pattern and task analysis must also be used to design out potential bottlenecks in a space vehicle.

To be avoided are the bottlenecks experienced on Skylab missions. They were insufficient passage room in areas with workstations, too much activity in one place (e.g., conflicting placement of shower and tool kit), and the inability to use the waste management equipment if there was someone using the hand-washing equipment (NASA, 1983).

### 4.1.3  Windows

All habitable volumes should include windows that are adequate for terrestrial and celestial references. Windows are necessary for observation of scientific phenomena, monitoring of EVA, observation of the vehicle exterior, photography, and general viewing. Sufficient window locations should always be provided to view Earth for both Earth observation experiments and crew recreation and well-being.

All viewing windows and the area adjacent to them should be considered a crew workstation. Sufficient workspace and restraint equipment should be provided at view ports for one or more crew members to perform assigned tasks. A window should be installed in the pressure hatch that allows the flight crew to observe the EVA crew in the airlock. Windows that are to be utilized for special photography and scientific experiments must be designed with an aperture size that is compatible with the equipment and tasks specified for that location. Space flights have shown window gazing to be the prime off-duty activity for crew members. Window viewing has been a treasured pastime on all missions to date. Astronauts use photography as a way to connect to Earth events and a challenging pastime for some crew members. Between 2001 and 2005, astronauts took 144,180 images of Earth, 84.5% of which were self-initiated (Robinson et al., 2006, 2011)).

Design of windows should provide handholds and equipment restraints. Failure to do so for a window (Figure 3) in the science module of the ISS led to crew members using a flexible air hose as a handhold. After numerous uses, a hole popped open, causing a slow air leak that took weeks to detect and repair (Banke, 2004).

The design of viewing windows should not impose difficult housekeeping tasks on the crew. Cleaning equipment should be provided for removal of fingerprints and other stains that may accumulate. The equipment must be compatible with the coating(s) on the window and not scratch or affect the optical quality of the window or disturb any surface coating.

Each window should have a sufficiently clear area around it to permit a variety of body positions for viewing. A positive means of defogging the windows should be provided. All window covers and/or shutters should be operated by a device that is easy for any crew member to use. All viewing windows should be provided with a crew-operated, opaque sunshade located



**Figure 3**  An astronaut carries out Earth observation activities through the window in the ISS science laboratory.

within the interior of the spacecraft that is capable of restricting all sunlight from entering the habitable compartments (NASA, 1983).

### 4.1.4  Color

Color should be used to provide visual stimulation for the vehicle occupants and to create different moods for relieving the monotony of prolonged confinement. Factors required in color planning are room volume, function, architecture materials, safety, and required color coding. As the Skylab mission grew in length, the interior color scheme became less acceptable. The crew of the 84-day mission felt that the color scheme was too drab and suggested that accent colors should be used more extensively. Color coding should be used as a supplement to nomenclature to enhance discrimination and to assist the crew in rapid identification of functions. Coding of EVA equipment should be used with colors that will not deteriorate from solar exposure. All EVA handrails should be a standard color. The color should have a high contrast ratio with the background (NASA, 1983).

### 4.1.5  Workstations

A *workstation* is defined as any location in the space vehicle where a dedicated task or activity is performed exclusive of the recreation, personal maintenance, and sleep areas. Tasks and activities include vehicle stabilization and control, systems management, experiments, science, and maintenance (equipment repair). With any workstation, analysis should be done to determine the tasks, operator activities, tools, and equipment necessary for each workstation. To make efficient use of space, multiuse workstation can be considered.

All necessary equipment, tools, restraints, lights, and power outlets should be provided at each workstation. Adequate space should be provided for the crew to perform the assigned tasks efficiently and safely. Where possible, workstations and associated equipment should be standardized throughout the entire vehicle to aid in the efficiency of tasks. Part of the workstation analysis should cover adjacent workstations and any impact that might arise from two crew members working at adjacent workstations at the same time. An analysis of traffic flow should be completed to determine placement of a workstation without bottlenecks.

Flight experience has shown that anything "usable" will be used as a kickoff point or as a grabbing point to change direction of travel. All workstations should be planned to limit inadvertent control activation and/or deactivation by passing crew members. A restraint system should be incorporated into a workstation design with compatibility to the task to be done (NASA, 1983).

### 4.1.6  Off-Duty Areas

There should be a dedicated area for off-duty activities, with a minimum space for the entire crew. This allows for socialization. Stowage areas should be provided in a dedicated recreation area and in the personal space area for items to be used during recreation activity and off-duty time (NASA, 1983). There has been agreement

from crew members on U.S. missions and also from crew during analog studies that they do not like to have the same table used for dining as well as a maintenance bench and/or as a biology work area (Mount, 2002). The psychological need for a separate wardroom/dining table was realized by the ISS Expedition 1 crewmembers so they built a table from scratch using sheet metal and vice grips rather than waiting the delivery of a table in a future Progress mission (Jones, 2004; NASA, 2000).

### 4.1.7  Stowage

Stowage space must be provided. For efficient use the space should be near the stations where the stowed items will be used. A method should be provided for locating stowed equipment and supplies. This is extremely important for a mission like the ISS, where crews are changed out periodically, but large quantities of the stowed equipment and supplies stay (NASA, 1983). Substantial time is spent moving and unpacking stowed supplies and equipment; containers floating in the translation pathways as shown in Figure 4 may provide a hazard if an emergency were to occur.

### 4.1.8  Ambient Lighting

For the most part, lighting follows the same requirements as for an Earth structure, but spacecraft hardware designers face a few human factors challenges not usually encountered in earthbound environments. In general, design of any space vehicle must take into account the constraints of power and weight limitations. This has an impact on the number of lights and their specifications. General lighting for all vehicles designed and built in the U.S. space program have been fluorescent luminaires. LEDs (light emitting diodes) are being considered due to reduced mass and power required for a given amount of light. Fluorescent lighting has to be sealed to contain the mercury in case of breakage. The use of fixed luminaires for general illumination within the relatively small habitable volume of a spacecraft implies that an astronaut may frequently find one or more of these light sources in her or his field of view as she or he floats in microgravity. This creates potential direct glare sources. Additionally, many astronauts are old enough to have experienced typical symptoms of presbyopia. The loss of the full range of accommodation in their viewing close and distant objects is often simply compensated for by their use of corrective eyeglasses or contact lenses. These means are not available to an astronaut during EVAs in a spacesuit, however. The dry, low-pressure, high-oxygen content environment within the spacesuit precludes the use of contact lenses, and the helmet does not provide adequate interior space for eyeglasses. If the helmet were roomy enough to allow eyeglasses to be worn, it is likely that internal light reflections between the lenses of the eyeglasses and the interior of the faceplate would prove problematic. This means that when planning an EVA task, lack of eyeglasses and light levels must be taken into account.

While in low Earth orbit there is a change from light to dark every 90 min. This affects the EVA task planning, due to the changes in light and shadows. The

**Figure 4** Astronaut Alvin Drew moves a stowage container through the laboratory, while other storage bags are temporarily tethered to walls.

Graphics Research and Analysis Facility at Johnson Space Center uses an accurate lighting model to produce realistic images of this complex, ever-changing environment. Measured data are used to develop models of shuttle and station artificial lights along with the natural lighting from sun and Earth shine. This information is incorporated into the task analysis for EVA tasks.

### 4.2 Considerations for Self-Sustenance

The spacecraft must be designed to provide for all aspects of life. For long-duration missions, private compartments are used for sleep and certain personal activities, such as recreational reading or communicating with family and friends. Since the sleep compartment is the single location in which the crew member spends the most time (presleep and sleep), it has been found to be most effective to shield the compartment heavily against radiation.

#### 4.2.1 Sleep

An individual sleep compartment should be provided for each crew member. The private sleeping accommodations should have a privacy curtain, partitions, and stowage lockers. Each sleep area should be located as far as possible from noise, activity, and public area. Since there is no up or down in weightlessness, the position of the body does not matter during sleep (Figure 5). Some astronauts have been bothered by an effect known as *head nod*. If the head is not secure when fully relaxed during sleep, the head develops a nodding motion. Astronauts can secure the sleep restraint (sleeping bag) to limit this nod. Skylab sleep restraints were similar to sleeping bags with neck holes and arm slits. Straps were on the front and back so the crew member could be tightened for a steady, snug position. The space shuttle missions sometimes split the crew into two shifts

to enable around-the-clock science. The ISS is now equipped with six individual sleep stations that have their own lighting, ventilation, and soundproofing (Figure 6). The crew members enjoy the privacy and quiet sleeping environment the sleep stations afford. Many crew members decorate the inside of the sleep station with their family pictures and personal items.

#### 4.2.2 Food

The problem of ensuring astronauts consume sufficient calories and nutrients to maintain health and performance over missions of increasing duration has been a challenge to human systems integration (Perchonok and Bourland, 2002). Since the first food was consumed in orbit in 1962, improvements and developments have been made and are continuing to be made in the food systems for manned space flight. The food system for the Mercury flights was limited in scope and purpose. Food was used in most cases to obtain general information on the effects of null gravity on food ingestion and digestion and to determine types of food and packaging for longer duration space flights. Food for Mercury flights consisted of purees in aluminum tubes, coated tubes, and rehydratables.

The Gemini food system began with an all-dehydrated food system The food consisted of bite-size cubes with an expanded variety and rehydratable foods which included beverages, pudding, soups, fruits, and vegetables. The initial Apollo food system was based on the dehydrated system used for Gemini; however, greater attention was focused on astronaut preference. The availability of hot water increased the selection of foods and enhanced the palatability. The thermostabilized food in a flexible pouch, fresh bread, canned fruit and puddings, and frozen sandwiches for launch day were some of the items introduced on Apollo. Results

**Figure 5** Two examples of using the shuttle sleep restraint while sleeping in the shuttle middeck.



**Figure 6** Astronaut Peggy Whitson straddles the doorway of the temporary sleep station on the ISS.

from Apollo proved that food could be consumed from an open container using normal utensils in microgravity.

A completely new food system was designed for the Skylab program. The new system was required because (1) the food was launched with the orbiting laboratory and would be exposed to unusual environmental extremes and long-term on-orbit storage; (2) the metabolic studies on board required precise intakes of several nutrients; (3) all water had to be launched, so rehydratables offered no weight advantage; and (4) refrigerators, freezers, and food warmers would be available. To meet the long-shelf-life requirement, all Skylab foods were packaged in full-panel pull-out aluminum cans. Cabin pressure required that the aluminum cans be "overcanned" in canisters to withstand pressure variances. This resulted in the rehydratables being packaged in three containers: a plastic pouch, a can, and a canister. Beverages were packaged in a polyethylene collapsible container which expanded on reconstitution. Menus for Skylab were repeated every six days.

The design of the shuttle package significantly reduced the production process and eliminated numerous failure points. Most of the package production steps are automated or semiautomated. At the present time, research is ongoing to look into advanced technology for future food systems for lunar and/or Mars long-term missions.

The ISS menu composition is an extension of the menu system established for the space shuttle/Mir Phase 1 program, which consisted of 50% Russian and 50% American foods (Kloeris and Bourland, 2003).

The next possible step after ISS is long-duration manned space flights beyond low Earth orbit. The duration of these missions may be as long as 2.5 years and will probably include a stay on a lunar or planetary surface. The primary goal of the food system in these long-duration exploratory missions is to provide the crew with a palatable, nutritious, and safe food system and minimize volume, mass, and waste. The paramount importance of the food system in a long-duration manned exploration mission should not be underestimated. During long-duration space missions, several physiological effects may occur, including weight loss, fluid shifts, dehydration, constipation, electrolyte imbalance, calcium loss, potassium loss, decreased red blood cell mass, and space motion sickness. The

menu will provide the crew with changes in the nutrient levels that may be required due to the longer duration mission.

The acceptability of the food system is of much greater importance due to the longer mission durations and the partial energy intake often observed in space flight. The decreased energy intake might significantly compromise the survival of the crew.

### 4.2.3  Personal Hygiene

Managing personal waste and cleaning the skin and hair are problematic because of the lack of gravity and the cost of lifting water to orbit. Except for Skylab, dedicated volumes for various activities have been very limited. Early bodily waste management systems can be described succinctly as "baggies." Since Skylab, there have been a variety of suction-based toilets for collecting fecal matter and urine. The principal systems for personal hygiene for each major spacecraft are described below.

*Skylab*  Personal hygiene for the Skylab crew members was supported in the waste management compartment (WMC). The WMC included a fecal–urine collector, a hand washer, stowage for personal hygiene items and kits, and a drying station. There was also a shower aboard the Skylab. Pressurized water flow combined with a suction device to collect the water caused the water to flow "down." It was considered a pleasant experience but was very time consuming, about 45 min from start to finish. This included cleanup activity.

*Mir*  The Mir personal hygiene subsystem consisted of toilets for body waste management, hand washing units, a shower, and personal hygiene kits. For the last two years the shower was on board, and it was used as an air shower (sauna). It was removed to make way for other required equipment.

*Shuttle*  For washing, the shuttle crew is provided with a personal hygiene system hose located in the waste collection system (WCS) compartment. Water is squirted onto a washcloth using the hose. Some crew prefer to use the hygiene port provided at the galley because it provides hot water. The hose for the galley hygiene port is long enough to be extended to the WCS for cleansing and grooming. The crew is provided with no-rinse body bath and no-rinse shampoo.

*ISS*  The Russian segment is generally the same as for Mir, without a shower. In the U.S. segment the personal hygiene subsystem provides a WMC. Wet wipes and towels are used from the Russian segment. Occasionally, ISS crew members have rigged up a bathing device for their use. There are differing opinions on the results (Mohanty, 2001).

### 4.2.4  Exercise

Exercise regimens prescribed for space missions have required gradually longer and more frequent periods of exercise, particularly as the length of missions has increased. On the first prolonged (18-day) Soviet manned flight, *Soyuz 9*, physical exercises were performed by the cosmonauts for two 1-h periods each

day. In subsequent 24-day flights, 2.5 h of exercise per day was employed, including walking and running on a treadmill. By 1975, the standard program involved three exercise periods per day, with a variety of equipment, for a total of 2.5 h, with the selection of exercises on the fourth day being optional. Over the three missions of the Skylab program a similar increase in exercise quantity was imposed, although the total amounts were less than those used by the Soviets. On the last manned Skylab mission, a treadmill was provided which allowed more vigorous exercise.

Throughout the Skylab missions, successive improvements were seen in postflight leg strength and volume changes, orthostatic tolerance and recovery time, and cardiac output and stroke volume, even though each mission lasted four weeks longer than the last. *Skylab 4* was an 84-day mission. Results of exercise on Soviet missions have shown a similar pattern of reduced physiological deconditioning in response to more strenuous exercise programs (NASA, 1982).

The exercise requirement for ISS is 2.5 h daily with 1.0 h for aerobic exercise (cycle ergometry or treadmill locomotion) and 1.5 h for resistive exercise condition. Each time segment includes 15 min for setup and 15 min for set-down of equipment. Usually astronauts exercise six days a week, with day 7 as active rest (the astronauts can exercise if they want to). They usually start exercise conditioning after space motion sickness has resolved and all transfer of payload has occurred. The Russians do not start exercise countermeasures until flight day 30. The shuttle requirements are different and depend on mission length and crew member roles. They apply only to use of the cycle ergometer.

### 4.2.5  Recreation

With any space vehicle design for a long-term mission, an area for recreation should be designated to provide for social interaction, Earth viewing, games, videotape viewing, music, and active and passive participatory activities. A quiet area should be provided for a crew member to read, listen to music, and write. Currently on the ISS astronauts and cosmonauts listen to their music on computers or personal MP3 players and uplinked streaming video has replaced videotapes and DVD movies. There is an electronic keyboard and guitar on the ISS. And some astronauts have brought their own instruments, like a trumpet and even a didgeridoo. Holiday decorations and materials are located in a specific locker to keep the morale of the crew high on their holidays. Astronauts also have access to a standard load and individual recreational software on their computers. These examples of recreation are continually being upgraded on the ISS.

### 4.3  Vehicle Maintenance

With the exception of Skylab and ISS, in-flight maintenance provisions and planning on U.S. space programs have not been supported by definitive program requirements. The Skylab mission acknowledged a substantive role for maintenance to achieve mission objectives. The wisdom of this decision was validated by the major repair and maintenance tasks required during the brief lifetime of the program. The shuttle program was to have

no in-flight maintenance, with all maintenance tasks planned to be done on the ground. Over the life of the program this has changed, due to the necessity of preventive maintenance, even on the short missions, and unanticipated problems (Mount, 1989).

On-orbit maintenance was recognized as an essential consideration within the ISS program (NASA, 2004). A three-tiered maintenance concept was adopted that is similar to that employed by military organizations. The primary mode of on-orbit maintenance was designated as organizational maintenance and consisted primarily of removal and replacement of orbital replaceable units (ORUs) (comparable to line replaceable units in military applications). This was supplemented by in situ maintenance for systems that did not lend themselves to the modular ORU design approach, such as utility lines and secondary structure. The option was retained for intermediate-level maintenance, which would consist of on-orbit repair of ORUs. Intermediate-level maintenance has been employed to a limited extent in applications such as replacement of circuit cards within avionics ORUs. Crew member training for maintenance has focused on the development of general skills and on types of maintenance tasks. However, extensive training on highly specific actions is done in some specific instances.

Future missions will be challenged by their extended duration, limited or no resupply opportunities once the mission has begun, and extended round-trip communication times (Watson et al., 2003). These factors will require such missions to be almost entirely self-sufficient. An additional constraint will be the need to carefully control and minimize the mass and volume of equipment and supplies used to support maintenance activities. It is expected that maintenance will be performed at the level of piece parts, so that the required replacement parts will be as small as possible. However, performing maintenance at this level carries significant implications from multiple perspectives.

First, hardware must be designed to enable crew members to perform the required maintenance. Not only must the equipment be accessible but also it must be possible for units to be disassembled as necessary to enable piece-part replacement. Additionally, commonality and standardization of piece parts must be imposed to obtain mass and volume benefits. If not, the number of unique piece parts could be so great as to negate any potential benefit. This maintenance concept will also require more extensive diagnostic capabilities than have been used heretofore in space. Every effort should be made to incorporate these capabilities within the systems themselves to minimize the amount of stand-alone test equipment that is required. Preparation of all potential maintenance procedures in advance will probably be prohibitively expensive, so means must be available to provide crew members with necessary information and guidance when needed. An attractive concept would be capable of automatically generating needed procedures based on input from diagnostic systems and from hardware design information stored onboard. Finally, maintenance at this level will require the ability to perform quality assurance tests (Watson, 2003).

Future missions will probably require operations in multiple gravitational environments, including the microgravity environment of Earth orbit or in-space transit, lunar gravity (approximately $0.17g$), and Martian gravity (approximately $0.38g$). Design for maintenance must take these environments into account. For example, a microgravity environment offers three-dimensional freedom of motion, facilitating access to all areas within a spacecraft volume. However, a microgravity environment introduces significant challenges from the standpoint of reacting forces that must typically be applied during maintenance tasks. Fractional-$g$ environments will restrict mobility and access to some degree (e.g., restricting access to hardware in overhead locations) but will facilitate the application of forces by crew members. Another subtle advantage to working in fractional-$g$ environments is that unrestrained parts and tools remain where placed and do not tend to float away and become lost.

With longer missions maintenance must be planned and all contingencies must be anticipated. Simple maintenance tasks take on great complexities when in microgravity. What might be considered a simple task on Earth, such as using a slot-head screwdriver, could be impossible in space. Automation is being developed to save crew time and increase productivity, but we need to know all the ramifications when the automation (and robotics) breaks down (Mount, 1989). As automated capabilities become increasingly prominent in maintenance operations, the potential for their failure and appropriate fallback positions must be considered. Tasks and hardware for which robotic intervention is planned should retain manual intervention as a backup capability. Designs should not preclude manual troubleshooting even if embedded diagnostics are planned. Interchangeability of hardware within and among spacecraft should be a key design objective.

Considerations to be given for support of maintenance in space fall into four categories (Mount, 1989):

1. Crew provisions, which includes interfaces, restraints, physical and visual access, tools and equipment, procedures and references, and personal protective equipment
2. Hardware design, which includes design for maintainability; use of common connectors, fasteners, and mounts; structural interfaces; and replacement parts
3. Software, which includes architecture design for maintainability and reconfigurability, fault detection and recovery, integrated training support, and inventory control and management
4. Supporting disciplines and processes, especially safety, reliability, and quality assurance

### 4.4 Restraints

Launch and reentry require significant structural strength; loads of up to $5g$ are experienced in nominal conditions. But once in orbit, the microgravity environment enables objects to be held in place with very little force; hook and loop fasteners dot the surfaces.

**Figure 7** Operating the remote manipulator system requires a stable restraint carefully adjusted.



**Figure 8** Brief activities can be performed with simpler restraints.

On the other hand, some force must be provided to hold anything in place. Restraints are needed for both personnel and equipment in microgravity. The most common restraint for crew members is a foot restraint. In a location where a person will be working for extended periods of time, platforms can be used that tilt to accommodate a neutral posture, with the feet angled down and with height adjustments.

Tasks of various durations requiring various degrees of force or dexterity require different types of restraints. Short, easy tasks can often be performed with toes stuck under a handle or one hand on a handhold. Tasks such as attaching a module to the ISS using the remote manipulator system, which take many hours and a high degree of hand–eye coordination, require a restraint such as that shown in Figure 7. This restraint provides support for the feet and thighs. Another example of restraints is shown in Figure 8, illustrating use of existing hardware for a temporary restraint.

## 5 SLEEP AND CIRCADIAN RHYTHM

### 5.1 Sleep Shifting and Light

Circadian and sleep components, two physiological processes, interact in a dynamic manner to regulate changes in alertness, performance, and timing of sleep. Light can aid in shifting circadian rhythms to an

earlier or later time within the biological day. Also, use of bright light during nighttime can result in significant improvement in performance and alertness levels (Campbell and Dawson, 1990). Astronauts in space are exposed to variable light levels due to the non-24-h orbital cycle (day/night) of space operations, such as the 90-min orbital cycle of the shuttle. Additionally, light levels in the space environment can be variable. Field data have shown that light levels aboard spacecraft can be as low as 10 lux during the highest activity portions of the day and as high as 79,433 lux on the flight deck (Dijk et al., 2001). The Soviets recommended 400–500 lux of full-spectrum light for work on spacecraft, and results demonstrated an improvement in performance when the location of lights on *Salyut*-7 was changed to maximize lighting (Bluth, 1984).

Barger et al. (2008) reported on their progress in collecting in-flight data about actual sleep patterns using Actiwatches and sleep logs. Preliminary results of this and other sleep studies are available in the report "Risk of Performance Errors Due to Sleep Loss, Circadian Desynchronization, Fatigue and Work Overload" (Whitmire et al., 2010).

NASA currently uses light treatment to help crew members adapt their circadian system prior to missions,

allowing the astronauts to be physiologically alert when critical tasks are required. The timed use of bright light to facilitate circadian phase shifts was effective in the STS-35 mission, the first mission requiring both dual shifts and a night launch. Subjective reports indicated that crew members were able to obtain better quality sleep during the day and remain more alert during the night after using bright-light exposure to facilitate their schedule inversion prior to the launch dates (Czeisler et al., 1999). Czeisler is currently testing the hypothesis that exposure to short-wavelength light will synchronize circadian rhythms to a shifted sleep schedule within four to five days.

Around-the-clock operational tasks for some missions have required splitting crews into two separate shifts, which required that half the crew invert their sleep–wake cycles. A procedure called *slam shifting*, which involves abrupt shifts of up to 12 h, has been used to align the sleep–wake schedules of shuttle and ISS crews upon docking, in conjunction with EVAs, and also with Progress and Soyuz dockings. The ISS crew is normally on Greenwich Mean Time (GMT) and will be shifted either to Moscow time or Houston time for activities under those perspective ground controllers and then returned to GMT after the activity is completed. Staggered sleep schedules on an eight-day mission did not work, since the crew tended to retain ground-based work–rest cycles and the schedules resulted in increased fatigue and irritability. On a one-year flight, where sleep times for docking operations were shifted by 4.5–5.0 h a total of 14 times, asthenia, end-of-day fatigue, and sleep disruptions were documented (Grigor'yev et al., 1990).

Current astronaut crew scheduling guidelines allow for astronauts' schedules to be lengthened by no more than 2 h (phase delay) and shortened by no more than 30 min (phase advance) within a given day (NASA, 1992). Schedules can be lengthened only if there is an operational requirement. For example, if the shuttle is going to dock with ISS during a time that the ISS crew is scheduled to be sleeping, operations would require the ISS crew to shift to a new schedule in the days preceding in order to be awake and alert for the docking (Mallis and DeRoshia, 2003).

### 5.2 Mars Day Circadian Entrainment

With NASA's continuing support of a manned mission to Mars, the effects of a Mars light–dark cycle must be investigated to determine a person's ability to adapt to a Mars cycle and its impact on physiological alertness. The Martian day, otherwise known as a sol, is about 39 min longer than an Earth day (a sol period is 24.6 h). Although this period length is well within the circadian range of entrainment according to previous studies conducted in relatively bright light (23–27 h) (Aschoff and Wever, 1981), preliminary laboratory results have suggested that in dim-light conditions, such as found indoors, humans cannot reliably entrain to a 24.6-h Mars sol. People differ as to their circadian rhythm, and the 25% of the population who have periods shorter than 24 h will have the greatest challenges acclimatizing to a Mars sol (Mallis and DeRoshia, 2003). Another laboratory study by Gronfler et al. (2007) showed that,

while 25 lux did not entrain subjects, 100 lux and a modulated light exposure of 25, 100, and 9500 lux did entrain the subjects to a 25-h day. This light exposure entrainment was confirmed in a study conducted in an operational environment with Phoenix Mars Lander scientists who lived on a Mars sol during their 90-day mission in 2008 (Thompson, 2008).

## 6  PERCEPTION AND COGNITION

One driver of human spaceflight is the perceptual–cognitive abilities of the crew. Perception includes both the sensing and interpreting of stimuli. Cognitive capabilities range from attention to spatial skills to executive decision making. Consideration must be given to how spacecraft architecture and design can help or hinder perceptual and cognitive capabilities across all task demands and crew capability (NASA, 2009b).

The major difference between optimal perceptual and cognitive functioning on Earth and long-duration space flight is the challenge of microgravity. Microgravity produces fluid shifts which in turn change otolith regulation and vestibulo-ocular reflexes and affect the congruence of the vestibular system to other receptors. Eye–hand coordination and gaze transitions can be temporary perceptual problems while adapting to microgravity (Paloski et al., 2008). During this time of adjustment, postural cues are disorienting and body movement may be awkward if not unbalanced. Re-entry to a $1g$ environment will also cause temporary fluid shifts as the body returns to Earth normal. Whole-body vibration, such as seen during launch or landing, hurts perceptual accuracy (Conway et al., 2006). The added component of adapting to a different gravity may lead to additional head movement hypersensitivity or illusions of self-motion (NASA, 2010d). Reduced capacity to easily make sense of perceptual cues will lead to cognitive inefficiency. Inability to perceive accurately will reduce spatial cognition, decision making, and problem solving. Attention that could be focused on other problems will be used trying to make sense of the ambiguous stimuli.

Over a decade of work on the ISS has begun to shed light on individual variations in perceptual and cognitive functioning across time in microgravity. After an individual adapts to space, the crew members can successfully complete basic cognitive tasks (Fowler et al., 2008). How long-term space flight affects an individual's complex mental functioning is being assessed by an experiment sponsored by the European Space Agency led by L. Balazs and Guy Cheron (Balazs et al., 2009). Current evidence indicates that crew members adapt and return to, or near, baseline functioning. Nonetheless, the space traveler will need easily available reviews of training in-flight for complex or seldom used procedures. The designer of robotic aids must take into account the complete isolation, autonomy, confinement, and noise that the spacefarer will endure as well as understand human problem-solving processes (NASA, 2010c). The military has Earth-based training and robotic aids that can be modified for enhanced cognitive ability in space.

In human space flight, cognition includes adapting to new situations no matter how unexpected or novel, coping with off-nominal tasks, and resolving technical and social challenges. The human will make executive decisions and must be technically competent and healthy enough to maintain cognitive skills. Poor air quality, compromised immune systems, poor task design or displays,and user-unfriendly or user-incompatible robotic aids will all decrease cognitive efficiency (Manzey, 2000).

The configuration of the ISS is such that what is "up" in one module may be down in another. Smoke or hazy conditions will exacerbate the spatial and visual challenges of working across modules, thereby increasing the need for redundant auditory aids and clear interfaces and controls. Exposure to toxins, excessive radiation, infection, or poor air quality can easily reduce cognitive clarity, making the crew member more reliant on clarity of presentation and robotic aids. Visual perception is dependent upon good lighting without ambiguous shadows or blinding glare. This is especially true in off-nominal events or in the exhaustion of EVA.

Reviews of the stressors to cognitive functioning can be found in Kanas and Manzey (2008) and Bourne and Yaroush (2003). Individual and team cognition is affected by individualistic versus collective cultures. The design of habitat and communication interfaces, training of social roles, and on-board aids for multi-cultural group or metacognition are important to avoid cognitive inefficiency. Earth-based ground crews will also be multinational, often working according to space-craft time while living on earthbound time (Schmidt et al., 2009).

Off-nominal tasks increase the cognitive workload at the same time that the crew must work under the physical demands of microgravity. On long-duration spaceflight, cognitive workload could also be minimal, easily leading to boredom and automaticity. To maintain optimal performance and safety, human factor experts need to consider task design and analysis, monitoring, and intervention for either too high or too low workload.

Spacecraft and habitat must be compatible with normal human perception and cognitive processing. Signal-to-noise clarity is increased with easily recognized, consistent orientation and location across the environment, with clear visual, sound, and tactile contrast.

Impaired perceptual or cognitive functioning because of psychiatric reasons, head trauma, or infection will require on-board medical intervention or stabilization. Early recognition or prevention of such incidents is clearly the better option. Although the same is true on Earth, there is no way to quickly extract an impaired crew member. If perceptual or cognitive impairment occurs, the system infrastructure needs to be robust enough to support the crew member during reduced functioning.

## 7 ASTRONAUT SELECTION

Today's astronauts come from an international pool of candidates, including the European Space Agency (ESA), the Russian program, Canadian Space Agency (CSA), Japanese Aerospace Exploration Agency (JAXA), and the U.S. program. Each country selects its own astronaut candidates according to its own criteria. Planning the first U.S. astronaut selection in 1958, Allen O. Gamble, one of the psychologists on the medical team, realized that there needed to be some job or task analyses—a difficult challenge, as no one had ever flown in space before. He listed the duties of the first astronauts as the following (Link, 1965):

1. To survive; that is, to demonstrate the ability of humans to fly in space and return safely
2. To perform; that is, to demonstrate the human capacity to act usefully under conditions of space flight
3. To serve as backup for automatic controls and instrumentation; that is, to add reliability to the system
4. To serve as scientific observers; that is, to go beyond what instruments and satellites can observe and report
5. To serve as engineering observers and, acting as true test pilots, to improve the flight system and its components

Since the late 1950s there has always been some system of psychological selection, although there have been many changes in criteria and procedure. Originally, psychological assessment was extensive, requiring 30 h of psychological testing, plus interviews and evaluation by a team made up of a psychiatrist, an industrial–organizational psychologist, and management. In the 1960s the Lovelace clinic tested several women; 25 female pilots completed the same psychological evaluations as those given the males chosen for the Mercury project. Of these, 13 of them enrolled in an unofficial astronaut training program; none were declared official astronaut candidates.

From 1958 through 1969, astronaut selection occurred at least four more times. Since applicants already had extensive, often hazardous, flight experience, criteria emphasized emotional stability, motivation and energy, self-concept, and quality of interpersonal relationships. Psychological testing now required only 6.5 h, and the clinical evaluation was primarily psychiatric rather than psychological. This shift toward clinical content paralleled a shift away from research, reducing the data available for systematic scientific selection into astronaut selection. By 1983, Jones and Annes (1983) could write: "Presently, no psychological testing is done." Instead, the evaluation consisted of two consulting psychiatrists who separately interviewed each candidate for 2 h. This screening, although completed by expert aviation psychiatrists, did not have specific and objective criteria by which to rate each candidate.

After a hiatus of nine years, in 1978, astronaut selection began again for the space shuttle program, including nonpilots, scientists, and women. It was not until the 1980s that NASA hired its own psychiatrist and, soon thereafter, a psychologist to work in the

operational arena. From 1988 through 1990, a newly established in-house group met to improve the selection process. This first NASA Working Group on Psychiatric and Psychological Selection of Astronauts in 1988 distinguished between the roles of psychology and psychiatry and rewrote NASA psychiatric standards to include disqualifying psychiatric disorders based on the then-current American Psychiatric Association's *Diagnostic and Statistical Manual*. In addition, the working group defined the "best" psychological make-up for the job of astronaut and the "best" crew psychological mix, particularly for extended-duration space flights. Three attributes—aptitude, motivation, and sensitivity (referred to as select-in criteria)—were defined as being of equal importance in the selection of astronauts (Santy, 1994). Aptitude includes the psychological traits of intelligence and technical aptitude, history of professional success, adaptability and flexibility, being a team player, ability to represent NASA effectively, stress or discomfort tolerance, ability to function despite personal danger, ability to compartmentalize, ability to tolerate separation from loved ones, and ability to tolerate isolation. Motivation includes the psychological traits of achievement/goal orientation, hardworking/self-starting, mastery, persistence, optimism, no unhealthy motivation, healthy sense of competition, capacity to tolerate boredom, mission orientation, and healthy risk-taking behaviors. Finally, sensitivity to self and others includes the psychological traits of overall emotional maturity and stability, self-esteem, ability to form stable and quality interpersonal relationships, expressivity, sense of humor, insight and self-awareness, appropriate assertiveness, and cultural sensitivity. These three attributes coincide with the known astronaut tasks of systems management, sequence monitoring, motor tasks like steering vehicles and remote arm activities, repair and maintenance, conduct experiments, assembly, public relation activities, and self care (Santy, 1994). Beginning in 1989, NASA began using teams of external psychiatrists and psychologists to assist with the mental health portion of astronaut selection. These teams were asked to assess the applicants from a psychiatric and psychological viewpoint with the tasks of the astronauts. Subsequent behavioral health and performance selection meetings to review and update selection procedures were convened: NASA Psychiatric Astronaut Selection Standards meeting in 2001 to review psychiatric standards, NASA Select-Out review meeting in 2003, and an Astronaut Selection Working Group in 2008.

Holland (1999) notes that, by 1989, clinical testing had returned, giving some objective data to be used by the psychiatrists, but it was still a medical model. By the 1994–1995 selection cycle, nonmedical evaluations based on industrial–organizational principles and techniques were added to the clinical and medical models. Based on these organizational studies, Galarza and Holland (1999, p. 4) have listed the critical psychological proficiencies needed for space flight: "mental/emotional stability, ability to perform under stressful conditions, group living skills, teamwork skills, ability to cope with prolonged family separations, motivation,

judgment/decision making, conscientiousness, communication skills, leadership capability." These proficiencies or critical skills have continued to be assessed in all subsequent astronaut selection cycles (Hysong et al., 2007). Between 1978 and 2010 NASA had another13 selection group cycles that selected 257 U.S. astronauts. In the same timeframe, 94 Russian cosmonauts, 38 ESA astronauts, 14 Chinese astronauts, and 11 Canadian astronauts have been selected by their respective countries using very similar techniques of psychological testing, medical physicals, interviews, and practical exercises or simulations.

Long-duration missions aboard the ISS currently last six months. Training for long-duration missions is very arduous and takes approximately two to three years. This training requires extensive travel, including long periods away in other countries training with our international partners. Travel to and from the ISS will be by space shuttle until its retirement, which is expected in 2011. Following the shuttle retirement, all trips to and from the ISS will be aboard the Russian Soyuz vehicle.

Information about applying to the NASA astronaut selection process is available at http://nasajobs.nasa.gov/astronauts/content/broch00.htm (NASA, 2010a).

## 8 CONCLUSIONS

After 50 years of human space flight and 10 years on the ISS, we have gathered great quantities of information dealing with the crew and their interfaces. With a new mission in front of us, going beyond low Earth orbit, we must learn more about the challenges of long-term missions. We must gather much more data from ISS missions. Additionally, we must take advantage of analogs that are consistent with the perceived challenges of long-term missions and glean what we can to augment our knowledge base.

## REFERENCES

Allen, C. S., and Goodman, J. R. (2003), "Preparing for Flight: The Process of Assessing the ISS Acoustic Environment," in *Proceedings of NOISE-CON 2003*, Cleveland, OH.

Aschoff, J., and Wever, R. (1981), "The Circadian System of Man," in *Handbook of Behavioral Neurobiology*, Vol. 4, *Biological Rhythms*, J. Aschoff, F. A. King, and E. M. Stricker, Eds., Plenum, New York, pp. 311–331.

Balazs, L., Cheron, G., Achimowicz, J., Karmos, G., Barkaszi, I., Leurs, F., Bengoetxea Arrese, A. M., McIntyre, J., Berthoz, A., Molnar, M., Cebolla Alvarez, A. M., Nagy, E., Czigler, I., Pató, L., and de Saedeleer, C. (2009), "The Effect of Gravitational Context on EEG Dynamics: A Study of Spatial Cognition, Novelty Processing and Sensorimotor Integration (NEUROSPAT)," available: http://www.nasa.gov/mission_pages/station/science/experiments/Neurospat.html, accessed July 2010.

Banke, J. (2004), "Convenient Crew Handhold Caused Station Air Leak," available: http://www.space.com/missionlaunches/exp8_update_040123.html.

Barger, L. K., Wright, K. P., Jr., Czeisler, and C. A., "The Sleep of Crewmembers on Nine Space Shuttle Missions," paper presented at the NASA Human Research Program

Investigators' Workshop, South Shore Harbour Resort and Conference Center, League City, Texas, February 2008, Conference Proceedings, NASA Human Research Program Investigators' Workshop, February 2008.

Bluth, B. J. (1984), "Human Systems Interfaces for Space Stations," in *AIAA/NASA Space Systems Technology Conference Proceedings*, American Institute of Aeronautics and Astronautics, Costa Mesa, CA, National Aeronautics and Space Administration, Washington, DC, pp. 40–49.

Bond, R. L. (1977), *Application of Skylab Workday Analysis to Future Programs*, JSC 12856, Johnson Space Center, Houston, TX.

Bourne, L. E., Jr. and Yaroush, R. A. (2003), *Stress and Cognition: A Cognitive Psychological Perspective*, Final Report for NAG2-1561, NASACR-2003-212282, National Aeronautics and Space Administration, Washington, DC.

Campbell, S. S., and Dawson, D. (1990), "Enhancement of Nighttime Alertness and Performance with Bright Ambient Light," *Physiological Behavior*, No. 48, pp. 317–320.

Clark, J., and Allen, C. (2008), "Acoustic Issues" in *Principles of Clinical Medicine for Space Flight*, M. R. Barratt, and S. L. Pool, Eds., Springer, New York

Conway, G. E., Szalma, J. L. Saxton, B. M., Ross, J. M., and Hancock, P. A. (2006), "The Effects of Whole-Body Vibration on Human Performance: A Meta-Analytic Examination," in *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*, pp. 1741–1745.

Czeisler, C. A. (1999), *Circadian Entrainment, Sleep-Wake Regulation and Neurobehavioral Performance during Extended Duration Space Flight*, Report 20000029467, National Aeronautics and Space Administration Center for AeroSpace Information, Moffett Field, CA.

Czeisler, C. A., Dijk, D., and Neri, D. F. (1999), "Ambient Light Intensity, Actigraphy, Sleep and Respiration, Circadian Temperature and Melatonin Rhythms and Daytime Performance of Crew Members during Space Flight on STS-90 and STS-95 Missions" [abstract], in *Proceedings of the First Biennial Space Biomedical Investigators' Workshop*, League City, TX, National Aeronautics and Space Administration, Houston, TX.

Dijk, D. J., Neri, D. F., and Wyatt, J. K. (2001), "Sleep, Performance, Circadian Rhythms, and Light-Dark Cycles during Two Space Shuttle Flights," *American Journal of Physiology: Regulatory, Integrative, and Comparative Physiology*, Vol. 281, pp. R1647–R1664.

Fowler, B., Meehan, S., and Singhal, A. (2008), "Perceptual-Motor Performance and Associated Kinematics in Space," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 50, No. 6, pp. 879–892.

Galarza, L., and Holland, A. W. (1999), "Critical Astronaut Proficiencies Required for Long-Duration Space Flight," in *Proceedings of the 29th International Conference on Environmental Systems*, ICES 1999CD, Paper No. 1999-01-2096, Society of Automotive Engineers, Denver, CO.

Goodman, J. R. (1992), "Space Shuttle Crew Compartment Debris/Contamination," ICES Paper 921345, in *Proceedings of the 22nd International Conference on Environmental Systems*, Seattle, WA.

Grigor'yev, A. I., Bugrov, S. A., and Bogomolov, V. V. (1990), "Review of the Major Medical Results of the 1-Year Flight on Space Station 'Mir,'" *Kosmichesleaya Biologiya i Aviakosmichesleaya Meditsina*, Vol. 24, No. 5, pp. 3–10.

Gronfler, C., Wright, K. P., Kronauer, R. E., and Czeisler, C. A. (2007), "*Entrainment of the Human Circadian Pacemaker to Longer-Than-24-h Days*," *Proceedings in the National Academy of Sciences*, Vol 104, No. 21, pp. 9081–9086.

Harwood, W. (2009), "Station's Carbon Dioxide Scrubber Back On Line," available: http://spaceflightnow.com/shuttle/sts127/090726fd12/index.html

Holland, A. W. (1999), "Psychology of Spaceflight," in *Human Spaceflight: Mission Analysis and Design*, W. J. Larson and L. K. Pranke, Eds., McGraw-Hill, New York.

Hysong, S. J., Galarza, L., and Holland, A. W. (2007), "A Review of Training Methods and Instructional Techniques: An Implication for Behavioral Skills Training in U.S. Astronauts," available: http://ston.jsc.nasa.gov/collections/TRS/_techrep/TP-2007-213726.pdf.

Jahweed, M. M. (1994), "Muscle Structure and Function," in *Space Physiology and Medicine*, 3rd ed., A. E. Nicogossian, C. L. Huntoon, and S. L. Pool, Eds., Lea and Febiger, Philadelphia, PA.

Jones, D. R., and Annes, C. A. (1983), "The Evolution and Present Status of Mental Health Standards for Selection of USAF Candidates for Space Missions," *Aviation, Space, and Environmental Medicine*, Vol. 54, No. 8, pp. 730–734.

Jones, T. (2004), "The First 1000 Days," *Air & Space Magazine*, Smithsonian, Washington, DC, available: http://www.airspacemag.com/space-exploration/days.html.

Kanas, N., and Manzey, D. (2008), *Space Psychology and Psychiatry*, 2nd ed., Space Technology Library, Microcosm Press,

El Segundo CA. Kloeris, V. L., and Bourland, C. T. (2003), "The Food System for the International Space Station: The First Five Increments," in *Proceedings of the 33rd International Conference on Environmental Systems*, Society of Automotive Engineers, Denver, CO.

Link, M. M. (1965), *Space Medicine in Project Mercury*, Special Publication 4003, National Aeronautics and Space Administration, Washington, DC.

Mallis, M. M., and DeRoshia, C. S. (2003), "Circadian Rhythms, Sleep, and Performance in Space: A Focus of NASA Research," in *Special Supplement of Aviation, Space and Environmental Medicine*, *Proceedings of the New Directions in Behavioral Health: Integrating Research and Application Conference*, December 2-3.

Manzey, D. (2000), "Monitoring of Mental Performance During Spaceflight," *Aviation, Space and Environmental Medicine*, Vol. 7, pp. A69–A75.

Mohanty, S. (2001), *Design Concepts for Zero-G Whole Body Cleansing on ISS Alpha*, Part II, *Individual Design Project*, NASA CR-2001-208931, National Aeronautics and Space Administration, Washington, DC.

Mount, F. E. (1989), "Session Summary: Maintenance," in *Proceedings of the Manned System: A Human Factors Symposium and Workshop*, American Astronautical Society, Houston, TX.

Mount, F. E. (2002), "Habitability: An Evaluation," in *Isolation: NASA Experiments in Closed-Environment Living*, H. W. Lane, R. L. Sauer, and D. L. Feeback, Eds., Univelt, San Diego, CA, pp. 87–116.

Mount, F. E., Adam, S., McKay, T., Whitmore, M., Merced-Moore, D., Holden, T., Wheelwright, C., Koros, A., O'Neal, M., Toole, J., and Wolf, S. (1994), *Human Factors Assessments of the STS-57 SpaceHab 1 Mission*, NASA TM-104802, National Aeronautics and Space Administration, Washington, DC.

Mount, F. E., Whitmore, M., and Stealey, S. L. (2003), *Evaluation of Neutral Body Posture on Shuttle Mission STS-57 (SPACEHAB-1)*, NASA TM-2003-104805, National Aeronautics and Space Administration, Washington, DC.

National Aeronautics and Space Administration (NASA) (1982), *Space Physiology and Medicine*, NASA SP-447, NASA, Washington, DC.

National Aeronautics and Space Administration (NASA) (1983), *Crew Interface Panel Space Station Habitability Requirements Document*, JSC 19517, NASA, Johnson Space Center, Houston, TX.

National Aeronautics and Space Administration (NASA) (1992), *Appendix K of the Space Shuttle Crew Procedures Management Plan*, Mission Operations Directorate, Operations Division, NASA, Johnson Space Center, Houston, TX.

National Aeronautics and Space Administration (NASA) (1995), *Man-Systems Integration Standards*, NASA STD-3000, NASA, Washington, DC.

National Aeronautics and Space Administration (NASA) (2000), "Expedition One December Crew Log," available: http://spaceflight.nasa.gov/station/crew/exp1/exp1shepdec.html.

National Aeronautics and Space Administration (NASA) (2003), *Astronaut Fact Book*, NP-2003-07-008JSC, NASA, Johnson Space Center, Houston, TX.

National Aeronautics and Space Administration (NASA) (2004), *ISS System Specification*, SSP 41000AR, NASA, Johnson Space Center, Houston, TX.

National Aeronautics and Space Administration (NASA) (2009a), *Space Flight Human System Standard*, Vol. 1, *Crew Health*, NASA-STD-3001, NASA, Washington, DC.

National Aeronautics and Space Administration (NASA) (2009b), *Human Exploration of Mars Design Reference Mission Architecture* 5.0. NASA-SP-566, http://www.nasa.gov/pdf/373665main_NASA-SP-2009-566.pdf, p. 67.

National Aeronautics and Space Administration (NASA) (2010a), "Astronaut Selection," available: http://nasajobs.nasa.gov/astronauts/content/broch00.htm.

National Aeronautics and Space Administration (NASA) (2010b), "Water on the Space Station-NASA Science," available: http://science.nasa.gov/science-news/science-at-nasa/2000/ast02nov_1/.

National Aeronautics and Space Administration (NASA) (2010c), *Human Integration Design Handbook*, NASA/SP-2010-3407, NASA, Washington, DC, available: http://ston.jsc.nasa.gov/collections/TRS/_techrep/SP-2010-3407.pdf.

Whitmire, A. M., Leveton, L. B., Barger, L., Brainard, G., Dinges, D.F., Klerman, E., Shea, C. "Risk of Performance Errors Due to Sleep Loss, Circadian Desynchornization, Fatigue and Work Overload," available: http://humanresearch.jsc.nasa.gov/elements/smo/hrp_evidence_book.asp.

National Aeronautics and Space Administration (NASA) (20011), *Space Flight Human System Standard*, Vol. 2, Human Factors, Habitability, and Environmental Health, NASA-STD-3001, NASA, Washington, DC.

Paloski, W. H., Oman, C. M., Bloomberg, J. J., Reschke, M. F., Wood, S. J., Harm, D. L., Peters, B. T., Mulavara, A. P., Locke, J. P. and Stone, L. S. (2008). "Risk of Sensory-Motor Performance Failures Affecting Vehicle Control during Space Missions: A Review of the Evidence," *Journal of Gravitational Physiology*,. Vol. 15, No. 2, 00. 1–29.

Perchonok, M., and Bourland, C. (2002), *NASA Food Systems: Past, Present, and Future*, National Space Biomedical Research Institute, Houston, TX.

Robinson, J. A., Slack, K., Olson, V. A. Trenchard, M., Willis, K., Baskin, P., and Ritsher, J. B. (2006), "Patterns in Crew-Initiated Photography of Earth from ISS–Is Earth Observation a Salutogenic Experience?," paper presented at the International Astroautical Congress, Valencia, Spain.

Robinson, J. A., Slack, K. J., Olson, V., Trenchard, M., Willis, K., Baskin, P., and Boyd, J. (2011), "Patterns in Crew-Initiated Photography of Earth from ISS--Is Earth Observation a Salutogenic Experience?," in *Psychology of Space Exploration: Contemporary Research in Historical Perspective*, D. A. Vakoch, Ed., NASA History Series, National Aeronautics and Space Administration, Washington, DC.

Santy, P. A. (1994), *Choosing the Right Stuff: The Psychological Selection of Astronauts and Cosmonauts*, Praeger, Westport CT.

Schmidt, L. L. Keeton, K., Slack, K. J., Shea, C., and Leveton, L. B. (2009), "Risk of Performance Errors due to Poor Team Cohesion and Performance, Inadequate Selection/Team Composition, Inadequate Training, and Poor Psychosocial Adaptation," in *Human Health and Performance Risks of Space Exploration Missions*, J. C. McPhee and J. B. Charles, Eds., NASA SP-2009-3405, pp. 53–84, http://humanresearchroadmap.nasa.gov/evidence/reports/EvidenceBook.pdf.

Schneider, V. S., LeBlanc, A. D., and Taggard, L. C. (1994), "Bone and Mineral Metabolism," in *Space Physiology and Medicine*, 3rd ed., A. E. Nicogossian, C. L. Huntoon, and S. L. Pool, Eds., Lea and Febiger, Philadelphia, PA.

Thompson, A. (2008), "Living on Mars Time: Scientists Suffer Perpetual Jet LAG," available: http://www.space.com/scienceastronomy/080729-st-mars-time.html.

Watson, J. K.. (2003), "Supportability Concepts for Long-Duration Human Exploration Missions," AIAA Paper 2003-6240, presented at the AIAA Space 2003 Conference, Houston, TX.

Wheelwright, C. D., Lengel, R. D., and Koros, A. S. (1994), "Noise, Vibration and Illumination," in *Space Biology and Medicine*, Vol. II, *Life Support and Habitability*, American Institute of Aeronautics and Astronautics, Washington, DC.

Young, K. S, Mesloh, M., and Rajulu, S. L. (2010), "Preliminary Results of the Effect of Spinal Elongation in Microgravity on Seated Height," paper presented at the Aerospace Medicine Association 81st Annual Scientific Meeting, Phoenix, AZ.

# PART 6
# PERFORMANCE MODELING

# CHAPTER 32

# MODELING HUMAN PERFORMANCE IN COMPLEX SYSTEMS

**K. Ronald Laughery, Jr., Beth Plott, Michael Matessa, and Susan Archer**
**Alion Science and Technology, MA&D Operation**
**Boulder, Colorado**

**Christian Lebiere**
**Carnegie Mellon University**
**Pittsburgh, Pennsylvania**

## 1 INTRODUCTION

Over the past few decades, human factors and ergo-nomics practitioners have been called upon increasingly early in the system design and development process. Early inputs from all disciplines result in better and more integrated designs, as well as lower costs, than if one or more disciplines are solely in charge, find out late in the development stage that changes are required, and then call upon the expertise of the other disciplines. Our goal as human factors and ergonomics practitioners should be to provide substantive and well-supported input regarding the human(s), his or her interaction(s) with the system, and the resulting total performance. Total performance includes a number of converging measures, including task latency, type and probability of errors, quality of performance, and workload measures. Furthermore, we should be prepared to provide this input from the earliest stages of system concept development and then throughout the entire system or product life cycle.

To meet this challenge, many human factors and ergonomics tools and technologies have evolved over the years to support early analysis and design. Two specific types of technologies are design guidance (e.g., Boff et al., 1986; O'Hara et al., 1995) and high-fidelity

rapid prototyping of user interfaces (e.g., Dahl et al., 1995). Design guidance technologies, in the form of either handbooks or computerized decision support systems, put selected portions of the human factors and ergonomics knowledge base at the fingertips of the designer, often in a form tailored to a particular problem, such as nuclear power plant design or Unix computer interface design. However, design guides have the shortcoming that they do not often provide methods for making quantitative trade-offs in *system* performance as a function of design. For example, design guides may tell us that a high-resolution color display will be better than a black-and-white display, and they may even tell us the value in terms of increased response time and reduced error rates. However, this type of guidance will rarely provide good insight into the value of this improved element of the human's performance to the *overall system's* performance. As such, design guidance has limited value for providing concrete input to *system-level performance prediction*.

Rapid prototyping, on the other hand, supports analysis of how a specific design and task allocation will affect human and system-level performance. The disadvantage of prototyping, as with all human subject experimentation, is that it can be slow and costly. In particular, prototypes of hardware-based systems, such

as aircraft and machinery, are very expensive to develop, particularly at early design stages when there are many widely divergent design concepts. Despite the expense, hardware and software prototyping is an important tool for the human factors practitioner, and its use is growing in virtually every application area.

Although these technologies are valuable to the human factors practitioner, what is often needed is an integrating methodology that can extrapolate from the base of human factors and ergonomics data, as reflected in design guides and the literature, to support system-level performance predictions as a function of design alternatives. This methodology should also bind with rapid prototyping and experimentation in a mutually supportive and iterative way. As has become the case in many engineering disciplines, a prime candidate for this integrating methodology is computer modeling and simulation.

Computer modeling of human behavior and performance is not a new endeavor. Computer models of complex cognitive behavior have been around for over 20 years (e.g., Newell and Simon, 1972; Card et al., 1983) and tools for computer modeling of task-level performance have been available since the 1970s (e.g., Wortman et al., 1978). However, three trends have emerged in the past decade to promote the use of computer modeling and simulation of human performance as a standard tool for the practitioner. First is the rapid increase in computer power and the associated development of easier-to-use modeling tools. People with an interest in predicting human performance through simulation can select from a variety of computer-based tools. For a comprehensive list of these tools, see the Defense Technical Information Center (DTIC) Directory of Design Support Methods (DDSM). The DDSM contains references to human systems integration (HSI) design and interface tools, techniques, databases, guides, and standardization documents. Second is the increased focus by the research community on the development of *predictive* models of human performance rather than simply descriptive models. For example, the goals–operators–methods–selection rules (GOMS) model (Gray et al., 1993) represents the integration of research results into a model for making predictions of how humans will perform in a realistic task environment. Another example is the research in cognitive workload that has been represented as computer algorithms (e.g., McCracken and Aldrich, 1984; Farmer et al., 1995). Given a description of the tasks and equipment with which humans are engaged, these algorithms support assessment of when workload-related performance problems are likely to occur and often include identification of the quantitative impact of those problems on overall system performance (Hahler et al., 1991). These algorithms are particularly useful when embedded as key components in computer simulation models of the tasks and the environment. Third is the integration of those algorithms into cognitive architectures that integrate cognition, perception, and action into a single computational framework that can be applied to a broad range of tasks, from basic laboratory experiments used to validate the architectural mechanisms to

predicting operator performance on complex practical tasks (Gray et al., 1997).

Perhaps the most powerful aspect of computer modeling and simulation is that it provides a method through which the human factors and ergonomics team can "step up to the table" with the other engineering disciplines, which also rely on quantitative computer models. What we discuss in this chapter are the methods through which the human factors and ergonomics community can contribute early to system design trade-off decisions.

## 1.1 Chapter Objectives

In this chapter we discuss some existing computer tools for modeling and simulating human–system performance. It is intended to provide the reader with an understanding of the types of human factors and ergonomics issues that can be addressed with modeling and simulation and some of the tools that are now available to assist the human factors and ergonomics specialist in conducting model-based analyses and an appreciation of the level of expertise and effort that will be required to use these technologies. We begin with two caveats. The first is that we are not yet at a point where computer modeling of human behavior allows sufficiently accurate predictions that no other analysis method (e.g., prototyping) is needed. In the early stages of system concept development, high-level modeling of human–system interaction may be all that is possible. As the system moves through the design process, human factors and ergonomics designers will often want to augment modeling and simulation predictions with prototyping and experimentation. In addition to providing high-fidelity system performance data, these data can be used to constrain, enhance, and refine the models. This concept of human performance modeling supporting and being supported by experimentation with human subjects is represented in Figure 1. In essence, simulation provides the human factors and ergonomics practitioner with a means of extending the knowledge base of human factors and of amplifying the effectiveness of limited experimentation.

The second caveat is that the technologies discussed here are evolving rapidly. We can be certain that every tool discussed is undergoing constant change and that new modeling tools are being developed. We are discussing computer-based tools, and we expect the pace of change in these tools to mirror the pace in other software tools, such as word processors, spreadsheets, presentation and productivity tools, and Internet-based applications. These detailed discussions of several of the modeling tools are included to facilitate better understanding of human performance modeling tools. We encourage the reader to follow citations in this chapter to assess the current state of any tool. Most of these modeling tools have large, active user communities that maintain websites to provide introductory tutorials, software downloads, validated models, and published papers. These resources are invaluable both for the experienced modeler trying to stay abreast of recent developments and the novice user attempting to get up to speed on a new technology.
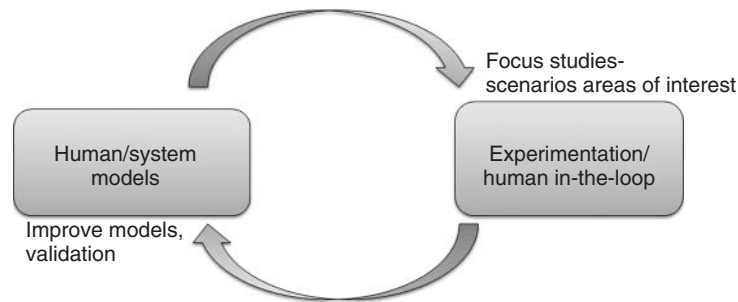
**Figure 1** Synergy between modeling and experimentation.

## 2 QUESTIONS ADDRESSED BY HUMAN PERFORMANCE MODELS

Below are a few classes of problems to which human–system modeling has been applied:

- How long will it take a human or team of humans to perform a set of tasks as a function of system design, task allocation, and individual capabilities?
- What are the performance trade-offs for different combinations of design, task allocation, and individual capability selections?
- What are the workload demands on the human as a function of system design and automation?
- How will human performance and resulting system performance change as the demands of the environment change?
- How many people are required on a team to ensure safe, successful performance?
- How should tasks be allocated to optimize performance?
- How will environmental stressors such as heat, cold, or vibration affect human–system performance?

The list above is a sample rather than an exhaustive list. The tools we discuss in this chapter are inherently flexible and we consistently discover that these tools can be used to solve problems that the tool developers never conceived. To assess the potential of simulation to answer questions, in every potential human performance modeling project we should first determine the specific questions that the project is trying to answer. Then we can conduct a critical assessment of what is important in the human–machine system being modeled. This will define the required content and fidelity of the model. The questions that should be considered about the system include:

1. *Human Performance Representation*. What time or duration of performance is important? How is human performance initiated, and what resolution of behavior is required? What aspects

of human performance, including task management, load management, and goal management, are expected? How much is known and constrained about the knowledge and strategies that human users bring to bear on this task?

2. *Equipment Representation*. What equipment is used to accomplish the task? To what level of functional and physical description can and should equipment be represented? Is it operable by more than one human or system component?

3. *Interface Requirements*. What information needs to be conveyed to the humans and when? Is transformation of information required? How often is information updated and monitored?

4. *Control Requirements*. What processes need to be controlled by the human and to what level of resolution? How much attention is required by the human to perform control changes?

5. *Logical and Physical Constraints*. How is performance supported through equipment operability and procedural sequences? What alarms and alerts should be represented?

6. *Simulation Driver*. What makes the system function? The occurrence of well-defined events (e.g., a procedure), the passage of time (e.g., the control of a vehicle), or a hybrid of both?

In using human performance models, perhaps the most significant task of the human factors practitioner is to determine what aspects of the human–machine system to include in the model and what to leave out. By defining the purpose of the model and then answering the questions above, the human factors practitioner will get a sense of what is important in the system and therefore what may need to be represented in a model. Many modeling studies have failed because of the inclusion of too many factors that, although a part of human–system performance, were not system performance drivers. Consequently, the models become overly complex and expensive to develop. In our experience, it is better to begin with a model with too few aspects of the system represented and then add to it than to begin a modeling project by trying to model everything. The first approach may succeed, whereas the second is often doomed. It is also important that the

level of detail is consistent with the types of data that are available.

Additionally, the human factors practitioner should consider the measures of effectiveness of the system that the model should be designed to predict. In building the model, it is important to remember that the goal will be to predict measures of human performance that will affect system performance. Therefore, a clear definition of what is important to performance is necessary. The following aspects of performance measures should be considered:

1. *Success Criteria.* What operational success measures are important to the system? Can these be stated in relative terms or must they be measured in absolute terms?

2. *Range of Performance to Be Studied.* What experimental variables are to be explored by the model? How important is it to establish a range of performance for each experimental condition as a function of the stochastic (i.e., random) behavior of the system?

By asking the foregoing questions prior to beginning a modeling project, the human factors practitioner can develop a better sense of what is important in the system in terms of both aspects that drive system performance and the measures of effectiveness that are truly of interest. Then, and only then, can a human performance modeling project begin with a reasonable hope of success.

In the remainder of this chapter we discuss two classes of modeling tools for human performance simulation, then report on recent efforts to unify those two complementary classes in order to leverage their strengths and alleviate their shortcomings. After discussing each class of modeling tool, we provide specific examples of a modeling tool and then provide case studies about how these tools have been used in answering real human performance questions.

## 3   CLASSES OF SIMULATION MODELS

Human performance can be highly complex and involve many types of processes and behavior. Over the years many models have been developed that predict sensory processes (e.g., Gawron et al., 1983), aspects of human cognition (e.g., Newell, 1990), and human motor response (e.g., Fitts' law). The current literature in the areas of cognitive engineering, error analysis, and human–computer interaction contains many models, descriptions, methodologies, metaphors, and functional analogies. However, in this chapter we are not focusing on the models of these individual elements of human behavior but rather, on models that can be used to describe human performance in systems. These human–system performance models typically include some of these elemental behavioral models as components but provide a structural framework that allows them to be integrated with each other and put in the context of human performance of tasks in systems.
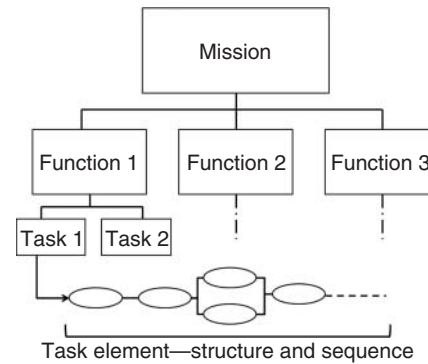


**Figure 2**   Reductionist models of human performance.

We separate the world of human–system performance models into two general categories that can be described as reductionist models and first-principle models. *Reductionist models* use human–system task sequences as the primary organizing structure, as shown in Figure 2. The individual models of human behavior for each task or task element are connected to this task-sequencing structure. We refer to it as reductionist because the process of modeling human behavior involves taking the larger aspects of human–system behavior (e.g., "perform the mission") and then reducing them successively to smaller elements of behavior (e.g., "perform the function," "perform the tasks"). This continues until a level of decomposition is reached at which reasonable estimates of human performance for the task elements can be made. One can also think of this as a top-down approach to modeling human–system performance. The example of this type of modeling that we use in this chapter is *task network modeling*, where the basis of the human–system model is a task analysis.

*First-principle models* of human behavior are structured around an organizing framework that represents the underlying goals, principles, and mechanisms of human performance (Figure 3). Tools that support first-principle modeling of human behavior have structures embedded in them that represent elemental aspects of human performance. For example, these models might directly represent processes such as goal-seeking behavior, task scheduling, sensation and perception, cognition, and motor output. In turn, those processes might invoke fundamental actions such as shifts of attention, memory retrieval, and conflict resolution among competing courses of action. To use tools that support first-principle modeling, one must describe how the system and environment interacts with the human processes being modeled. In this chapter we focus on the adaptive control of thought–rational (ACT-R) cognitive architecture (Anderson and Lebiere, 1998).

It is worth noting that these two modeling strategies are not mutually exclusive and, in fact, can be mutually supportive in any given modeling project. Often, when one is modeling using a reductionist approach, one needs models of basic human behavior to represent behavioral phenomena accurately and therefore must draw on
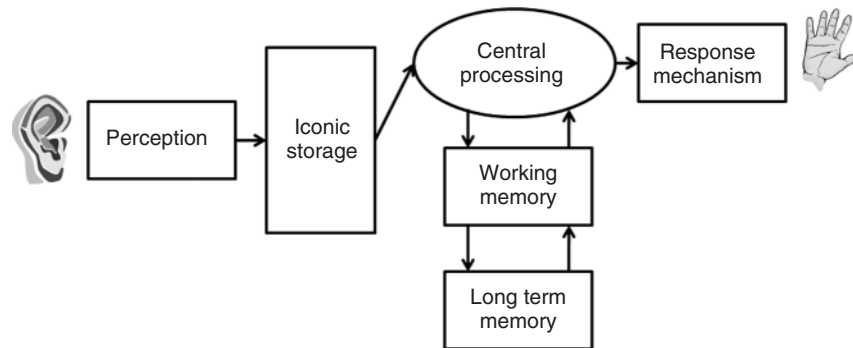
**Figure 3** First-principle models of human performance.

elements of first-principle models. Alternatively, when one is modeling human–system performance using a first-principle approach, some aspects of human–system performance and interrelationships between tasks may be more easily defined using a reductionist approach. Both classes of model have been used to model individual and team performance. It is also worth noting that recent advances in human performance modeling tool development are blurring the distinctions between these two classes (e.g., Hoagland et al., 2001). Increased emphasis on interoperability between models has caused researchers and developers to focus on integrating reductionist and first-principle models. In the final section of this chapter we present one such attempt at integrating the ACT-R cognitive architecture with the Improved Performance Research Integration Tool (IMPRINT).

## 4 REDUCTIONIST APPROACH: TASK NETWORK MODELING

One technology that has proven useful for predicting human–system performance is *task network modeling.* In a task network model, human performance is decomposed into tasks. The fidelity of this decomposition can be selective, with some functions being decomposed several levels and others just one or two. This is, in human factors engineering terms, the task analysis. The sequence of tasks is defined by constructing a *task network*. This concept is illustrated in Figure 4, which presents a sample task network for driving while talking on a cell phone.

Task network modeling is an approach to modeling human performance in complex systems that has evolved for several reasons. First, it is a reasonable means for extending the human factors staple: the task analysis. Task analyses organized by task sequence are the basis for the task network model. Second, task network models can include sophisticated submodels of the system hardware and software to create a closed-loop representation of relevant aspects of the human–machine system. Third, task network modeling is relatively easy to use and understand. Recent advancements in task network modeling technology have made

this technology more accessible to human factors practitioners. Finally, task network modeling can provide efficient, valid, and useful input to many types of issues. With a task network model, the human factors engineer can examine a design (e.g., control panel redesign) and address questions such as "How much longer will it take to perform this procedure?" and "Will there be an increase in the error rate?" Generally, task network models can be developed in less time and with substantially less effort than would be required if a prototype were developed and human subjects used. However, as stated before, for revolutionary designs, modeling may not alleviate the need for empirical data collection.

Task network models of human performance have been subjected to validation studies with favorable results (e.g., Lawless et al., 1995; Engh et al., 1998). However, as with any modeling approach, the real level at which validation must be considered is with respect to a particular model, not with respect to the general approach.

### 4.1 Components of a Task Network Model

To represent complex, dynamic human–system behavior, many aspects of the system may need to be modeled in addition to simply task lists and sequence. In this section we use the task network modeling tool *Micro Saint Sharp* as an example. The basic ingredient of a Micro Saint Sharp task network model is the task analysis as represented by a network or series of networks. The level of system decomposition (i.e., how finely we decompose the tasks) and the amount of the system that is simulated depend on the particular problem. For example, in a power plant model, one can create separate networks for each of the operators and one for the power plant itself. Although the networks may be independent, performance of the tasks can be interrelated through shared variables. The relationships among different components of the system, represented by different segments of the network, can then communicate through changes in these shared variables. For example, when an operator manipulates a control, this may initiate an "open valve" task in a network representing the plant. This could ripple through to a network representing other operators and subsystems and their
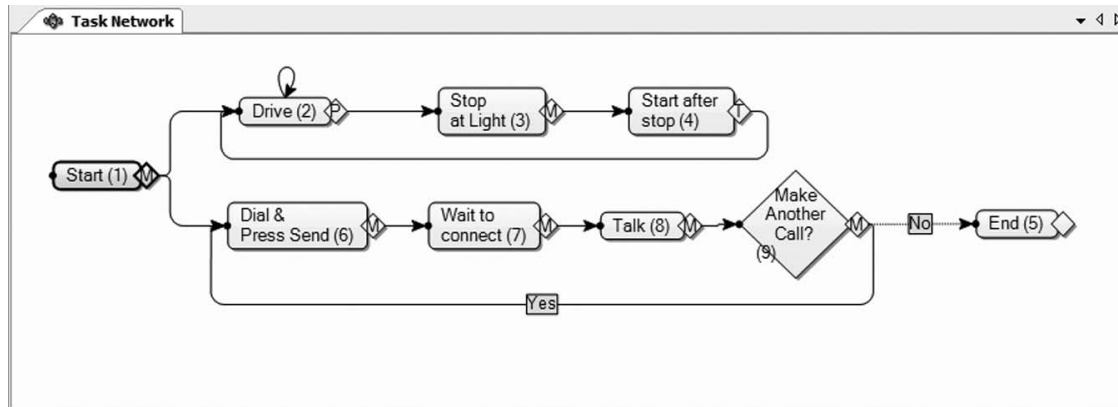
**Figure 4**   Task network model representing a human driving while talking on a cell phone.

response to the open valve. This basic task network is built in Micro Saint Sharp via a point-and-click drawing palette. Through this environment, the user creates a network as shown in Figure 5. Networks can be embedded within networks, allowing for hierarchical construction. In addition, the shape of the nodes on the diagram can be chosen to represent specific types of activity.

To reflect complex task behavior and interrelationships, more detailed characteristics of the tasks need to be defined. By double clicking on a task, the user opens up the task description window, as shown in Figure 6. Below are descriptions of each of the items on the tabs in this window.

- *Task ID*. This value is an arbitrary number for task referencing.
- *Task Name*. This parameter contains a text string used to identify the task.
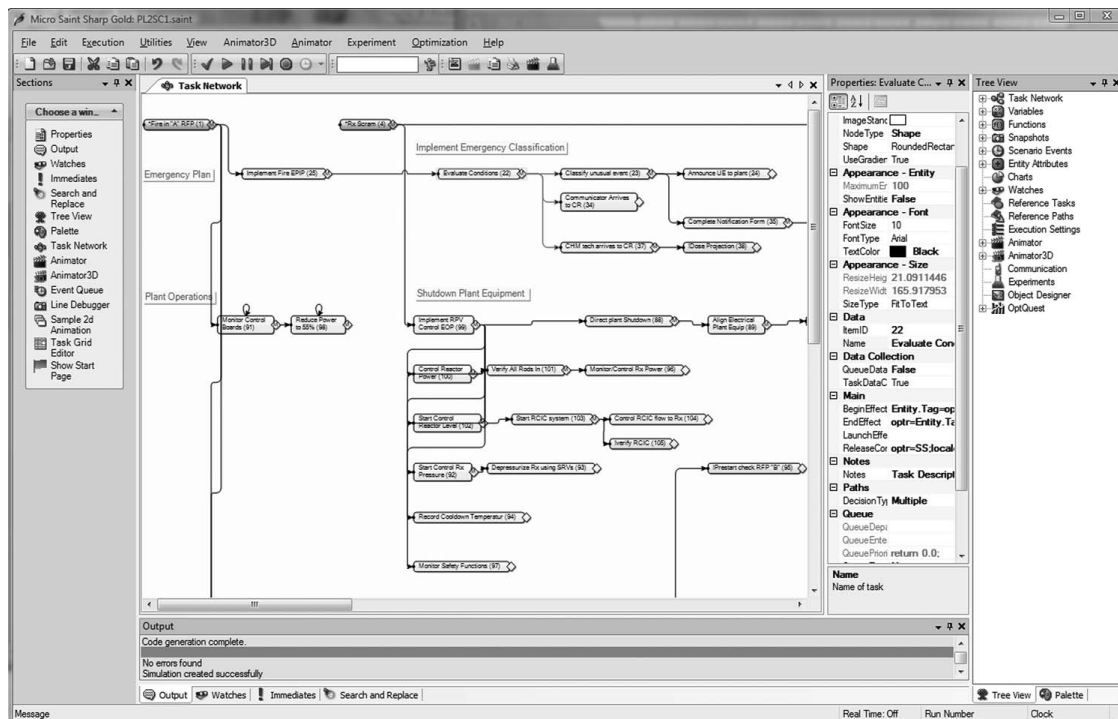


**Figure 5**   Main window in Micro Saint Sharp for task network construction and viewing.

**Figure 6** User interface in Micro Saint Sharp for providing input on a task.

- *Time Distribution*. Micro Saint Sharp conducts Monte Carlo simulations with task performance times sampled from a distribution as defined by this option (e.g., normal, beta, exponential).
- *Mean Time*. This parameter defines average task performance time for this task. This can be a number, equation, or algorithm, as can all values in the fields described below.
- *Standard Deviation*. This value contains the standard deviation of the task performance time, assuming that the user has chosen a distribution that is parameterized by a standard deviation.
- *Release Condition*. Data in this field determine when a task begins executing. For example, a condition stating that this task will not start before an operator is available might be represented by a release condition such as the following:

  OperatorBusy == false;

  In other words, for the task to begin, the value of the variable "OperatorBusy" must be false. This task would wait until the condition was true

before beginning execution, which would probably occur as a result of the operator completing the task he or she is currently performing.

- *Beginning Effect*. This field permits the user to define how the system will change as a result of the commencement of this task. For example, if this task used an operator that other tasks might need, we could set the following condition to show that the operator is unavailable while he or she performed this task:

  OperatorBusy = true;

  Assignment and modification of variables in beginning effects are one principal way in which tasks are interrelated.

- *Launch Effect*. This data element is similar to a task beginning effect but is used to launch high-resolution two- (2D) and/or three- (3D) dimensional animation of the task.
- *Ending Effect*. This field contains the definition of how the system will change as a result of the completion of this task. From the previous example, when this task was complete and the

operator became available, we could set the ending effect as follows:

$$\text{OperatorBusy} = \text{false};$$

At this point, another task waiting for an operator to become available could begin. Ending effects are another important way in which tasks can be interrelated through the assignment and modification of variables.

Another notable aspect of the task network diagram window shown in Figure 5 is the diamond-shaped icon that follows every task. This icon encapsulates data that describe the paths and the associated logic that will be executed when this task is completed. Often, this logic represents a human decision-making process. In that case, the branches align to potential courses of action that the modeled human could select. To define the decision logic, the Micro Saint Sharp user would use the "Paths" tab on the task description dialogue, as shown in Figure 7. There are three general types of decisions to model:

- *Probabilistic*. In probabilistic decisions, the human will begin one of several tasks based on a random draw weighted by the probabilistic branch value. These weightings can be dynamically calculated to represent the current context of the decision. For example, this decision type

might be used to represent human error likelihoods and would be connected to the subsequent tasks that would be performed.

- *Tactical*. In tactical decisions, the human will begin one of several tasks based on the branch with the highest "value." This could be used to model the many types of rule-based decisions that humans make, as illustrated in Figure 7.
- *Multiple*. This would be used to begin several tasks at the completion of this task, such as when one human issues a command that begins other crew members' activities.

The expression fields in Figure 7 represent the values associated with each branch. The values can be numbers, expressions, or complicated algorithms defining the probability (for probabilistic branches) or the desirability (for tactical and multiple branches) of taking each branch in the network. Again, any value on this screen can be not simply numbers but also variables, algebraic expressions, logical expressions, or groups of algebraic and logical expressions that would, essentially, form a subroutine. As the model executes, Micro Saint Sharp includes a parser that evaluates the expressions included in the branching logic when it is encountered in the task network flow. This results in a dynamic network in which the flow through the tasks can be controlled with variables that represent equipment state, scenario
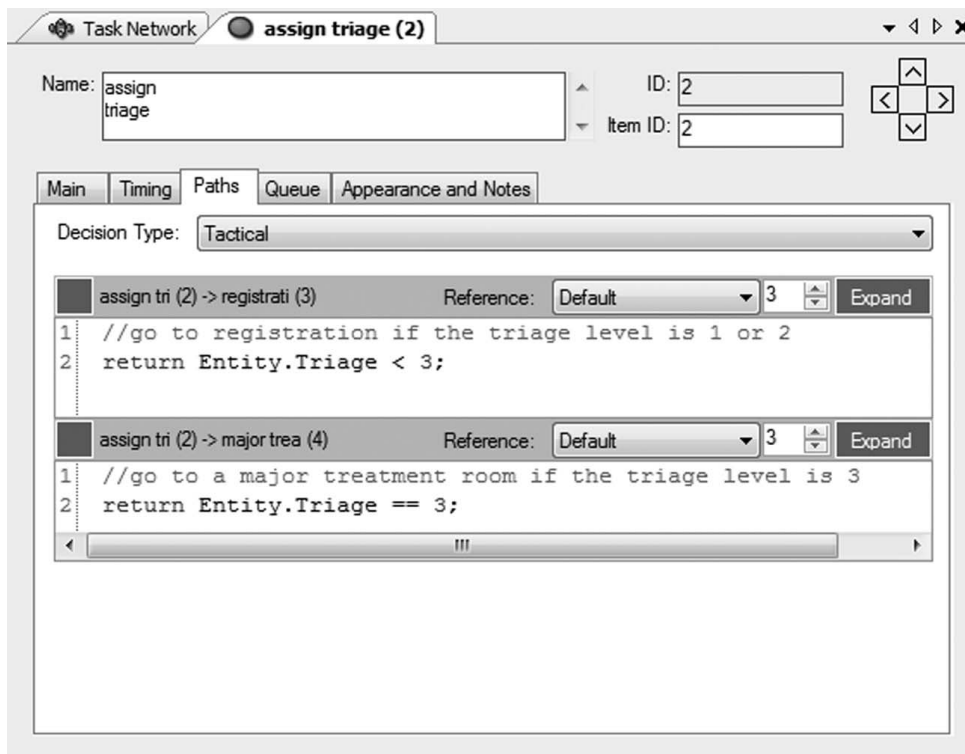


**Figure 7**   User interface in Micro Saint Sharp for defining task-branching decision logic.

context, or the task loading of the humans in the system, to name a few examples. It is the power of this parser that provides many task network models with the ability to address complex problems.

The research community has been inspired by the Micro Saint Sharp architecture, and several efforts have enhanced the capability of the task network model to more accurately represent theoretical advances. Specifically, Warwick and Santamaria (2009) extended the available decision types described above to a "recognition-primed" decision type (Klein, 1998). This new decision type represents human decisions that are experience driven and in which the subsequent action choices are driven by recognition of aspects of the scenario, rather than by a rule-based process. To implement this decision type, the model develops and maintains an ongoing representation of the human's memory, which either can be preloaded prior to model start or can be populated (i.e., trained) as the model proceeds. Preliminary validation work on this model has been encouraging and provides opportunities for more complete representations of "natural" human behavior.

The Command, Control, and Communications— Techniques for Reliable Assessment of Concept Execution (C3TRACE) tool has also expanded the available decision types to allow branching based on message communication type (digital, face to face, written, etc.) or decision quality (Plott et al., 2004).

There are other aspects of task network model development. Some items define a simulation scenario defining continuous processes within the model and queues in front of tasks. Further details of these features can be obtained from the Micro Saint Sharp *User's Guide* (Alion Science and Technology, 2009). As a model is being developed and debugged, the user can execute the model to test it and collect data. The user can rearrange, open, and close a variety of windows to represent a variety of display modes providing differing levels of information during execution. The simulation speed can also be controlled to include pausing after every simulated task. Typically, during execution the user will display the task network on the screen, and tasks that are currently executing will be highlighted. In this mode, the analyst can get a very clear picture of what events are occurring in what sequence in the model, greatly aiding debugging. Figure 8 presents a sample display during model network animation. Additionally, 2D and 3D animator modes are available. In these modes, the user can create a graphical representation of the system. Changes on the graphical background can be tied to the task flow, providing a powerful method to communicate the model's findings to stakeholders. Once a model is executed and data are collected, the analyst has a number of alternatives for data analysis. The data created during a model execution can be reviewed within Micro Saint Sharp or can be exported to statistical and graphics packages for post processing.

As stated before, the basis for task network models of human performance is the mainstay of human engineering analysis, the task analysis. Much of the information discussed above is generally included in the task analysis. Task network modeling greatly increases
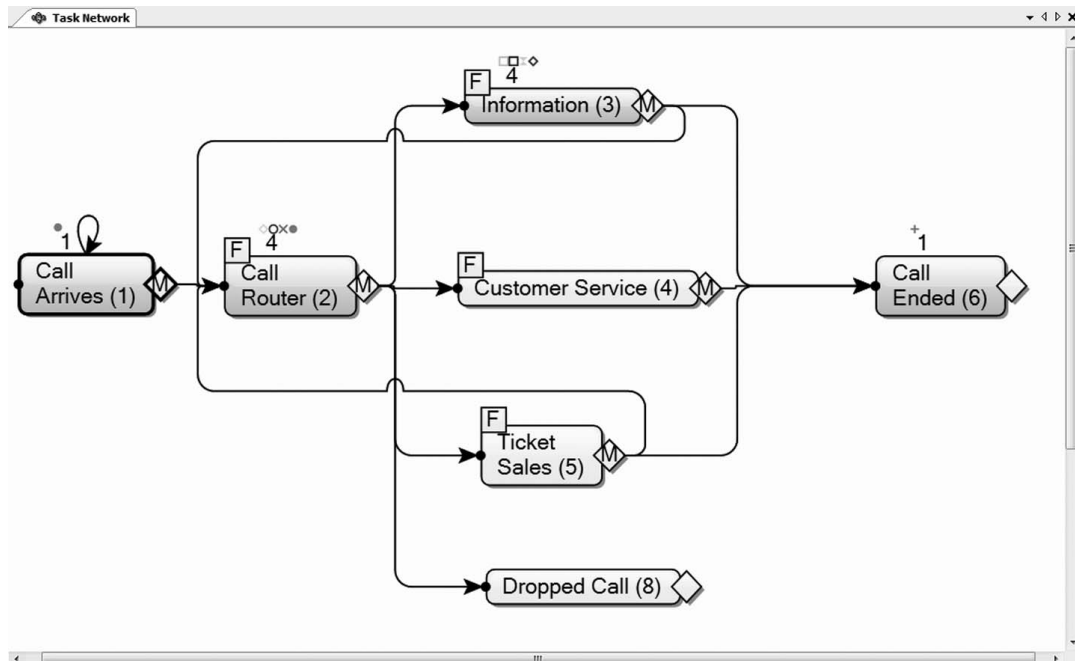


**Figure 8** Task network animation during model execution in Micro Saint Sharp.

the power of task analysis since the ability to simulate a task network with a computer permits *prediction* of human performance rather than simply the *description* of human performance that a task analysis provides. What may not be as apparent, however, is the power of task network modeling as a means of modeling human performance in *systems*. Simply by describing the system's activities in this step-by-step manner, complex models of the system can be developed where the human's interaction with the system can be represented in a closed-loop manner. The preceding discussion, in addition to being an introduction to the concepts, is also intended to support the argument that task network modeling is a mature technology ready for application in a wide range of problem domains.

### 4.2  Task Network Model of a Process Control Operator

This simple hypothetical example illustrates how many of the basic concepts of task network modeling can be applied to studying human performance in a process control environment. It is intended to illustrate many of the concepts described above. The simple human task that we want to model is of an operator responding to an annunciator. The procedure requires that the operator compare readings on two meters. Based on the relative values of these readings, the operator must either open or close a valve until the values on the two meters are nearly the same. The task network in Figure 9 represents

the operator activities for this model. Also, to allow the study of the effects of different plant dynamics (e.g., control lags), a simple one-node model of the line in which the valve is being opened is included in Figure 10.

The operator portion of the model will run the "monitor panels" task until the values of the variables "meter1" and "meter2" are different. The simulation could begin with these values being equal and then precipitate a change in values based on what is referred to as a *scenario event* (e.g., an event representing the effects of a line break on a plant state). This event could be as simple as

$$\text{meter } 1 = \text{meter } 1 + 2.0;$$

or as complex as an expression defining the change in the meter as a function of line break size, flow rates, and so on. An issue that consistently arises in model construction is how complex the plant system model should be. If the problem under study is purely operator performance, simple models will usually suffice. However, if overall plant behavior is of interest, the models of plant dynamics, such as meter values, are more important. Again, we recommend the "start simple" approach whenever possible.

When the transient occurs and the values of meter1 and meter2 start to diverge, the annunciator signal will trigger. This annunciator would be triggered in the plant
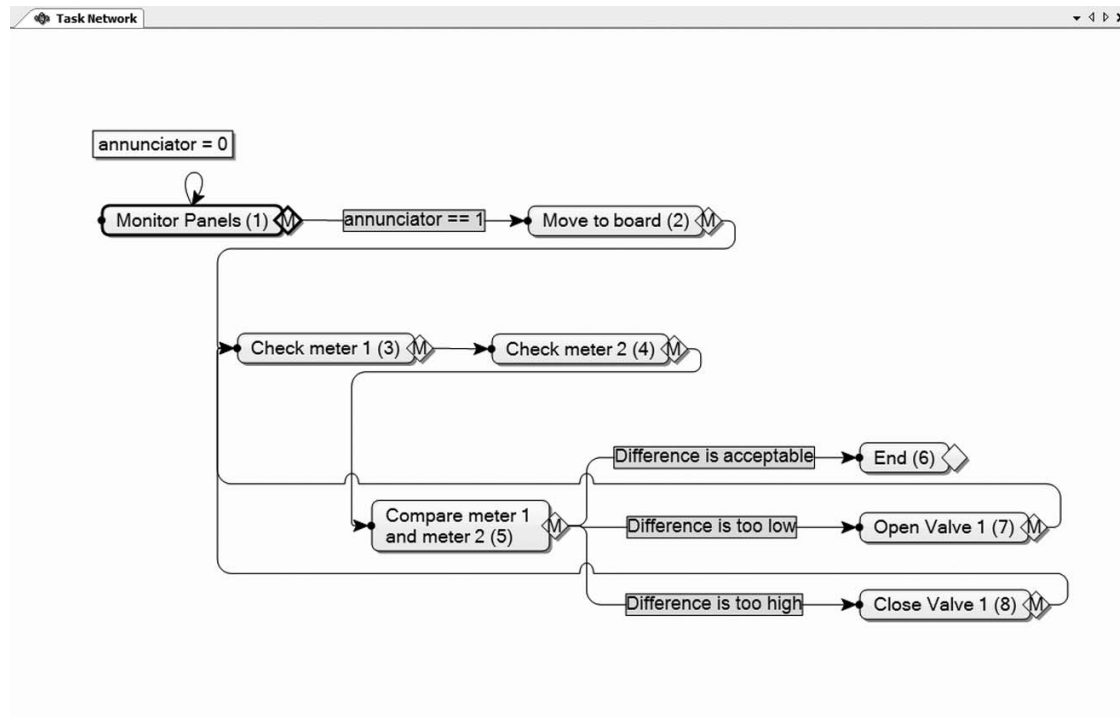


**Figure 9**　Task network model of a process control operator responding to an annunciator.
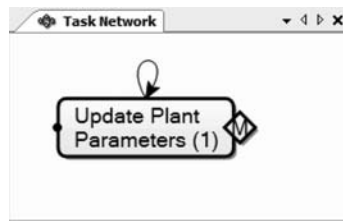
**Figure 10** Simple one-node model of the plant integrated with the detailed operator model.

portion of the model by a task-ending effect such as

if (meter 1 ! = meter 2) then annunciator = 1;

Once the plant model sets the value of the variable annunciator to 1, the operator will begin to move to the appropriate board. Then the operator will continue through a loop to check the values for meter1 and meter2 and open valve 1, close valve 1, or make no change. The determination of whether to make a control input is determined by the difference in values between the two meters. If the value is less than the acceptable threshold, the operator would open the valve further. If the value is greater than the threshold, the operator would close the valve. This opening and closing of the valve would be represented by changes in the value of the variable valve1 as a task-ending effect of the tasks open valve1 and close valve1. In this simple model, operators do not consider rates of change in values for meter1 and therefore would get into an operator-induced oscillation if there were any response lag. A more sophisticated operator model could use rates of change in the value for meter1 in deciding whether to open or close valves.

Again, this is a very small model reflecting simple operator activity on one control via a review of two displays. However, it illustrates how large models of operator teams looking at numerous controls and manipulating many displays could be built via the same building blocks used in this model. The central concepts of a task network and shared variable reflecting human–system dynamics remain the same.

Given a task network model of a process control operator in a "current" control room, how might the model be modified to address human-centered design questions? Some examples are (1) modifying task times based on changes in the time required to access a new display; (2) modifying task times and accuracies based on changes in the content and format of displays; (3) changing task sequence, eliminating tasks, and/or adding tasks based on changes in plant procedures; (4) changing allocation of tasks and ensuing task sequence based on reallocation of tasks among operators; and (5) changing task times and accuracies based on stressors such as sleep loss or the effects of circadian rhythm. This is not intended as a definitive list of all the ways that these models may be used to study design or operations concepts but should illustrate how these models can be used to address design and operational issues.

### 4.3 Use of Task Network Modeling to Address Specific Design Concerns

In this section we examine two case studies in the use of task network simulation for studying human performance issues. The first case study explores how task network modeling can be used to assess task allocation issues in a cognitively demanding environment. The second example explores how task network modeling has been used to extend laboratory and field research on human performance under stress to new task environments. We should state clearly that these examples are intended to be representative of the types of issues that task network modeling can address as well as approaches to modeling human performance with respect to these issues. *They are not intended to be comprehensive with respect to either the issues that might be addressed or the possible techniques that the human factors practitioner might apply.* Simulation modeling is a technology whose application leaves much room for creativity on the part of the human factors practitioner with respect to application areas and methods. These two case studies are representative.

#### 4.3.1 Crew Workload Evaluation

Perhaps the greatest contributor to human error in many systems is the extensive workload placed on the human operator. The inability of the operator to cope effectively with all of his or her information and responsibilities contributes to many accidents and inefficiencies. In recognition of this problem, new automation technologies have been introduced to reduce workload during periods of high stress. Some of these technologies are in the form of enhanced controls and displays, some are in the form of tools that "push" information to the operator and alert the operator in order to focus attention, and still others consist of adaptive tools that "take over" tasks when they sense that the operator is overloaded. Unfortunately, these technical solutions often introduce new tasks to be performed that affect the visual, auditory, and/or psychomotor workload of the operators.

Recently, new concepts in crew coordination have focused on better management of human workload. This area shows tremendous promise and is benefiting from efforts of human factors researchers. However, their efforts are hindered because there are limited opportunities to examine empirically the performance of different combinations of equipment and crew composition in a realistic scenario or context. Additionally, high workload is not typically caused by a single task but by situations in which multiple tasks must be performed and managed simultaneously. It is not simply that the quantity of tasks can lead to overload, but it also depends on the composition of those tasks. For example, two cognitive tasks being performed in parallel are much more effortful than a simple motor task and an oral communication task being performed together. The occurrence of these situations will not typically be discovered through normal human engineering task analysis or subjective workload analysis until there is a system to be tested. That is often too late to influence design. To rectify this problem, there has been a significant amount of

recent research and development aimed at human work-load *prediction* models. Predictive models allow the designers of a system to estimate operator workload *without human subject experimentation*. From this and other research, a solid theoretical basis for human work-load prediction has evolved, as is described in Wickens (1984).

In this section we discuss a study using task network modeling to predict the impact of task allocation on human workload. Although these examples are posed in the context of the design of a military system, the same techniques have been used in nonmilitary applications such as process control and user–computer interface design.

### 4.3.2 Modeling the Workload of a Future Command and Control Process

The Army command and control (C2) community is concerned with how new information technology and organizational changes projected for tomorrow's battle-field will affect soldier tasks and workload. To address this concern, an effort was undertaken to model soldier performance under current and future operational condi-tions. In this way, the impact of performance differences could be quantitatively assessed so that equipment and doctrine design could be influenced in a timely and effective manner.

In one C2 project, the primary concern was to deter-mine how tasks should be allocated and automated such that a C2 team could evaluate all the relevant data and make decisions within an environment with particularly high time pressure. Specifically, the effort was to address the following key questions:

- How many crew members do you need?
- How do you divide tasks among jobs?

- How does decision authority flow?
- Can the crew meet decision timeline require-ments?
- Is needed information usable and accessible?

Task network modeling was used to study crew member, task, and scenario combinations in order to examine these questions. Figure 11 shows the top-level diagram of the task network. Essentially, the crew members receive and monitor information about the system and the environment until an event occurs that pushes them out of the 10,000 and 20,000 networks into either a series of planning tasks or a series of evaluation, decision, direction, and execution tasks. The purpose of the planning task is to update tactical battle plans based on new information received from the system or the environment. Receipt of new intelligence data about the enemy's intention or capability is an example of an event that would cause crew members to undertake planning tasks. Similarly, receipt of information from the system about resource limitations might trigger the crew members to proceed down the alternative path (through evaluate to execute). Specifically, limited resources might cause crew members to evaluate whether the engagement is proceeding appropriately (30,000), decide how to adjust system parameters (40,000), direct the appropriate response to the correct level of command (50,000), and then execute the order (60,000). Upon completion, crew members would return to monitoring the system and situation.

Each rectangle in the task network shown in Figure 11 actually consists of a network of tasks. An example of the tasks that belong to network 10000 are shown in Figure 12. As described earlier, the tasks in network 10000 are linked by probabilistic and tactical decisions. Each of the tasks in the C2 task network
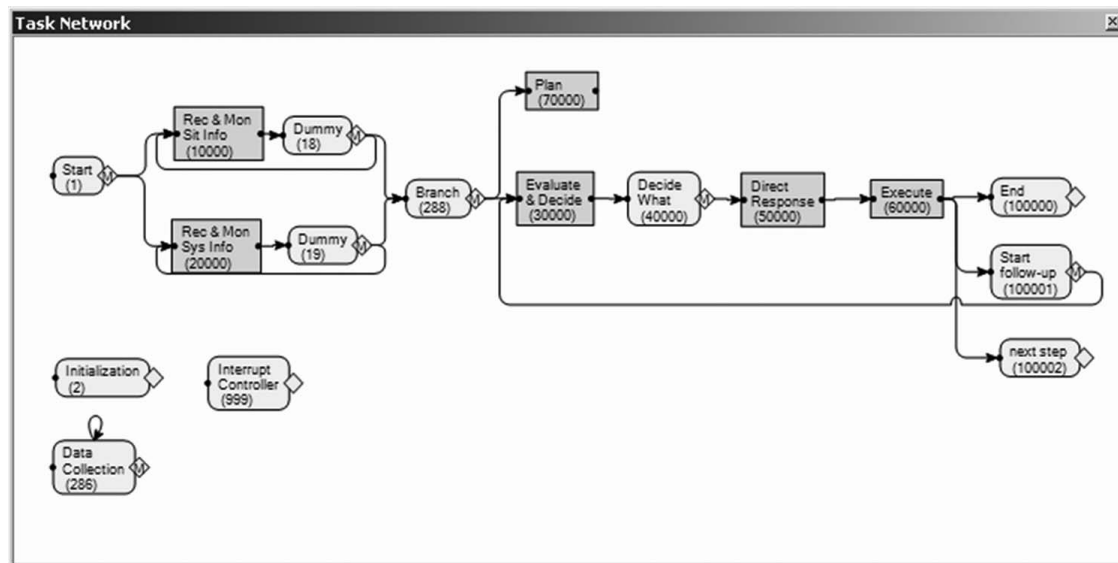


**Figure 11**  Upper level task network.
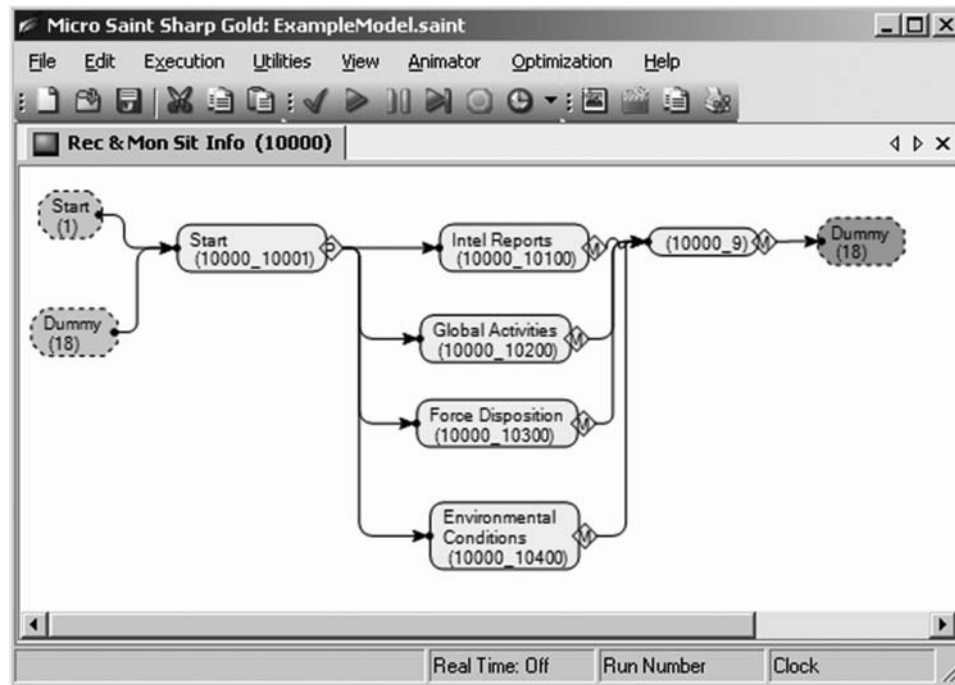
**Figure 12** Second-level task network.

is associated with several items of human performance data:

- *Task Performance Time.* These data consist of a mean, standard deviation, and distribution. The data were collected from a combination of three sources: (1) human factors literature (e.g., Fitt's law), (2) empirical studies during operator-in-the-loop simulator exercises, and (3) subject matter experts.
- *Branching Logic.* Although the task network indicates a general process flow, this particular model was designed to respond to scenario events. Because of that design decision, each task includes logic to determine the following task. For example, if the scenario is very intense and multiple target tracks are available, crew members would follow a different task flow than if they were performing routine system checks.
- *Release Rules.* Logic controlling the number and types of parallel tasks each crew member can perform is contained in each task's release condition.

Since one purpose of the model was to examine various task allocation strategies, the model was designed to incorporate several measures of crew member workload. The basis of this technique is an assumption that excessive human workload is not usually caused by one particular task required of the operator. Rather, the human

having to perform several tasks simultaneously leads to overload. Since the factors that cause this type of workload are intricately linked to these dynamic aspects of the human's task requirements, task network modeling provides a good basis for studying how task allocation and sequencing can affect operator workload.

However, task network modeling is not inherently a model of human workload. The only relevant output common to all task network models is the time required to perform a set of tasks and the sequence in which the tasks are performed. Time information alone would suffice for some workload evaluation techniques, such as Siegel and Wolf (1969), whereby workload is estimated by comparing the time available to perform a group of tasks to the time required to perform the tasks. Time available is driven by system performance needs, and time required can be computed with a task network model. However, it has long been recognized that this simplistic analysis misses many aspects of the human's tasks that influence both perceived workload and ensuing performance. At the very least, this approach misses the fact that some pairs of tasks can be performed in combinations better than other pairs of tasks.

One of the most promising theories of operator workload, which is consistent with task network modeling, is the multiple-resource theory proposed by Wickens (see Wickens et al., 1983). Simply stated, the multiple-resource theory suggests that humans have several different resources that can be tapped simultaneously and with varying levels of inter-resource conflict and competition. Depending on the nature of the information

processing tasks required of a human, these resources would have to process information sequentially (if different tasks require the same types of resources) or possibly in parallel (if different tasks required different types of resources). There are many versions of this multiple-resource theory in the workload literature (e.g., McCracken and Aldrich, 1984; Archer and Adkins, 1999). In this chapter we provide a discussion of the underlying methodology of the basic theory.

Multiple-resource workload theory is implemented in a task model in a fairly straightforward manner. First, each task in the task network is characterized by the workload demand required in each human resource, often referred to as a *workload channel*. Examples of commonly used channels include auditory, visual, cognitive, and psychomotor. Particular implementations of the theory vary in the channels that are included and the fidelity with which each channel is measured (high, medium, low vs. seven-point scale). As an example, the scale for visual demand is presented in Figure 13.

Similar scales have been developed for the auditory, cognitive, fine-motor, gross-motor, speech, and tactile channels. Using this approach, each operator task can be characterized as requiring some amount of each of the seven types of resources, as represented by a value (typically between 1 and 7). All operator tasks can be analyzed with respect to these demand values. In performing a set of tasks pursuant to a common goal (e.g., engage an enemy target), crew members frequently must perform several tasks simultaneously, or at least nearly so. For example, they may be required to monitor a communication network while visually searching a display for target track. Given this, the workload literature indicates that the crew member may either accept

the increased workload (with some risk of performance degrading) or begin dumping tasks perceived as less important. To factor these two issues into task network simulations, two approaches can be incorporated: (1) evaluate combined operator workload demands for tasks that are being performed concurrently and/or (2) determine when the operator would begin dumping tasks due to overload.

During a task network simulation, the model of the crew may indicate that they are required to perform several tasks simultaneously. The task network model evaluates total attentional demands for each human resource (e.g., visual, auditory, fine motor, gross motor, speech, tactile, and cognitive) by combining the attentional demands across all tasks that are being performed simultaneously. Intra- and interresource conflict values are then computed that indicate how much the different resources compete with each other. The conflict score is then used to increase the total attentional demand. This combination leads to an overall workload demand score for each crew member.

To implement this approach in Micro Saint Sharp, the task-beginning effect can be used to increment variables that represent the current workload score in each resource. Then, while the tasks are being performed, these variables track attentional demands. When the tasks are completed, the task-ending effects can decrement the values of these variables accordingly. Therefore, if these workload variables were recorded and then plotted as the model runs, the output would look something like as shown in Figure 14. This result can be used to identify points of high workload throughout the scenario being modeled. The human factors practitioner can then review the tasks that led to the points of



**Figure 13**  Visual workload scale.

high workload and determine whether they should be reallocated or redesigned in order to alleviate the peak. This is a common approach to modeling workload.

Once the task networks were verified with knowledgeable crew members, they became part of the human factors team's analytical test bed. Figure 15 shows the overall method that can be used to examine aspects of crew member performance across a wide variety of operational scenarios and crew configuration concepts. The center of this diagram, labeled the task network, represents the tasks that the crew performs. The network itself, representing the flow of the tasks, does not change between model runs. Rather, the model has been parameterized so that an event scenario stimulates the network. The left side of the diagram illustrates the types of data that are used to drive the task network model. In this case, those data include crew configurations, or allocations of tasks to different crew members and automation devices, as well as scenario events. The scenario events represent an externally generated time-ordered list of the events that trigger the crew members to perform tasks

in the task network. The right side of Figure 15 represents the types of outputs that can be produced from this task network model. One of the primary outputs is a crew member workload graph, such as that shown in Figure 14. Another is operator utilization, as shown in Figure 16.

### 4.3.3 Extensions to Other Environments

The workload analysis methodology described above has been developed into a stand-alone task network modeling tool by the Army Research Laboratory (ARL) Human Research and Engineering Directorate (HRED) as part of the IMPRINT (Archer and Adkins, 1999). IMPRINT integrates task network modeling software with features that specifically support the multiple-resource theory of workload discussed above. It provides the human factors practitioner with an environment that supports the analysis of task assignment to crew members based on four factors:
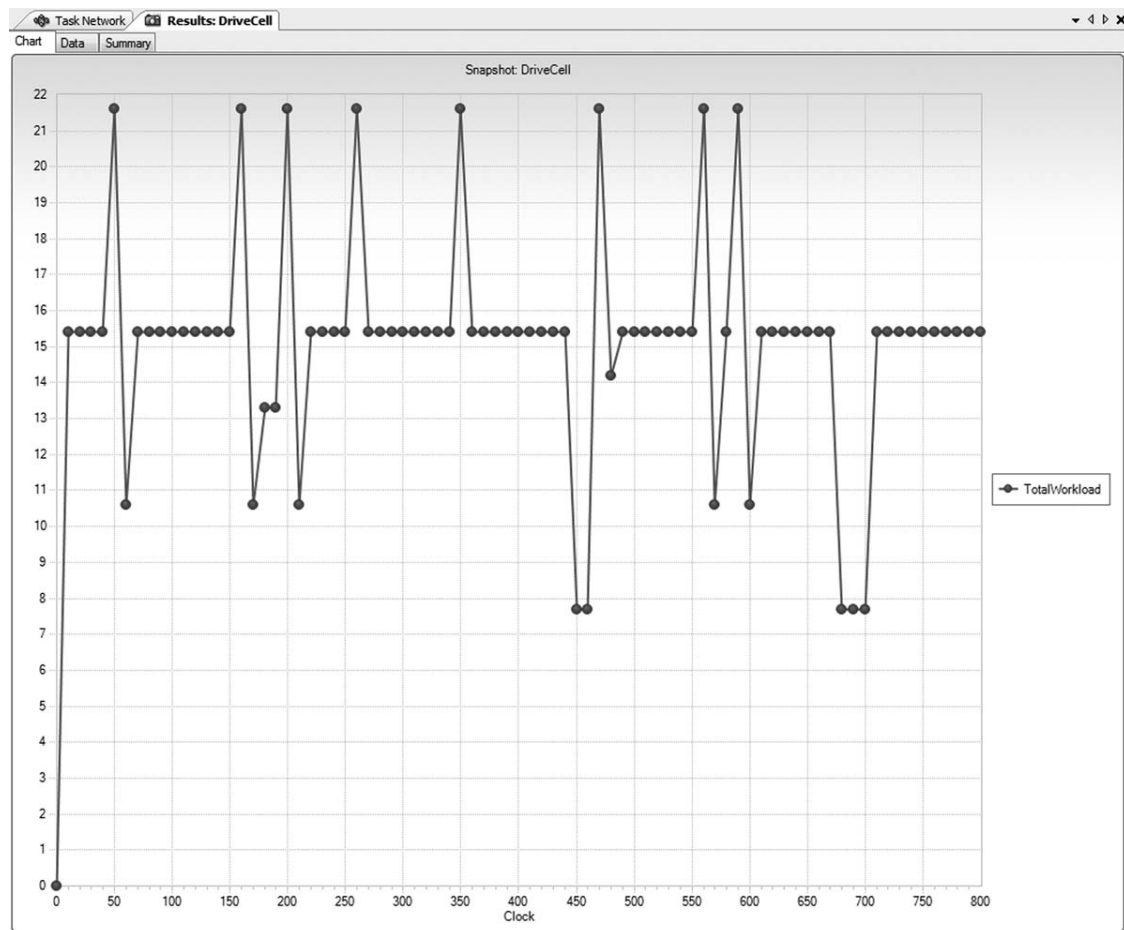


Figure 14 Workload output from a task network model.
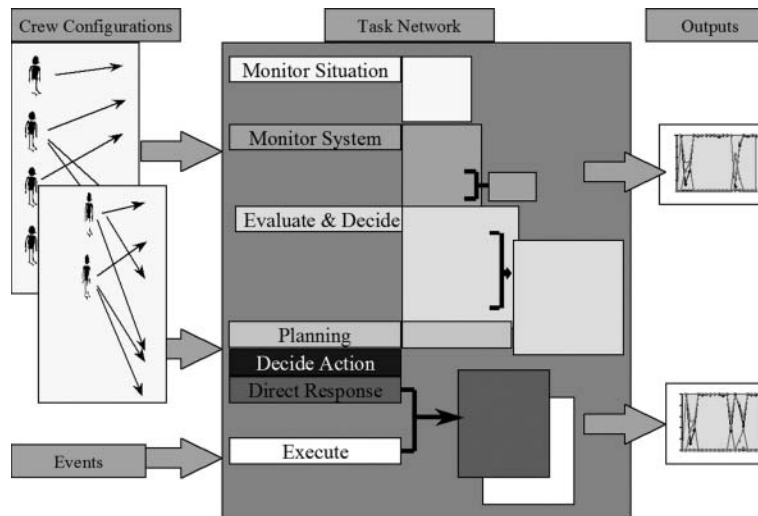
**Figure 15**   Overall method for examining workload in a complex system.
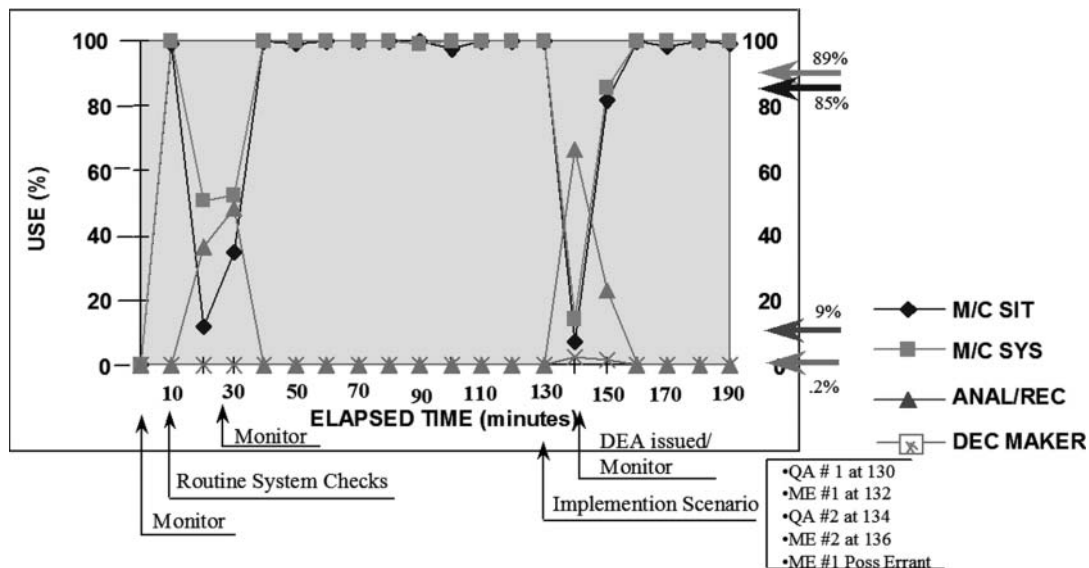


**Figure 16**   Percent utilization taken over 10-min intervals.

1.  *Workload of Crew Members.* Tasks should be assigned to minimize the amount of time that crew members will spend in situations of excessive workload.

2.  *Time Performance Requirements.* Tasks must be assigned and sequenced so that they are completed within the available time. This consideration is essential since time constraints often will drive the need to perform several tasks simultaneously.

3.  *Likelihood of Successful Performance and Consequences of Failure.* Tasks must be assigned and sequenced so that they can be completed within a specified accuracy measure.

4.  *Access to Controls and Displays.* Tasks cannot be assigned to crew members that do not have access to the necessary controls and displays.

Of course, there are numerous theoretical questions regarding this simplistic approach to assessing workload in an operational environment. However, even the use of this simple approach has been shown to provide useful

insight during design. For example, in a study conducted by the Army (Allender, 1995), a three-man crew design was evaluated using a task network model. The three-man model was constructed using data from a prototype four-man system. From this model-based analysis, the three-man design was found to be unworkable. Later, experimentation using human subjects verified that the model's workload predictions were sufficiently accurate to point the design team in a valid direction.

IMPRINT also includes built-in constructs for simulating *workload management* strategies that operators would employ to accommodate points of high operator workload (Plott, 1995). The ultimate result of simulating the workload management strategies is that the operator task network being modeled is *dynamic*. In other words, the task sequence, operator assignments, and individual task performance may change in response to excessive operator workload as the task network model executes. These changes may be as simple as one operator handing tasks off to another operator to reduce workload to an acceptable level or as complex as the operator beginning to time share tasks in order to complete all the tasks assigned, potentially with associated task performance penalties. Ultimately, the tool provides an estimate of system-level performance as a result of these realistic workload management strategies. This innovation in modeling provides greater fidelity in efforts that model human behavior in the context of system performance, particularly in high-workload environments such as complex system control and management.

### 4.3.4 Extending Research Findings to New Task Environments

Task network modeling was used by LaVine et al. (1995) to extend laboratory data and field data collected on one set of human tasks to predicting performance on similar tasks. The problem of extending laboratory or field human performance data to other tasks has plagued the human engineering community for years. We know intuitively that human performance data can be used to predict performance for similar tasks. However, it is often the case that the task whose performance we want to predict is similar in some ways but different in others. The approach described below uses a skill taxonomy to quantify task similarity and therefore provides a means for determining how other tasks will be affected when exposed to a common stressor on human performance. Once functional relationships are defined between a skill type and a stressor, task network modeling is used to determine the effect of the stressor on performance of a complex task that uses many of these skills simultaneously.

The specific approach below is being used by the U.S. Army to predict crew performance degradation as a function of a variety of stressors. It is not intended to represent a universally acceptable taxonomy for simulating human response to stress. The selection of the best taxonomy would depend on the particular tasks and stressors being studied. What this example is intended to illustrate is another way that task network modeling can be used to predict human performance by *making a series of reasonable assumptions* that

can be played together in a model for the purpose of making predictions that would be impossible to make otherwise. The methodology for predicting human performance degradation as a function of stressors consists of three parts: (1) a *taxonomy* for classifying tasks according to basic human skills, (2) *degradation functions* for each skill type for each stressor, and (3) *task network models* for the human-based system whose performance is being predicted. Conceptually, either laboratory or field data can be used to develop links between a human performance stressor (e.g., heat, fatigue) and basic human skills. By selecting a skills taxonomy that is sufficiently discriminating to make this assumption reasonable, one can assume that the effects of the stressor on all tasks involving the skill will be approximately the same. The links between the level of a stressor (e.g., fatigue) and resulting skill performance (e.g., the expected task time increase from fatigue) are defined mathematically as the degradation function. The task network model is the means for linking these back to complex human–system performance.

***Taxonomy*** The basic premise behind the taxonomy is that the tasks that humans perform can be broken down into basic human skills or atomic tasks (Roth, 1992). The taxonomy that was used by Roth consists of five skill types described by Roth as follows:

1. *Attention:* the ability to attend actively to a stimulus complex for extended periods of time in order to detect specified changes or classes of changes that indicate the occurrence of some phenomenon that is critical to task performance
2. *Perception:* the ability to detect and categorize specific stimulus patterns embedded in a stimulus complex
3. *Psychomotor skill:* the ability to maintain one or more characteristics of a situation within a set of defined conditions over a period of time, either by direct manipulation or by manipulating controls that cause changes in the characteristics
4. *Physical skill:* the ability to accomplish sustained, effortful muscular work
5. *Cognitive skill:* the ability to apply concepts and rules to information from the environment and from memory in order to select or generate a course of action or a plan (includes communicating the course of action or plan to others)

These five skills covered most of the tasks that were of interest to the Army for this study and still provided a manageable number of categories for an analyst to use.

***Degradation Functions*** The degradation functions quantitatively link skill performance to the level of a stressor. The degradation functions can be developed from any data source, including standard test batteries or actual human tasks. Through statistical analysis, one can build skill degradation functions for each taxon. These functions map the performance decrement expected on

**Performance multipliers**
as a function of time since sleep
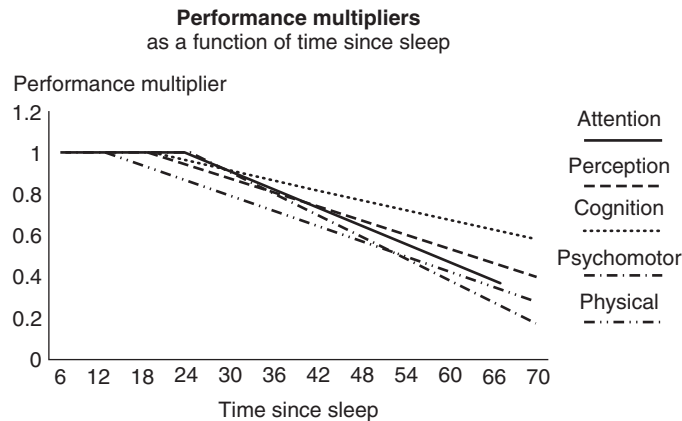


**Figure 17** Performance degradation functions associated with each of the human skills from the taxonomy.

a skill based on the parameters of the performance-shaping factor (e.g., time since sleep). An example of these functions is presented in Figure 17.

***Incorporating the Degradation Functions into Task Network Models to Predict Overall Human–System Performance Degradation*** The key to making this approach useful to predicting complex human performance is the task network model of the new task. In the task network model of the human's activities, all tasks are defined with respect to the percentage of each skill required from the taxonomy. For example, the following are ratings for tasks faced by a console operator responding to telephone contacts:

| Detect ring | 50% attention, 50% perception |
| Select menu item using a mouse | 40% attention, 60% psychomotor |
| Interpret customer's request for information | 100% cognitive |

In building the task network model, mathematical expressions can be developed that degrade a specific task's performance through an arithmetic weighting of skill degradation multipliers that are derived from the degradation functions. For example, if the fatigue parameter was "time since sleep" and the value of that parameter was "36 hours since sleep," the task time performance multipliers would be as follows in the example above:

| Attention performance multiplier | 0.82 |
| Perception performance multiplier | 0.808 |
| Cognition performance multiplier | 0.856 |
| Psychomotor performance multiplier | 0.784 |
| Physical performance multiplier | 0.727 |

Based on these multipliers and the task weightings above, the specific task effects would be:

● Detect ring (50% attention, 50% perception)

$$\text{Task multiplier} = 0.5 \times 0.82 + 0.5 \times 0.808$$
$$= 0.814$$

● Select menu item using a mouse (40% attention, 60% psychomotor)

$$\text{Task multiplier} = 0.4 \times 0.82 + 0.6 \times 0.784$$
$$= 0.7984$$

● Interpret customer's request for information (100% cognitive)

$$\text{Task multiplier} = 0.856$$

In a model of the complex tasks being examined by LaVine et al. (1995), the task networks consisted of several dozen or even several hundred tasks. Through the approach described above, each task in a model exhibited a unique response to a stressor depending on the particular skills that it required. The task network model then provided the means for relating the individual task performance to overall human–system performance as a function of stressor level (e.g., the time to perform a complex series of tasks involving decision making and error correction). Through this type of analysis, LaVine et al. were able to develop curves such as that shown in Figure 18 relating human performance to a stressor. These relationships would have been virtually impossible to develop experimentally.

Again, there were a number of simplifying assumptions that were made in this research. However, by being willing to accept these assumptions, LaVine et al. were able to characterize how complex human–system performance would be affected by a variety of stressors over a wide range in a relatively short time. As such, they were able to estimate the effects of stressors that would have otherwise been pure guesswork.
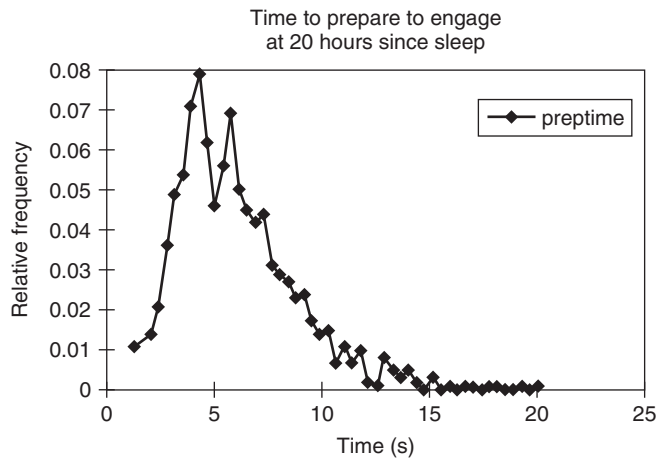
**Figure 18** Frequency distribution of expected human performance as a function of time since sleep that was derived using task network modeling.

### 4.3.5 Incorporation of Advanced Attention Model

In recent years advanced theories of selective attention have been advanced to the point that they can be embedded into human performance models to more accurately predict how a human might perform in a visually complex environment. As the fidelity of human performance models improve, so too does the need for integrating complex human attention models in order to predict how operators will allocate their visual attention. Attention models that have shown preliminary success when being linked to human performance models are the SEEV (salience, effort, expectancy, and value) model (Horrey et al., 2006) and the newer N-SEEV (noticing-SEEV) model (Wickens and McCarley, 2008). SEEV is a computational, plausible model that accounts for how four quantifiable elements do and/or should drive operator's attention around a complex working environment. SEEV models the operator's visual scanning pattern attending to a given "area of interest" (AOI) that supports the tasks that need to be performed. The SEEV and N-SEEV algorithms are described in detail in Chapter 5.

The Man–Machine Integration Design and Analysis System (MIDAS) has recently been updated to incorporate the SEEV algorithm (Gore et al., 2009). The integration of the SEEV model into MIDAS allows dynamic scanning behaviors by calculating the probability that the operator's eye will move to a particular AOI given the tasks the operator is engaged in within the multitask context. It also better addresses allocation of attention in dynamic environments such as flight and driving tasks. In MIDAS, effort, expectancy, and value are assigned values between 0 and 1, while salience is left unconstrained. Effort, expectancy, and value drive the human operator's visual attention around the displays. However, if a salient event occurs, then $P(\text{AOI})$ may be offset by the display exhibiting the salient event until the display location of the salient event has been

fixated and detected. The improved predictive capability of information-seeking behavior that resulted from the implementation of the validated SEEV model leaves MIDAS better suited to predict performance in complex human–machine systems.

### 4.3.6 Summary

Once again, the above are intended to serve as examples, not a catalog of problems or approaches that are appropriate for task network modeling. Task network modeling is an approach to extend task and systems analysis to make predictions of human–system performance. The creative human factors and ergonomics practitioner will find many other useful applications and approaches.

## 5 FIRST-PRINCIPLE APPROACH: ADAPTIVE CONTROL OF THOUGHT–RATIONAL COGNITIVE ARCHITECTURE

The other fundamental approach to modeling human performance is based on the mechanisms that underlie and cause human behavior. Since this approach is based on fundamental principles of the human and his or her interaction with the system and environment, we have designated them as *first-principle models*. By integrating these models with models of the system and environment, the human factors specialist can predict the full behavior of large-scale interactive human–machine systems. The ACT-R cognitive architecture (Anderson and Lebiere, 1998) is a production system theory that models the steps of cognition by a sequence of production rules that fire to coordinate retrieval of information from the environment and from memory. It is a cognitive architecture that can be used to model a wide range of human cognition. It has been used to model tasks from memory retrieval (Anderson et al., 1998) to visual search (Anderson et al., 1997). The range of models developed, from those purely concerned with internal

cognition to those focused on perception and action, makes ACT-R a plausible candidate to model complex tasks involving the interaction of one (or more) human operator with complex systems with the goal of evaluating the design of those systems. In all domains, ACT-R is distinguished by the detail and fidelity with which it models human cognition. It makes claims about what occurs cognitively every few hundred milliseconds in performance of a task. ACT-R is situated at a level of aggregation above those of basic brain processes (targeted by other modeling approaches, such as neural networks) but considerably below such complex tasks as air traffic control. The new version of the theory has been designed to be more relevant to tasks that require deploying significant bodies of knowledge under conditions of time pressure and high information-processing demand. This is because of the increased concern with the temporal structure of cognition and with the coordination of perception, cognition, and action.

## 5.1  ACT-R

ACT-R is a unified architecture of cognition developed over the last 30 years at Carnegie Mellon University. At a fine-grained scale it has accounted for hundreds of phenomena from the cognitive psychology and human factors literature. The most recent version, ACT-R 6.0 (Anderson et al., 2007), is a modular architecture composed of interacting modules for declarative memory, perceptual systems such as vision and audition modules, and motor systems such as manual and speech modules, all synchronized through a central production system (see Figure 19). This modular view of cognition is a reflection both of functional constraints and of recent advances in neuroscience concerning the localization of brain functions. ACT-R is also a hybrid system that combines a tractable symbolic level that enables the easy specification of complex cognitive functions with a subsymbolic level that tunes itself to the statistical structure of the environment to provide the graded characteristics of cognition such as adaptivity, robustness, and stochasticity.

The central part of the architecture is the production module. A production can match the contents of any combination of buffers, including the goal buffer, which holds the current context and intentions; the retrieval buffer, which holds the most recent chunk retrieved from declarative memory; the visual and auditory buffers, which hold the current sensory information; and the manual and vocal buffers, which hold the current state of the motor and speech module. The highest rated matching production is selected to effect a change in one or more buffers, which in turn triggers an action in the corresponding module(s). This can be an external action (e.g., movement) or an internal action (e.g., requesting information from memory). Retrieval from memory is initiated by a production specifying a pattern for matching in declarative memory. Each chunk competes for retrieval, with the most active chunk being selected and returned in the retrieval buffer. The activation of a chunk is a function of its past frequency and recency of use, the degree to which it matches the pattern requested, plus stochastic noise. Those factors confer memory retrievals, and behavior in general, desirable "soft" properties such as adaptivity to changing circumstances, generalization to similar situations, and variability (Anderson and Lebiere, 1998).

The current goal is a central concept in ACT-R, which as a result provides strong support for goal-directed behavior. However, the most recent version of the architecture is less goal focused than its predecessors
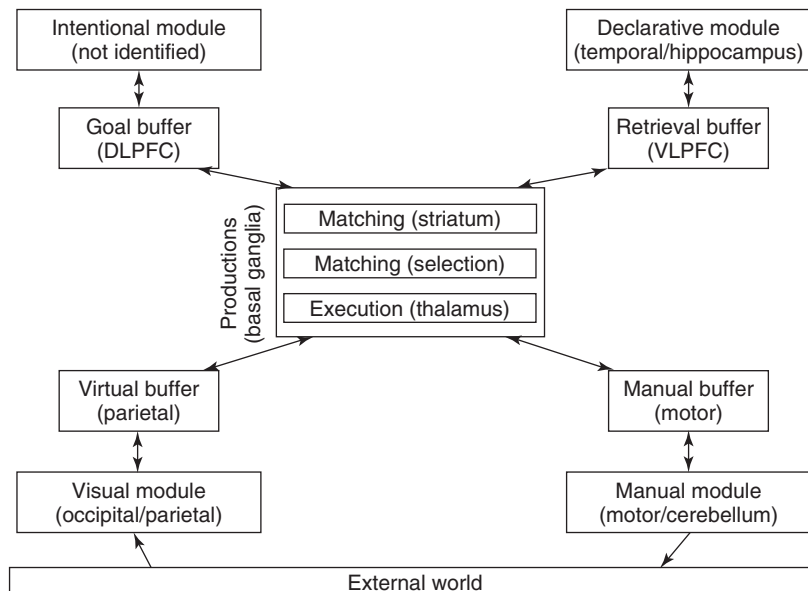


**Figure 19**  Modular view of ACT-R cognitive architecture.

by allowing productions to match to any source of information, including the current goal, information retrieved from declarative memory, objects in the focus of attention of the perceptual modules, and the state-of-the-action modules. The content of many of those buffers, especially the perceptual buffers, might have changed not as a function of an internal request but as a result of an external event happening, perhaps unexpectedly, in the outside world. This emphasis on asynchronous pattern matching of a wide variety of information sources better enables ACT-R to operate and react efficiently in a dynamic fast-changing world through flexible goal-directed behavior which gives equal weight to internal and external sources of information.

There are three main distinctions in the ACT-R architecture. First, there is the procedural–declarative distinction that specifies two types of knowledge structures: chunks for representing declarative knowledge and productions for representing procedural knowledge. Second, there is the symbolic level, which contains the declarative and procedural knowledge, and the subsymbolic level of neural activation processes that determine the speed and success of access to chunks and productions. Finally, there is a distinction between the performance processes by which the symbolic and subsymbolic layers map onto behavior and the learning processes by which these layers change with experience.

Human cognition can be characterized as having two principal components: (1) the knowledge and procedures codified through specific training within the domain and (2) the natural cognitive abilities that manifest themselves in tasks as diverse as memory, reasoning, planning, and learning. The fundamental advantage of an integrated architecture like ACT-R is that it provides a framework for modeling basic human cognition and integrating it with specific symbolic domain knowledge of the type specified by domain experts (e.g., rules specifying what to do in a given condition, a type of knowledge particularly well suited for representation as production rules). However, performance described by symbolic knowledge is mediated by parameters at the subsymbolic level that determine the availability and applicability of symbolic knowledge. Those parameters underlie ACT-R's theory of memory, providing effects such as decay, priming, and strengthening and make cognition adaptive, stochastic, and approximate, capable of generalization to new situations and robustness in the face of uncertainty. They also can account for the limitations of human performance, such as latencies to perform tasks and errors that can originate from a number of sources. Finally, they provide a basis for representing individual differences such as those in working memory capacity, attentional focus, motivation, and psychomotor speed as well as the impact of external behavior moderators such as fatigue (Lovett et al., 1999; Taatgen, 2001; Gunzelmann et al., 2009) through continuous variations of those subsymbolic architectural parameters that affect performance in complex tasks.

Because they influence quantitative predictions of performance so fundamentally, we describe in some more detail the subsymbolic level in which continuously varying quantities are processed, often in parallel, to produce much of the qualitative structure of human cognition. These subsymbolic quantities participate in neural-like activation processes that determine the speed and success of access to chunks in declarative memory as well as the conflict resolution among production rules. ACT-R also has a set of learning processes that can modify these subsymbolic quantities. Formally, activation reflects the log posterior odds that a chunk is relevant in a particular situation. The *activation* $A_i$ of a declarative chunk $i$ is computed as the sum of its base-level activation $B_i$ plus its context activation:

$$A_i = B_i + \sum_j W_j S_{ji}$$

In determining the context activation, $W_j$ designates the attentional weight given the focus element $j$. An element $j$ is in the focus, or in context, if it is part of the current goal chunk (i.e., the value of one of the goal chunk's slots); $S_{ji}$ stands for the strength of association from element $j$ to chunk $i$. ACT-R assumes that there is a limited capacity of source activation and that each goal element emits an equal amount of activation. Source activation capacity is typically assumed to be 1 (i.e., if there are $n$ source elements in the current focus each receives a source activation of $1/n$). The associative strength $S_{ji}$ between an activation source $j$ and a chunk $i$ is a measure of how often $i$ was needed (i.e., retrieved in a production) when chunk $j$ was in the context. Associative strengths provide an estimate of the log likelihood ratio measure of how much the presence of a cue $j$ in a goal slot increases the probability that a particular chunk $i$ is needed for retrieval to instantiate a production. The *base-level activation* of a chunk is learned by an architectural mechanism to reflect the past history of use of a chunk $i$:

$$B_i = \ln \sum_{j=1}^{n} t_j^{-d} \approx \ln \frac{nL^{-d}}{1-d}$$

where $n$ stands for the number of references to chunk $i$, $t_j$ stands for the time elapsed since the $j$ reference to chunk $i$, $d$ is the memory decay rate, and $L$ denotes the lifetime of a chunk (i.e., the time since its creation). As Anderson and Schooler (1991) have shown, this equation produces the power law of forgetting (Rubin and Wenzel, 1990) as well as the power law of learning (Newell and Rosenbloom, 1981). When retrieving a chunk to instantiate a production, ACT-R selects the chunk with the highest activation $A_i$. However, some stochasticity is introduced in the system by adding Gaussian noise of mean zero and standard deviation $\sigma$ to the activation $A_i$ of each chunk. In order to be retrieved, the activation of a chunk needs to reach a fixed retrieval threshold $\tau$ that limits the accessibility of declarative elements. If the Gaussian noise is approximated with a sigmoid distribution, the *probability $P$ of chunk $i$ to be retrieved* by a production is

$$P = \frac{1}{1 + e^{-(A_i - \tau)/s}}$$

where s $= \sqrt{3}\sigma/\pi$. The activation of a chunk $i$ is related directly to the latency of its retrieval by a production $p$. Formally, *retrieval time* $T_{ip}$ is an exponentially decreasing function of the chunk's activation $A_i$:

$$T_{ip} = Fe^{-fA_i}$$

where $F$ is a time scaling factor. In addition to the latencies for chunk retrieval as given by the retrieval time equation, the total time of selecting and applying a production is determined by executing the actions of a production's action part, whereby a value of 50 ms is typically assumed for elementary internal actions. External actions, such as pressing a key, usually have a longer latency determined by the ACT-R/PM perceptual–motor module (Byrne and Anderson, 2001). In summary, subsymbolic activation processes in ACT-R make a chunk active to the degree that past experience and the present context (as given by the current goal) indicate that it is useful at this particular moment.

Just as subsymbolic activation processes control which chunk is retrieved from declarative memory, the process of selecting which production to fire at each cycle, known as conflict resolution, is also determined by subsymbolic quantities called utility that are associated with each production. The utility of a production is defined as

$$U_i(n) = U_i(n-1) + \alpha[R_i(n) - U_i(n-1)]$$

where $U_i(n)$ is the utility of a production $i$ after its $n$th application, $U_i(n-1)$ is its utility after its $(n-1)$st application, $R_i(n)$ is the reward the production receives, and $\alpha$ is the learning rate. Just as for retrieval, *conflict resolution* is a stochastic process through the injection of noise in each production's utility, leading to a probability of selecting a production $i$ given by

$$Probability(i) = \frac{e^{U_i/\sqrt{2s}}}{\sum_j e^{U_j/\sqrt{2s}}}$$

where the summation is over all the productions which are currently able to fire. The production with the highest utility (after noise is added) will be the one chosen to fire. Similar computations are at work in other modules, such as the perceptual–motor modules. Especially important are the parameters controlling the time course of processing as one attempts to execute a complex action or as one shifts visual attention to encode a new stimulus (Byrne and Anderson, 2001). Recent work has simplified the process of specifying the sequence of steps for complex actions by allowing ACT-R modelers to demonstrate actions on interfaces (John et al., 2004; Matessa and Mui, 2009). ACT-R can not only predict direct quantitative measures of performance such as latency and probability of errors but, from the same mechanistic basis, can also arise more global, indirect measures of performance, such as cognitive workload. Although ACT-R has traditionally

shied away from such meta-awareness measures and concentrated on matching directly measurable data such as external actions, response times, and eye movements, it is by no means incapable of doing so. For the purpose of the task described below, Lebiere (2001) proposed a measure of cognitive workload in ACT-R grounded in the central concept of unit task (Card et al., 1983). *Workload* is defined as the ratio of time spent in critical unit tasks to the total time spent on the task. *Critical unit tasks* are defined as tasks that involve actions, such as a goal to respond to a request for action with a number of mouse clicks, or tasks that involve some type of pressure, such as a goal to scan a display result from the detection of an event onset. The ratio is scaled to fit the particular measurement scale used in the self-assessment report. Lebiere (2001) describes possible elaborations of this basic measure.

## 5.2   AMBR

In this section we describe in some detail the constraints and requirements of the process of developing an ACT-R model for a task of moderate complexity and the range of quantitative predictions that one can expect from such a model. The task is a synthetic air traffic control simulation that was developed for the agent-based modeling of behavior representation (AMBR) comparison (Pew and Gluck, 2004) that arose from a report (Pew and Mavor, 1998) that highlighted the need for more robust, realistic human performance models (HPMs) for use in simulations for training and system acquisition

The AMBR project was designed to advance the state of the art in cognitive and behavioral modeling, especially models of integrative performance, requiring the coordination of memory, learning, multitasking, interruption handling, and perceptual and motor systems in order to scale more effectively to real-world environments. The program provided a structure to gather human performance data and evaluate the accuracy and predictiveness of the models. The AMBR program was organized as a series of comparisons among alternative modeling approaches including ACT-R but also the Air Force Research Laboratory's DCOG (Eggleston et al., 2001), CHI Systems, Inc.'s COGNET/iGEN (Zachary et al., 2001), and George Mason University's EASE (Chong, 2001).

The task designed to elicit the desired behaviors is a synthetic air traffic control simulation. This domain requires a controller to manage one sector of airspace, especially the transition of aircraft into and out of the sector. Scenarios can vary the number, speed, altitude, and type of aircraft requesting access to the sector and can be complicated by having them arrive from multiple directions and adjoining sectors. This is a rich enough infrastructure to create a variety of scenarios having variable task load levels and varying levels of planning complexity. Figure 20 displays a screen shot of the simulation. The main part of the screen on the left contains a graphical representation of the entire airspace, with the part controlled by the human or model agent contained in the central yellow square. The rest of the airspace is divided by the yellow lines in four
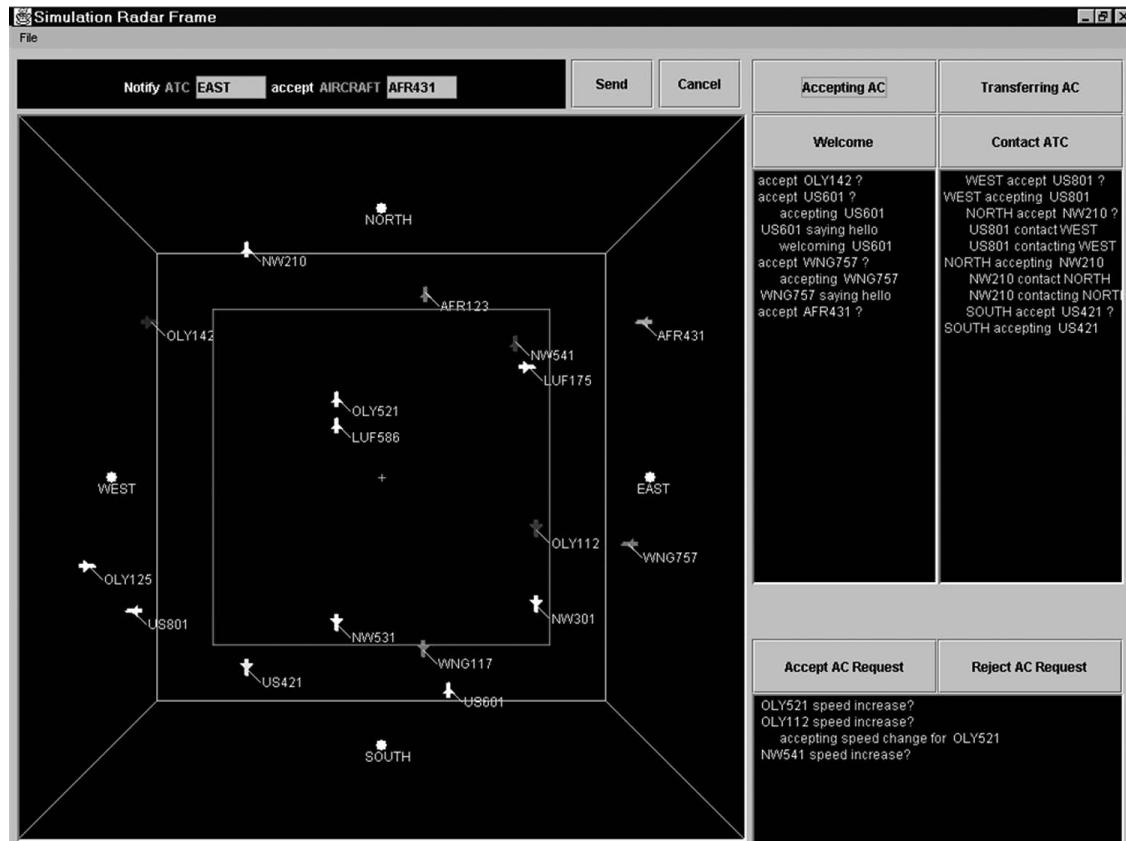
**Figure 20** Screen shot of the AMBR simulation.

regions, north, east, south, and west, each managed by a separate controller. At any point during the simulation a number of airplanes (the exact number being a parameter controlling the difficulty of the task) are present in the airspace, flying through the central region or entering or exiting it. The task of the central controller is to exchange messages with the airplanes (each tagged with its identifying code, e.g., UAL344) and neighboring controllers to manage their traversal of its airspace. Those messages are displayed in the text windows on the right of the screen, with each window dedicated to a specific message category. The top left window concerns messages sent when a plane is entering the central controller's region, while the top right window concerns messages sent when a place is exiting the central region. Both windows include messages exchanged between controllers as well as messages between the central controller and the plane itself. The bottom window concerns messages from and to planes requesting a speed increase, which should be granted unless that plane is overtaking another plane, which is the only airspace conflict that this simplified task allows.

A single event involves a number of messages being exchanged, all of which are appended to the relevant text window. For example, in the case of a plane about

to enter the central region, a message requesting permission to enter will first be sent to the central controller from the controller of the neighboring region from which the plane originates. The central controller must reply to the other controller in a timely manner to accept the plane, then contact the plane to welcome it to the airspace. Those two cannot be performed in immediate succession but, instead, require waiting for the first party contacted (in this case the other controller) to reply before taking the final action. This delay allows for the interleaving of unit tasks but also requires the maintenance of the currently incomplete tasks in working memory. Messages from other tasks can arrive when a task is being processed, thus requiring some search of the text window to identify the messages relevant to a task. A message is composed by clicking a button above the relevant text window (e.g., accepting AC), then clicking in the graphical window on the intended recipient (e.g., another controller) and optionally the target of the message (i.e., a plane, unless it is the intended recipient, in which case this is omitted), then the send button above the graphic window. The message being composed is displayed at the top left of the display in a text window.

To measure performance on the task objectively, penalties were assessed for a variety of failures to act

in a timely manner. To evaluate the impact of system design, a decision support condition contrasted with a support condition were implemented to dissociate two aspects of multitasking behavior. In the standard condition, subjects had to parse the messages printed in the text windows on the right side of the screen to determine which planes needed attention and which functions needed to be performed on them. In the assisted condition, planes that require assistance were color coded in the graphical display on the left side of the screen according to the task that needed to be performed (green for accept, blue for welcome, orange for transfer, yellow for contact, magenta for speed change, and red for holding). This helped the subjects track visually which tasks needed to be attended to and removed any necessity to parse the text windows on the left, a complex and time-consuming task. Therefore, it dissociated the maintenance and updating of the queue of to-be-attended tasks from the resolution of conflicts between high-priority tasks. Two sets of scenarios were created: One set was provided to the developers as a model on which to base their designs, and another set was reserved to be used at the time of the competitive validation (i.e., the fly-off). Human performance data on the first set of scenarios were provided to the developers to fine-tune their model. The data from the second set of scenarios were withheld until after the fly-off for comparison with the model performance. The range of behavior requirements of both sets had the same scope, but the ways in which those behaviors were exercised were not identical, to test the robustness and predictiveness of the models.

## 5.3   Model Development

If it is to justify its structural costs, a cognitive architecture should facilitate the development of a model in several ways. It should limit the space of possible models to those that can be expressed concisely in its language and work well with its built-in mechanisms. It should provide for significant transfer from models of similar tasks, either directly in the form of code or more generally in the form of design patterns and techniques. Finally, it should provide learning mechanisms that allow the modeler to specify in the model only the structure of the task and let the architecture learn the details of the task in the same way that human cognition constantly adapts to the structure of its environment. These architectural advantages not only reduce the amount of knowledge engineering required and the number of trial-and-error development cycles, providing significant savings in time and labor, but also improve the predictiveness of the final model. If the "natural" model (derived a priori from the structure of the task, the constraints of the architecture, and the guidelines from previous models of related tasks) provides a good fit to the empirical data, one can be more confident that it will generalize to unforeseen scenarios and circumstances than if it is the result of post hoc knowledge engineering and data analysis. That is the approach that we adopted in developing a model of this task and indeed, more generally, our design and use of the ACT-R architecture.

Of course, in domains involving a large body of expertise, it makes sense to encode in the cognitive model the accepted knowledge of the field. But in synthetic tasks or in tasks involving new system design, specific established knowledge is usually inexistent or inaccessible. Thus, we did not try to reverse engineer the subjects' strategies but instead tried to develop the simplest and most natural model for the architecture. We organized the model around a few goal types with their associated productions. Goal types correspond closely to the unit tasks in human–computer interaction (Card et al., 1983) as well as to the tasks in task network models (e.g., Allender et al., 1995). Five goal types, called color-goal, text-goal, scan-text, scan-screen, and process, were defined, together with a total of 36 very simple productions. Goals were simple and would hold just a few elements, such as the aircraft currently being handled together with related information such as its position and the action to be performed, in accordance with architectural constraints. Overall, such model development need not take more than a few days. Two basic modes of human interaction with the simulation were defined: one in which the operator had to rely mostly on text messages scrolling in windows to identify events that required action (the text condition) and one in which aircraft on the radar screen that required action would turn a color corresponding to the action (the color condition). The simulation also had three speeds (low, medium, and high) that controlled how much time the subjects would have (10, 7.5, and 5 min, respectively) to perform a given number of actions.

The goal type color-goal was the top goal for the color condition. Five productions were defined that applied to that goal. They scanned the radar screen continuously, identified an aircraft that had turned color, mapped the color into the required action by relying on five simple memory chunks encoding the instructions that the subjects were given regarding the color-action mappings, then created a goal to perform the given action on the aircraft. The goal-type process executed the sequence of mouse clicks required to perform the action. Twelve productions were defined to handle the five possible actions. This required clicking on a button identifying the action, then on the aircraft, then perhaps on a neighboring controller, then finally on the send button.

As expected, the text condition was both more difficult for the subjects and slightly more complicated for the model. The goal type text-goal was the top goal for the text condition. Four productions were defined to cycle through the three text windows and the radar screen looking for aircraft requiring action by creating goals of type scan-text and scan-screen, respectively. A goal of type scan-text would handle the scanning of a single text window for a new message from another controller requesting action. A production was defined to scan the window systematically for such a message. If one was found, another production would attempt to retrieve a memory of handling such a request. Memories for such requests would be created automatically by the architecture when the corresponding goal was completed, but their availability was subject to their subsymbolic parameters, which were in turn subject to decay as well as reinforcement. If no memory could

be retrieved, the window would be scanned for another message, indicating completion. If none could be found, a process goal would be created to perform the action requested. Note that this is the same goal as in the color condition. A key component of the model was an additional production that would detect the onset of a new message in another window and interrupt the current goal to scan that window instead. This allowed the model to be sensitive to new events and handle them promptly. Scanning the radar screen was accomplished in a similar manner by goals of type scan-screen and their eight associated productions.

Finally, all the architectural parameters that control the performance of the simulation were left at their default values provided by previous models. A key aspect of our methodology, which is also pervasive in ACT-R modeling, is the use of Monte Carlo simulations to reproduce not only the aggregate subject data (such as the mean performance or response time) but also the variation that is a fundamental part of human cognition. Especially when evaluating system design, it is essential not only to capture an idealized usage scenario but as broad a range of performance as possible. In that view, the model does not represent an ideal or even average subject, but instead, each model run is meant to be equivalent to a subject run, in all its variability and unpredictiveness. For that to happen, it is essential that the model not merely be a deterministic symbolic system but also be able to exhibit meaningful nondeterminism. To that end, randomness is incorporated in every part of ACT-R's subsymbolic level, including chunk activations, which control their probability and latency of retrieval; production utilities, which control their probability of selections; and production efforts, which control the time that they spent executing.

Moreover, as has been found in other ACT-R models (e.g., Lerch et al., 1999), that randomness is amplified in the interaction of the model with a dynamic environment: Even small differences in the timing of execution might mean missing a critical deadline, which results in an error condition, which requires immediate attention, which might cause another missed deadline, and so on. To model the variation as well as the mean of subject performance, the model was always run as many times as there were subject runs. For that to be a practical strategy of model development, it is essential that the model run very fast, ideally significantly faster than real time. Our model ran up to five times faster than real time on a desktop PC, making it possible to run a full batch of 48 scenarios in about an hour an a half, enabling a relatively quick cycle of model development.

### 5.4 Modeling Results

Because the variability in performance between runs, even of the same subject, is a fundamental characteristic of this task, we ran as many model runs as there were subject runs. Figure 21 compares the mean performance in terms of penalty points for subjects and model for color (left three bars) and text (right three bars) condition by increasing workload level. The model matches the
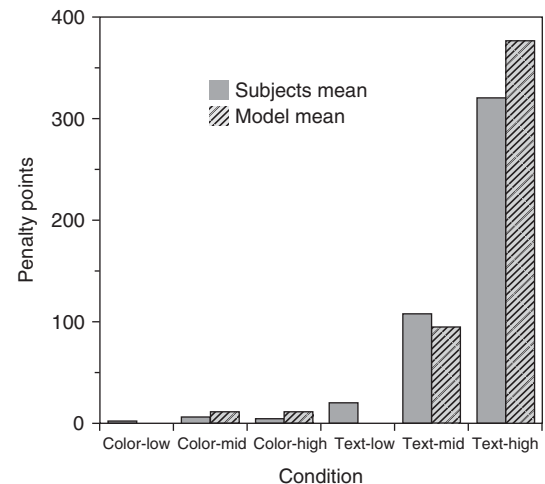


**Figure 21** Mean performance as a function of workload and system design.

data quite well, including the strong effects of the color-versus-text condition and of workload for the unaided (text) condition.

Because ACT-R includes stochasticity in chunk retrieval, production selection, and perceptual/motor actions, and because that stochasticity is amplified by the interaction with a highly dynamic simulation, it can reproduce a large part of the variability in human performance, as indicated by Figure 22, which plots the individual subject and model runs for the two conditions that generated a significant percentage of errors (text condition in medium and high workload). The range of performance in the medium-workload condition is reproduced almost perfectly other than for two outliers, and a significant portion of the range in the high condition is also reproduced, albeit shifted slightly too upward. It should be noted that each model run is the result of an identical model that differs from another only in its run time stochasticity. The model neither learns from trial to trial nor is modified to take into account individual differences.

The model not only reproduces the subject performance in terms of total penalty points but also matches well to the detailed subject profile in terms of penalties accumulated under eight different error categories, as plotted in Figure 23. It should be emphasized that those errors were not engineered in the model but, instead, resulted directly from the limitations of the cognitive architecture applied to a demanding, fast-paced dynamic task.

The model also fits the mean response times (RTs) for each condition, as shown in Figure 24, which plots the detailed pattern of latencies to perform a required action for each condition and number of intervening events (i.e., number of planes requiring action between the time of a given plane requiring action and the time the action is actually performed). The model predicts very accurately the degradation of RT as more
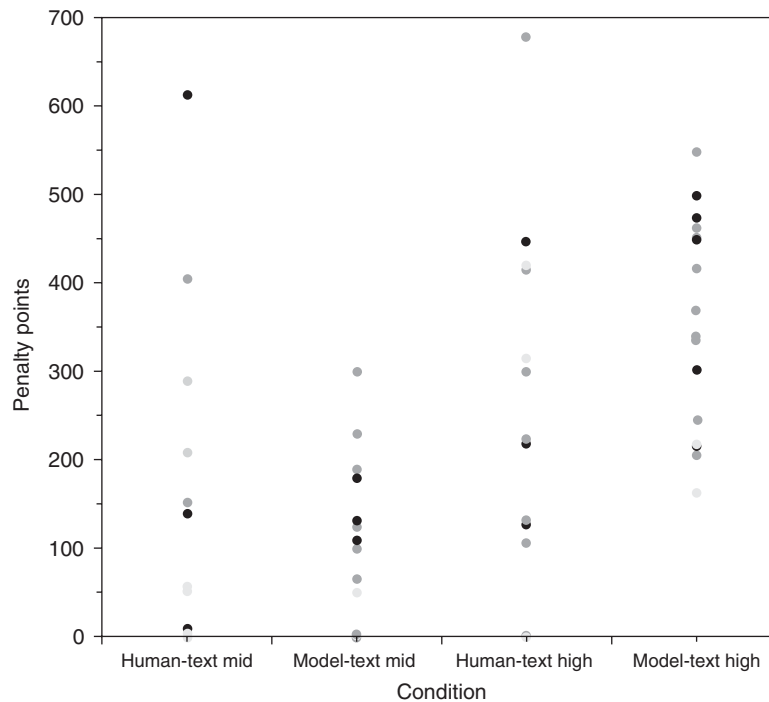
**Figure 22**  Mean performance as a function of workload and system design.
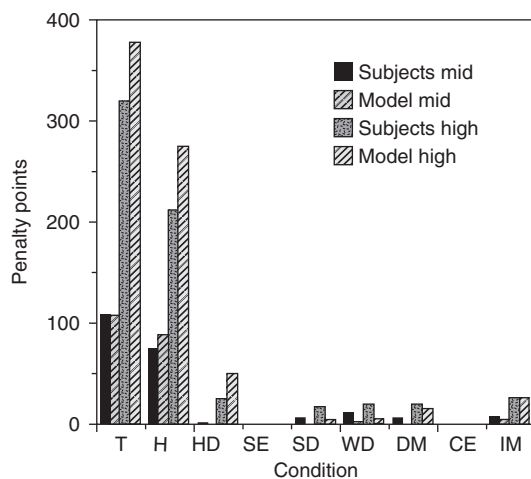


**Figure 23**  Penalty points for a variety of error categories.

events compete for attention, including the somewhat counterintuitive exponential (note that RT is plotted on a log scale) increase in RT as a function of number of events rather than a more straightforwardly linear increase. The differences in RT between conditions are

primarily a function of the time taken by the perceptual processes of scanning radar screen and text windows.

Finally, the model reproduces the subjects' answers to the self-reporting workload test administered after each trial. As shown in Figure 25, the simple definition of workload described in Section 5.3 captures the main workload effects, specifically effects of display condition and schedule speed. The latter effect results from reducing the total time to execute the task (i.e., the denominator) while keeping the total number of events (roughly corresponding to the numerator) constant, thereby increasing the ratio. The former effect results from adding to the process tasks the message-scanning tasks resulting from onset detection in the text condition, thus increasing the numerator while keeping the denominator constant, thereby increasing the ratio as well. Another quantitative effect that is reproduced is the higher rate of impact of schedule speed in the text condition (and the related fact that workload in the slowest text condition is higher than workload in the fastest color condition). This is primarily a result of task embedding [i.e., the fact that a process task can be (and often is) a subgoal of another critical unit task (scanning a message window following the detection of an onset in that window)], thus making the time spent in the inner critical task count twice.

Lebiere (2004) reports the results of a second phase of the AMBR comparison in which the model had to learn how to categorize airplanes properly based
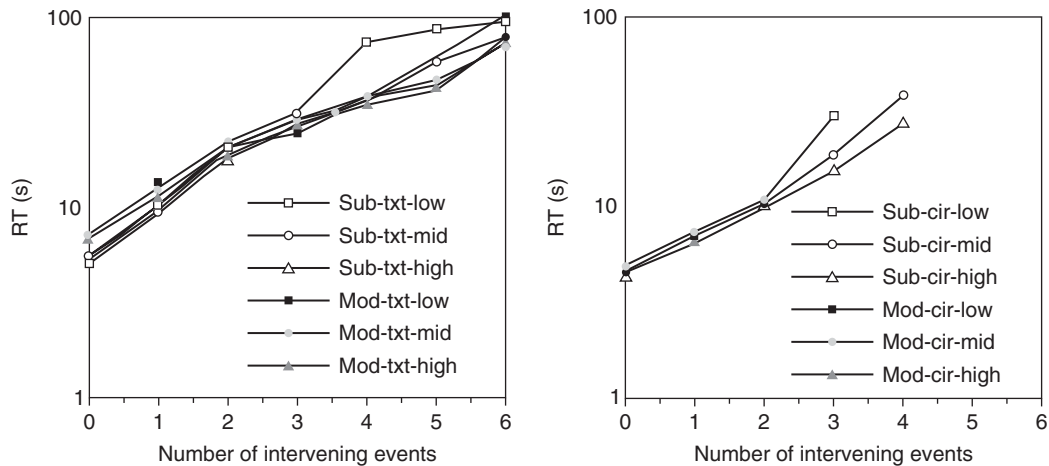
**Figure 24** Response time as a function of intervening events.
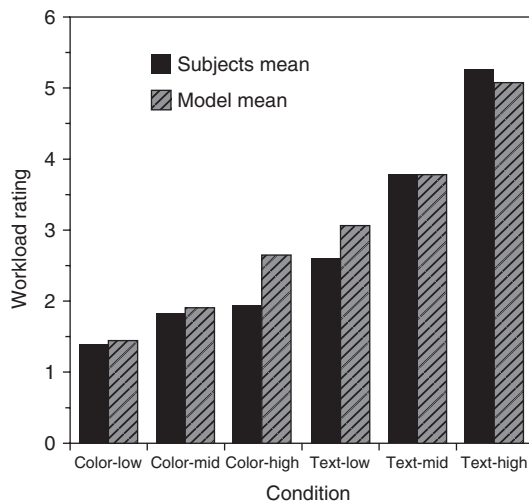


**Figure 25** Workload levels for various speed and display conditions.

on a simple pass–fail feedback. This model is similar to the one described here but leverages even more extensively the subsymbolic aspects of the architecture, especially the learning equations described in the ACT-R introductory section, to perform the learning task as a constrained component of the entire task. In summary, the advantages of this model are that it is relatively simple, required almost no parameter tuning or knowledge engineering, provides a close fit to both the mean and variance of a wide range of subject performance measures as well as workload estimates, and suggests a straightforward account of multitasking behavior within the existing constraints of the ACT-R architecture.

## 6 INTEGRATION OF APPROACHES

Because ACT-R and IMPRINT were targeted at different behavioral levels, they complement each other perfectly. IMPRINT is focused on the task level, how high-level functions break down into smaller scale tasks, and the logic by which those tasks follow each other to accomplish those functions. ACT-R is targeted at the "atomic" level of thought, the individual cognitive, perceptual, and motor acts that take place at the subsecond level. As shown in Figure 19 and in the previous example, the current goal is a central concept in ACT-R which corresponds directly to the concept of unit task. At each cycle, a production will be chosen that best applies to the goal, knowledge might be retrieved from declarative memory, and perceptual and motor actions may be taken. Those cycles will repeat until the current goal is solved, at which point it is popped and another one is selected. The ACT-R theory specifies in detail the performance and learning that takes place at each cycle within a specific goal but has comparatively little to say about the selection of those goals. Since goals in ACT-R correspond closely to tasks in IMPRINT, that weakness matches IMPRINT's strength perfectly. Conversely, since IMPRINT requires the characteristics of each task to be specified as part of the model, ACT-R can be used to generate those detailed characteristics in a psychologically plausible way without requiring extensive data collection. Thus, an integrated ACT-R/IMPRINT is structured along as pictured in Figure 26.

An IMPRINT model specifies the network of tasks used to accomplish the functions targeted by the model (e.g., landing a plane and taxiing safely to the gate). The network specifies how higher order functions are decomposed into tasks and the logic by which these tasks are composed together. As input, it takes the distribution of times to complete the task and the accuracy with which the task is completed. It can also take as input
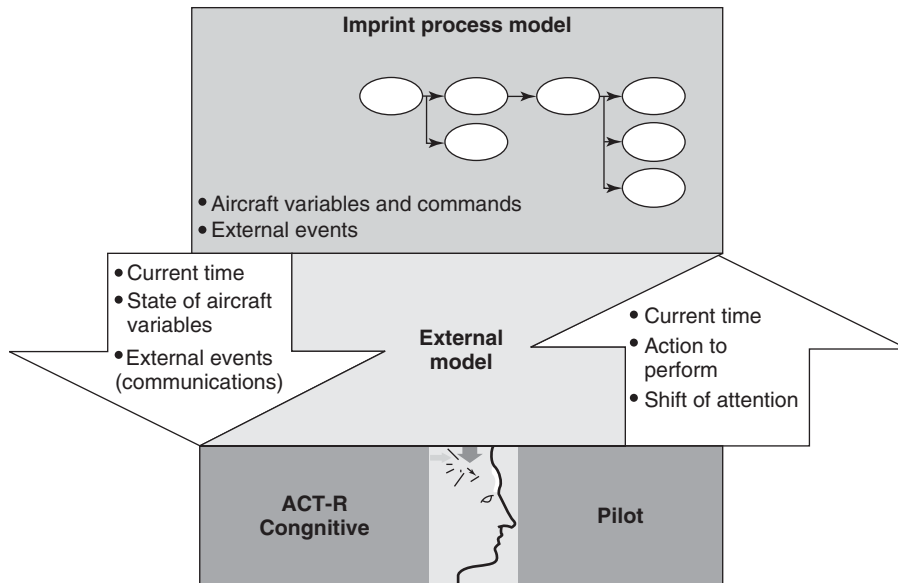
**Figure 26**  Integrated ACT-R/IMPRINT model.

the workload generated by each task. Additional inputs include events generated by the simulation environment. Finally, a number of additional general parameters, such as personnel characteristics, level of training, and familiarity and environmental stressors, can be specified. IMPRINT specifies the performance function by which these parameters modulate human performance. The outputs include mission performance data such as time and accuracy as well as aggregate workload data.

An ACT-R model specifies the knowledge structures, such as declarative chunks and production rules that constitute the user knowledge relevant to the tasks targeted by the model. It also specifies the goal structures reflecting the task structure and the architectural and prior knowledge parameters that modulate the model's performance. For each goal on which ACT-R is focused (i.e., made the current goal), it generates a series of subsecond cognitive, perceptual, and motor actions. The result of those actions is the total time to accomplish the goal as well as how the goal was accomplished, including any error that might result. Errors in ACT-R originate from a broad range of sources. They include memory failures, including the failure to retrieve a needed piece of information or the retrieval of the wrong piece of information; choice failures, including the selection of the wrong production rule; and attentional failures, such as the failure to detect the salient piece of information by the perceptual modules. Although those errors could arise because of faulty symbolic knowledge (either declarative or procedural), it is often not the case, especially in domains that involve highly trained crews. More often, those errors occur because the subsymbolic parameters associated with chunks or productions do not allow the

model to access them reliably or quickly enough to be deployed in the proper situation.

Moreover, because those parameters vary stochastically and their effect is amplified by the interaction with a dynamic environment, those times and errors will not be deterministic but will vary with each execution, as is the case for human operators. Thus, the ACT-R model for a particular goal can be run whenever IMPRINT selects the corresponding task to generate the time and error distribution for that task in a manner that reflects the myriad cognitive, perceptual, and motor factors that enter into the actual performance of the task. As seen in the previous example, ACT-R can also generate workload estimates for each goal that reflect the cognitive demands of the actions taken to perform that particular subtask, then pass those estimates to IMPRINT, which can then combine them into global workload estimates for the entire task. ACT-R and IMPRINT have been unified in a single integrated development environment with the Human Behavior Architecture (HBA) tool (Warwick et al., 2008).

### 6.1  Sample Applications

As a practical application of the IMPRINT and ACT-R integration, a complex and dynamic task was selected for a modeling effort. Researchers with the National Aeronautics and Space Administration (NASA) were interested in developing models of pilot navigation while taxiing from a runway to a gate. Research on pilot surface operations had shown that pilots can commit numerous errors during taxi procedures (Hooey and Foyle, 2001). NASA was hoping to reduce the number and scope of pilot error during surface operations by

using information displays that would improve the pilots' overall situation awareness.

NASA researchers provided the IMPRINT and ACT-R modeling teams with data describing pilot procedures during prelanding and surface taxi operations. These data included videotapes of pilots in the NASA Ames Advanced Concept Flight Simulator (ACFS), which is a simulated cockpit capable of duplicating pilot taxiing operations. A detailed, scaled map of Chicago's O'Hare airport was also provided, which included runway signage. Other types of documentation were provided to give the IMPRINT and ACT-R modeling team the information necessary to duplicate runway taxiing behavior by pilots.

The IMPRINT and ACT-R modeling teams used the scaled map of Chicago's O'Hare airport to estimate the time between runway taxi turns. IMPRINT handled the higher level, task-oriented parts of the taxiing and landing operations (i.e., turning, talking on radio, looking at instrumentation), while ACT-R handled the more cognitive and decision-making parts of the task (i.e., remembering where to turn, remembering the taxi route). By using the scaled map of the airport, the IMPRINT and ACT-R teams were able to determine the amount of time between each taxi turn (based on an estimated plane speed that was correlated with the simulated speeds from the videotape data) and then use those data to estimate the decay rate for the list of memory elements (i.e., runway names) that the pilot would have to remember.

Using this integrated architecture allowed the team to represent a complex, dynamic task, and by exploiting each architecture's strengths, the modeling process was enhanced and streamlined. The resulting model could account for a broad range of possible taxiing errors within a constrained first-principle framework, as was the case for the stand-alone AMBR model, but in addition benefited by integration with the task network model, which provided a convenient task-based organizing framework to minimize the authoring requirements for the cognitive model as well as to provide a high-productivity tool to simulate the environment and aircraft with which the cognitive model interacts.

Craig et al. (2002) performed a similar integration of ACT-R into the combat automation requirements tool (CART)* model (Brett et al., 2002), a task network model used in the acquisition process of the joint strike fighter. The task to be performed was target acquisition, more specifically, management of the shoot list, which allows a pilot to select potential targets to be identified by high-resolution radar. Using a methodology similar to that described above, specific subtasks were identified for which additional cognitive fidelity was required and reimplemented in the form of ACT-R goals and associated production rules. ACT-R then interacted with the CART model, providing plausible performance for cognitive subtasks such as prioritizing targets and recalling items identified previously.

---

* The reader should note that the CART model capabilities are now subsumed into the IMPRINT tool.

## 7 SUMMARY

In this chapter we have reviewed the need for simulating performance of complex human-based systems as an integral part of system design, development, testing, and life-cycle support. We have also defined two fundamentally different approaches to modeling human performance, a reductionist approach and a first-principle approach. Additionally, we have provided detailed examples of two modeling environments that typify these two approaches along with representative case studies. Finally, we described an integrated tool that attempts to leverage the advantages of both approaches into an efficient and principled modeling package.

As we have stated and demonstrated repeatedly throughout this chapter, the technology for modeling human performance in systems is evolving rapidly. Furthermore, the breadth of questions being addressed by models is expanding constantly. Necessity being the mother of invention, we encourage the human factors practitioner to consider how computer simulation can provide a better and more cost-effective basis for human factors analysis and in turn stimulate further developments in modeling and simulation tools to better serve their needs.

## REFERENCES

Alion Science and Technology (2009), *Micro Saint Sharp Version 3.0 User's Guide*, Alion Science and Technology/MA&D Operation, Boulder, CO.

Allender, L. (1995, December), personal communication.

Allender, L., Kelley, T., Salvi, L., Headley, D. B., Promisel, D., Mitchell, D., Richer, C., and Feng, T. (1995), "Verification, Validation, and Accreditation of a Soldier-System Modeling Tool," in *Proceedings of the 39th Human Factors and Ergonomics Society Meeting*, October 9-13, San Diego, CA, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1219–1223.

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press, New York.

Anderson, J. R., and Lebiere, C. (1998), *The Atomic Components of Thought*, Lawrence Erlbaum Associates, Mahwah, NJ.

Anderson, J. R., and Schooler, L. J. (1991), "Reflections of the Environment in Memory," *Psychological Science*, Vol. 2, pp. 396–408.

Anderson, J. R., Matessa, M., and Lebiere, C. (1997), "ACT-R: A Theory of Higher Level Cognition and Its Relation to Visual Attention," *Human-Computer Interaction*, Vol. 12, No. 4, pp. 439–462.

Anderson, J. R., Bothell, D., Lebiere, C., and Matessa, M. (1998), "An Integrated Theory of List Memory," *Journal of Memory and Language*, Vol. 38, pp. 341–380.

Anderson, J. R., Qin, Y., Jung, K-W., & Carter, C. S. (2007). "Information-Processing Modules and their Relative Modality Specificity," *Cognitive Psychology*, 54(3), pp. 185–217.

Archer, S. G., and Adkins, R. (1999), *Improved Performance Research Integration Tool (IMPRINT) Analysis Guide*, Army Research Laboratory Technical Report, Aberdeen Proving Ground, MD.

Boff, K. R., Kaufman, L., and Thomas, J. P. (1986), *Handbook of Perception and Cognition*, Wiley, New York.

Brett, B. E., Doyal, J. A., Malek, D. A., Martin, E. A., Hoagland, D. G., and Anesgart, M. N. (2002), *The Combat Automation Requirements Testbed (CART) Task 5 Interim Report: Modeling a Strike Fighter Pilot Conducting a Time Critical Target Mission*, AFRL-HE-WP-TR-2002-0018, Wright-Patterson Air Force Base, OH.

Byrne, M. D., and Anderson, J. R. (2001), "Serial Modules in Parallel: The Psychological Refractory Period and Perfect Time-Sharing," *Psychological Review*, Vol. 108, pp. 847–869.

Card, S. K., Moran, T. P., and Newell, A. (1983), *The Psychology of Human-Computer Interaction*, Lawrence Erlbaum Associates, Mahwah, NJ.

Chong, R. (2001), "Low-Level Behavioral Modeling and the HLA: An EPIC-Soar Model of an Enroute Air-Traffic Control Task," in *Proceedings of the 10th Conference on Computer Generated Forces and Behavior Representation*, Norfolk, VA, pp. 27–35, Orlando, FL, Division of Continuing Education, University of Central Florida.

Craig, K., Doyal, J., Brett, B., Lebiere, C., Biefeld, E., and Martin, E. (2002), "Development of a Hybrid Model of Tactical Fighter Pilot Behavior Using IMPRINT Task Network Model and ACT-R," in *Proceedings of the 11th Conference on Computer Generated Forces and Behavior Representation*, Orlando, FL.

Dahl, S. G., Allender, L., Kelley, T., and Adkins, R. (1995), "Transitioning Software to the Windows Environment: Challenges and Innovations," in *Proceedings of the 1995 Human Factors and Ergonomics Society Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, October, pp.1224–1227.

Eggleston, R. G., Young, M. J., and McCreight, K. L. (2001), "Modeling Human Work through Distributed Cognition," in *Proceedings of the 10th Annual Conference on Computer Generated Forces and Behavior Representation*, Norfolk, VA.

Engh, T., Yow, A., and Walters, B. (1998), *An Evaluation of Discrete Event Simulation for Use in Operator and Crew Performance Evaluation*, report to the Nuclear Regulatory Commission, NRC, Washington, DC.

Farmer, E. W., Belyavin, A. J., Jordan, C. S., Bunting, A. J., Tattershall, A. J., and Jones, D. M. (1995), *Predictive Workload Assessment: Final Report*, DRA/AS/MMI/CR95100/, Defence Research Agency, Farnborough, Hampshire, England, March.

Gawron, V. J., Laughery, K. R., Jorgensen, C. C., and Polito, J. (1983), "A Computer Simulation of Visual Detection Performance Derived from Published Data," in *Proceedings of the 2nd International Ohio State University Aviation Psychology Symposium*, Columbus, OH, April, pp. 465–471.

Gore, B. F., Hooey, B. L., Wickens, C. D., and Scott-Nash, S. (2009), "A Computational Implementation of a Human Attention Guiding Mechanism in MIDAS v5," in *Proceedings of the Human Interaction International 2009 Annual Conference*, San Diego, CA.

Gray, W. D., Young, R. M., and Kirschenbaum, S. S. (1997), Special Issue: Cognitive Architectures and Human-Computer Interaction, *Human-Computer Interaction*, Vol. 12, No. 4.

Gunzelmann, G., Byrne, M. D., Gluck, K. A., & Moore, L. R. (2009). "Using computational cognitive modeling to predict dual-task performance with sleep deprivation," *Human Factors*, Vol. 51, No. 2, pp. 251–260.

Hahler, B., Dahl, S., Laughery, R., Lockett, J., and Thein, B. (1991). "CREWCUT: A Tool for Modeling the Effects of High Workload on Human Performance," paper presented at the 35th Annual Human Factors Society Meeting, San Francisco.

Hoagland, D. G., Martin, E. A., Anesgart, M., Brett, B. S., Lavine, N., and Archer, S. G. (2001), "Representing Goal-Oriented Human Performance in Constructive Simulations: Validation of a Model Performing Complex Time-Critical-Target Missions," in *Proceedings of the Simulation Interoperability Workshop*, Orlando, FL, Spring.

Hooey, B. L., and Foyle, D. C. (2001), "A Post-Hoc Analysis of Navigation Errors during Surface Operations: Identification of Contributing Factors and Mitigating Strategies," in *Proceedings of the 11th Symposium on Aviation Psychology*, Ohio State University, Columbus, OH.

Horrey, W. J., and Wickens, C. D. (2003), "Multiple Resource Modeling of Task Interference in Vehicle Control, Hazard Awareness and In-Vehicle Task Performance," in *Proceedings of the Second International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Park City, Utah, pp. 7–12.

Horrey, W. J., Wickens, C. D., and Consalus, K. P. (2006) "Modeling drivers' visual attention allocation while interacting with in-vehicle technologies," in *Journal of Experimental Psychology: Applied*, Vol. 12, No. 2, pp. 67–86.

John, B. E., Prevas, K., Salvucci, D. D., & Koedinger, K. (2004). "Predictive human performance modeling made easy." in *CHI 2004, ACM Conference on Human Factors in Computing Systems, CHI Letters* 6(1).

Klein, G. (1998), *Sources of Power: How People Make Decisions*, The MIT Press, Cambridge, MA.

LaVine, N. D., Peters, S. D., and Laughery, K. R. (1995), *A Methodology for Predicting and Applying Human Response to Environmental Stressors*, Micro Analysis and Design, Inc., Boulder, CO, December.

Lawless, M. L., Laughery, K. R., and Persensky, J. J. (1995), *Micro Saint to Predict Performance in a Nuclear Power Plant Control Room: A Test of Validity and Feasibility*, NUREG/CR-6159, Nuclear Regulatory Commission, Washington, DC, August.

Lebiere, C. (2001), "A Theory-Based Model of Cognitive Workload and Its Applications," in *Proceedings of the 2001 Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*, National Defense Industrial Association (NDIA), Arlington, VA.

Lebiere, C. (2004), "Constrained Functionality: Application of the ACT-R Cognitive Architecture to the AMBR Modeling Comparison," in *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation*, R. W. Pew and K. A. Gluck, Eds., Lawrence Erlbaum Associates, Mahwah, NJ.

Lerch, F. J., Gonzalez, C., and Lebiere, C. (1999), "Learning under High Cognitive Workload," in *Proceedings of the 21st Conference of the Cognitive Science Society*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 302–307.

Lovett, M. C., Reder, L. M., and Lebiere, C. (1999), "Modeling Working Memory in a Unified Architecture: An ACT-R Perspective," in *Models of Working Memory*, A. Miyake and P. Shah, Eds., Cambridge University Press, Cambridge, MA.

Matessa, M., and Mui, R. (2009), "GRBIL: A Tool for Dynamic Operator and Environment Modeling," in *Advanced Decision Architectures for the Warfighter: Foundations and Technology*, P. McDermott and L. Allender, Eds., Army Research Laboratory, Boulder, CO, pp 237–254.

McCracken, J. H., and Aldrich, T. B. (1984), *Analysis of Selected LHX Mission Functions: Implications for Operator Workload and System Automation Goals*, Technical Note ASI 479-024-84(B), Anacapa Sciences, Ft. Rucker, AL, June.

Newell, A. (1990), *Unified Theories of Cognition*, Harvard University Press, Cambridge, MA.

Newell, A., and Rosenbloom, P. S. (1981), "Mechanisms of Skill Acquisition and the Power Law of Practice," in *Cognitive Skills and Their Acquisition*, J. R. Anderson, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1–56.

Newell, A., and Simon, H. A. (1972), *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, NJ.

O'Hara, J. M., Brown, W. S., Stubler, W. F., Wachtel, J. A., and Persensky, J. J. (1995), *Human-System Interface Design Review Guideline: Draft Report for Comment*, NUREG-0700 Rev. 1, U.S. Nuclear Regulatory Commission, Washington, DC.

Pew, R. W., and Gluck, K. A. (2004). *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation*, Lawrence Erlbaum Associates, Mahwah, NJ.

Pew, R. W., and Mavor, A. S. (1998), *Modeling Human and Organizational Behavior: Application to Military Simulations*, National Academy Press, Washington, DC.

Plott, B. (1995), *Software User's Manual for WinCrew, the Windows-Based Workload and Task Analysis Tool*, U.S. Army Research Laboratory, Aberdeen Proving Ground, MD.

Plott, B., Quesada, S., Kilduff, P., Swoboda, J., and Allender, L. (2004) "Using an Information-Driven Decision-Making Human Performance Tool to Assess U.S. Army Command, Control, and Communication Issues," in *Proceedings from the Human Factors and Ergonomics Society 48th Annual Meeting*.

Roth, J. T. (1992), *Reliability and Validity Assessment of a Taxonomy for Predicting Relative Stressor Effects on Human Task Performance*, Technical Report 5060-1, prepared under contract DNA001-90-C-0139, Micro Analysis and Design, Boulder, CO, July.

Rubin, D. C., and Wenzel, A. E. (1990), "One Hundred Years of Forgetting: A Quantitative Description of Retention," *Psychological Review*, Vol. 103, pp. 734–760.

Siegel, A. I., and Wolf, J. A. (1969), *Man-Machine Simulation Models*, Wiley-Interscience, New York.

Taatgen, N. A. (2001), "A Model of Individual Differences in Learning Air Traffic Control," in *Proceedings of the 4th International Conference on Cognitive Modeling*, Gray, W. Altmann, E., Cleeremans, and Schunn, C. Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 211–216.

Warwick, W., Archer, R., Hamilton, A., Matessa, M., Santamaria, A., Chong, R., Allender, L., and Kelley, T. (2008). "Integrating Architectures: Dovetailing Task Network and Cognitive Models," in *Proceedings of Behavior Representation in Modeling and Simulation (BRIMS '08)*, Orlando, FL, University of Central Florida.

Warwick, W., and Santamaria, A. (2009), "Modeling the Recognition Primed Decision," in *Advanced Decision Architectures for the Warfighter: Foundations and Technology*, P. McDermott and L. Allender, Eds., final report for DAD19-01-2-0009, Army Research Laboratory Advanced Decision Architectures Collaborative Technology Alliance, Boulder, CO, pp. 273–288.

Wickens, C. D. (1984), *Engineering Psychology and Human Performance*, Charles E. Merrill, Columbus, OH.

Wickens, C. D., and McCarley, J. S. (2008), *Applied Attention Theory*, Taylor & Francis, Boca Raton, FL.

Wickens, C. D., Sandry, D. L., and Vidulich, M. (1983), "Compatibility and Resource Competition between Modalities of Input, Central Processing, and Output," *Human Factors*, Vol. 25, pp. 227–248.

Wortman, D. B., Duket, S. D., Seifert, D. J., Hann, R. L., and Chubb, A. P. (1978), *Simulation Using SAINT: A User-Oriented Instruction Manual*, AMRL-TR-77-61, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, OH, July.

Zachary, W., Santarelli, T., Ryder, J., Stokes, J., and Scolaro, D. (2001), "Developing a Multi-Tasking Cognitive Agent Using the COGNET/iGEN Integrative Architecture," in *Proceedings of the 10th Annual Conference on Computer Generated Forces and Behavior Representation*, Norfolk, VA.

# CHAPTER **33**

# MATHEMATICAL MODELS IN ENGINEERING PSYCHOLOGY: OPTIMIZING PERFORMANCE

**Donald L. Fisher**
**University of Massachusetts Amherst**
**Amherst, Massachusetts**


**Richard Schweickert**
**Purdue University**
**West Lafayette, Indiana**


**Colin G. Drury**
**University at Buffalo: State University of New York**
**Buffalo, New York**

## 1 INTRODUCTION

Mathematical models of human behavior serve scientific and applied functions in human factors and ergonomics and have long done so (Gray, 2008; Pew, 2008). Their bearing on the advance of theoretical behavioral science (e.g., Townsend and Ashby, 1983) is well documented, and their utility in more applied areas such as biomechanics is well established (e.g., Chaffin et al., 1999). However, their utility for the engineering psychologist is often overlooked. In fact, there is a general sense that the engineering psychologist cannot yet use mathematical models to design an actual interface. In an issue of *Human Factors*, in an article in a special section on quantitative formal models of human performance, this sense is clearly articulated: "An aim of human factors research is to have models that allow for the advance of user-friendly environments. This is still a distant dream because existing models are not yet sufficiently sophisticated" (Jax et al., 2003).

Clearly, no model or technique can handle all situations. But we believe there exist models that play and will continue to play a central role in the design of user-friendly environments (Fisher, 1993). Unfortunately, they lie scattered throughout a varied literature, one not easily accessible to many readers. Moreover, designers aspire to consider the entire task and its environment, but available models are likely to be targeted for particular aspects of the situation, thereby not coming to the attention of the design community. Finally, many models that could be useful in design are not extended in that direction, perhaps because the optimization techniques needed to make this transformation are part

of a field, operations research, which generally does not overlap with engineering psychology. One purpose of this chapter is to bring to the attention of the more general reader models and optimization techniques that allow creation of real, user-friendly environments.

Design lies at the center of the link between theory (the scientific model) and practice (the application or environment). Understanding this link makes clear the limited but essential role mathematical models play in the design process, one that many models can now play. The interface between the human operator and the environment is the design element, natural or built. This design element can be as simple as a kitchen utensil or as complex as the displays in a nuclear power plant control room. The design element is frequently constrained but within those constraints can take on an infinite or very large number of different configurations. For example, consider just the location of the keys on an automated teller machine (ATM). Suppose there are four keys positioned on the ATM, one key near each edge of a small square display window and each key accessing a different option (Figure 1). Suppose that any option can be assigned to any key. And suppose the menu hierarchy is three levels deep, with four options in each display (only two levels are displayed in Figure 1). Given this configuration, at the bottom level of the

hierarchy there are 16 menus each with 4 options, or a total of 64 different terminal options. Then there are over 1 quadrillion possible arrangements of the options in each display (4! arrangements in each display, so $24^{1+4+16}$ arrangements overall).

The link from theory to practice through design is now easily made. Clearly, it is not possible to evaluate all of the different arrangements (designs) experimentally. This is where mathematical models have a critical role, in engineering in general and, more specifically, in human factors engineering (Byrne and Gray, 2003). They can be used to predict performance for each different configuration of an interface and, in this case, for each different arrangement of the menu options. By itself, this may be all that is required. One can simply iterate through all possibilities and identify the one or ones that optimize performance. However, when the number of different configurations gets too large or is infinite, it is necessary either to derive the optimal solution analytically or to use methods that can approximate it. This is where knowledge of optimization techniques becomes critical and why in this chapter some attention is given to such techniques.

Formally, we can treat optimization as finding the maximum (or minimum) of an objective function: for
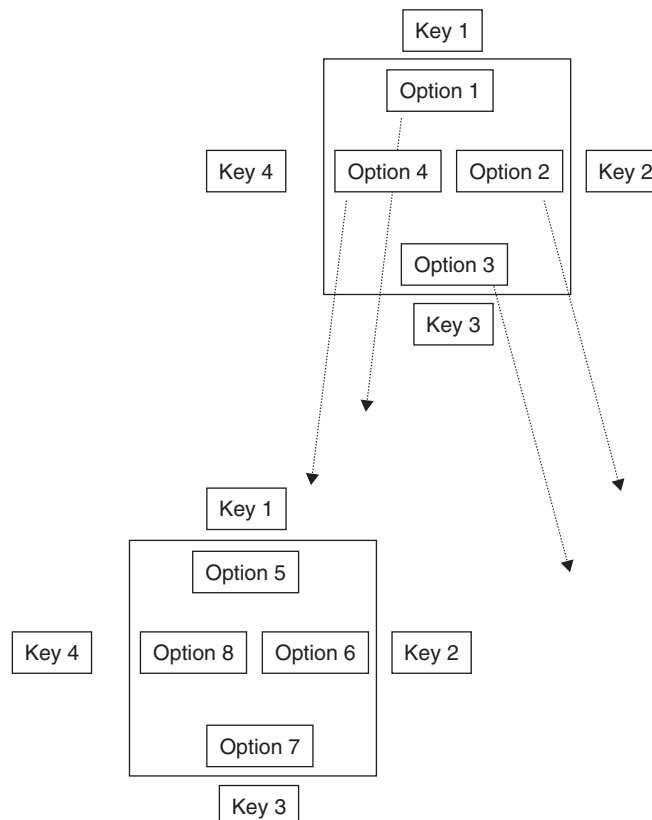


**Figure 1**  ATM arrangement of keys and options.

example, a weighted sum of $n$ key variables:

$$F(x) = \sum_{i=1}^{n} c_i x_i \qquad (1)$$

where the variables are $x_i$ and the weights on each variable are $c_i$. Of course, more complex and more general objective functions beyond the simple linear sum of equation (1) are possible. Generally, it is not possible to choose arbitrary values of the variables $x_i$, as these are constrained. In the ATM menu design example, we must choose arrangements that include one and only one placement of each of the four keys. In general, there will be a set of constraints, for example,

$$\sum_{i=1}^{n} a_i x_i \leq b_j \qquad (1 \leq j \leq m) \qquad (2)$$

where each of $m$ linear combinations of the variables $x_i$ and weights $a_i$ must be less than some constant $b_j$. Such optimization problems have been studied extensively in operations research, and solution procedures have been developed for many classes of problems: for example, linear programming (as above) or integer programming, where the $x_i$ values are constrained to integers. Not all mathematical models are optimization models, but many can be used in this way.

Although we focus on the use of mathematical models to optimize performance, we do not do so exclusively, for such models have a broader practical utility than simply the optimal design of an interface. First, the parameters of such models can indicate something about a quantity such as the relative speed of latent processes, a quantity that would be important if, say, one group has been exposed to a toxin and another group has not been so exposed (Smith and Langolf, 1981) or if, say, younger and older adults are being compared performing a particular task (Salthouse and Somberg, 1982). Second, models also have an important role to play before implementing an interface and incurring all the expenses that go along with such an implementation. Specifically, they can be used to estimate whether the interface will perform as desired (Gray et al., 1993). Third, models have still another, perhaps surprising role to play. They can identify situations where the intuitively most obvious course of action leads to paradoxical results (Meyer and Bitan, 2002). Fourth, because a mathematical model makes explicit the variables and parameters considered, the discipline of mathematical modeling also forces users to make explicit these variables and relationships, thus providing a solid foundation for the testing, extension, or perhaps ultimate rejection of the model. Somewhat paradoxically, there is an art to mathematical modeling, and that art is to abstract from the complex, real system of humans and devices those aspects most necessary for an accurate, yet economical prediction of performance. Of course, models have a theoretical purpose as well and we will make that clear as we go forward. Specifically, they are useful not only because they can provide testable predictions of complex theories (e.g., Sternberg, 1975;

Fific et al., 2010) but also because one can determine in certain cases whether models that appear on the surface as different indeed are different (Townsend, 1972).

The selection of examples below is necessarily limited by space, but also by a desire to give readers enough background material to understand how mathematical models, together with appropriate optimization techniques, can have a radical impact on design. An attempt has been made to be as broad as possible within the confines, giving to readers some sense of models currently used in design of interfaces as diverse as workstations, variable message signs, menu hierarchies, intelligent tutors, and warnings, among others.

Finally, we should end the introduction with a disclaimer. We recognize that some of the most creative design lies outside the scope of the procedures considered in this chapter. For example, consider the design of menu hierarchies. Recent efforts to improve performance include the compression of speech (Sharit et al., 2003) and, with cell phones and other technologies with very small display windows, the presentation of a portion of the hierarchy rather than just the current command (Tang, 2001). These more qualitative changes often bring about much larger changes than can be realized through the modeling and optimization techniques described below.

## 2  GENERAL MATHEMATICAL MODELS

We begin with some general tools that can be brought to bear on the design of an optimal interface. Perhaps the class of tools used most frequently by engineering psychologists are networks representing the processes involved in the performance of a task. Once the arrangement of the processes is understood, quantitative tools exist to estimate the response time. However, these tools often need to be augmented. For example, the output of an encoding process might not be perfect, changing from one trial to the next even though the objective stimulus did not change. The output of a decision process might change based on the payoff matrix; or the time to execute a response might change as a function of the number of possible responses. Tools to handle these and other more complex variations on behavior are discussed below.

### 2.1  Task Analysis and Activity Network Models

One can describe human performance in a task by starting with a general model of the human operator, identifying the components of the model critical for the task, and then using the model so constrained to predict performance. Introductions to the main general models for this purpose are discussed in Anderson (1993) and Byrne and Anderson (1998) for Adaptive Control of Thought–Rational (ACT-R), Liu et al. (2006) for Queuing Network–Model Human Processor (QN-MHP), Meyer and Kieras (1997a, b) for Executive-Process/Interactive Control (EPIC), and Newell (1990) for State Operator and Result (SOAR). It is often more direct to start by modeling the cognitive, perceptual, and motor activities in the specific task that the operator is performing. This approach is congruous with the state of

knowledge in experimental psychology. A psychologist may study memory, but the actual experiments will be conducted on a task such as recognition of an object. At this time, much knowledge in experimental psychology is organized as knowledge about the latent activities required to perform specific tasks such as searching a display or drawing a figure. A component of interest, such as perception, is emphasized, but the smallest unit of study is the task. This organizational scheme may be a temporary phase, or, if tasks turn out to be natural fundamental units, it may be permanent. In any case, most contemporary models are not for the entire human system, nor are they for isolated components. Most models are for a component within a task, even if presented in the literature as a model for the component alone. The implication is that if a model does not yet exist for performance in a new situation, it is unlikely that one can be made simply by snapping together existing models of components. A new model will usually need to be developed for the components in their new context.

### 2.1.1   Activity Networks

For context, the modeler needs a functioning model for the entire task. The handiest model for a task is often an activity network (e.g., Elmaghraby, 1977; Pritsker, 1979). An activity network indicates the arrangement of activities in the task, some following one upon another and some going on concurrently (see Figure 2). Each vertex $v_i$ in the network represents an activity $a_j$, and an arrow from one vertex to another indicates the order in which the activities must be performed. A path from a vertex $v_1$ to a vertex $v_2$, going along arrows in the proper directions, indicates that the activity represented at vertex $v_1$ precedes the activity represented at vertex $v_2$. Two activities $a$ and $b$ are called *sequential* if either $a$ precedes $b$ or $b$ precedes $a$. For example, in Figure 2 the activity "call begins," represented at vertex $v_1$, precedes the activity "system response," represented at vertex $v_2$. Two activities are called *concurrent* if and only if they are not sequential. For example, "system response" and "listen to beep" are concurrent. (Note that two activities are called concurrent if they could in principle be carried

out simultaneously, even if it happens that one of them is finished before the other starts.)

The duration of the task depends on the durations of the individual activities. Suppose all activities in the network must be completed for the task to be completed (there are other possibilities, of course). An activity network of this form is called a *critical-path network*. The task is completed when and only when all the activities on the longest path through the network are completed. The longest path through the network is the *critical path*, and the duration of the task is the sum of the durations of all the activities on the critical path. For example, in Figure 2 activity durations are listed immediately below the activity inside the vertex (assuming here these durations are constant). The durations of activities along the top path sum to 8; durations of activities along the bottom path sum to 11. Thus, the task duration is 11. If the durations of the activities are constants, then to reduce the task completion time, one must identify the activities on the critical path and minimize their durations. Shortening the duration of an activity not on the critical path has no effect.

***Application: Workstation Design***  When new workstations were designed for operators in a New York telephone company, the hope was that with faster displays and fewer keystrokes the time per call would drop. But, before the new workstations were installed, an attempt was made to determine whether this would actually be the case. Gray et al. (1993) modeled the way the new workstations would be used, extrapolating from videotapes of operators using the old workstations. The modelers used a technique called CPM-GOMS (John and Newell, 1989; John, 1990) to construct networks for the activities in phone calls. (CPM stands for both Cognitive Perceptual Motor and Critical Path Method. GOMS stands for Goals, Operators, Methods and Selection Rules.) The technique is an extension of the GOMS technique for task analysis in terms of goals, operators, methods, and selection rules (Card et al., 1983). The modeling predicted that the time per call would actually increase, and an increase was indeed found in later data from the new workstations.

What the modeling revealed was that activities that were performed more quickly with the new workstations would not shorten by much the overall time to complete the task, because the quicker activities would be going on concurrently with other slow activities. On the other hand, despite the need for fewer keystrokes with the new workstations, some new keystrokes would occur when there were no concurrent slow activities, so the time for the new keystrokes increased the call completion time. Here is an example where a model can be used not only to predict whether an interface will perform as desired but also why, if such is not the case, the performance will be less than desired. Given the right lead time, this in turn can suggest how to redesign the interface so that performance improves before it is implemented.

### 2.1.2   OP Diagrams

It was assumed above when discussing activity networks that the durations of the processes were constant. In practice, durations of activities are random variables, so
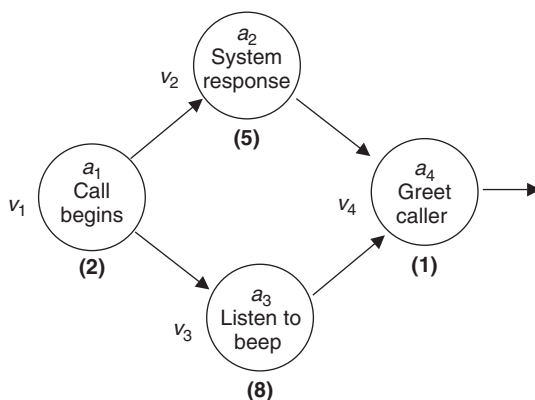


**Figure 2**   Activity network.

one is interested in the probability an activity is on the critical path (i.e., its *criticality*). With random activity durations, calculating the mean and variance of the task completion time is usually intractable, so simulations are carried out with programs such as MATLAB or programs especially designed for human factors, such as MICROSAINT (Laughery, 1985) and the recently developed SANLab (Patton and Grey, 2010). When durations have exponential or gamma distributions, exact formulas can be found with an OP (order-of-processing) diagram (Fisher and Goldstein, 1983). For example, if we let $T_i$ be the duration of activity $a_i$, then for the activity network in Figure 2 the expected value of the completion time is $E[\max\{T_1 + T_2 + T_4, T_1 + T_3 + T_4\}]$. This expectation can be computed easily if the foregoing conditions are met and the task is represented in an OP diagram, a discussion to which we now turn.

Figure 3 shows the beginning part of an OP diagram for a driver reading an electronic variable message sign that presents words one at a time. Each individual word is displayed, perceptually encoded, and comprehended. At any given time, a certain subset of activities will be executing; for example, the comprehension of the first word might go on simultaneously with the displaying of the second word. Such a set of activities executing simultaneously defines a state, and each possible state is represented by a vertex in the OP diagram. In the OP diagram in Figure 3, $w_1$ denotes the display of the first word, $e_1$ its encoding, $c_1$ its comprehension, and so on. In the state represented by the first vertex (labeled $s_1$), the first word is displayed ($w_1$) and the driver is encoding it ($e_1$). It is assumed that the durations of these activities are continuous random variables. Thus, the probability that two activities finish at the same time is zero, so such an event does not need to be represented in the OP diagram. In this case, one of these activities will finish first (not both), and when it does, the state is exited. The driver may finish encoding the first word before its display ends. In that case, in the next state ($s_2$) the first word is still being displayed, and the driver is comprehending it. The transition from the first state to this next state is indicated by an arrow labeled with the activity whose completion leads to this next state, in this case $e1$. The other way to exit the first state is for the display of the first word to finish before the driver has encoded it. The OP diagram indicates that in this case the driver does not complete the encoding of the

first word ($e_1$ appears in parentheses on the arc exiting from the state) and fails at reading the sign.

Every critical-path network can be converted to an OP diagram. In addition, situations such as two activities that must go on one at a time, but in any order, can be represented in an OP diagram but not in a critical-path network. After an OP diagram has been constructed, it can be used to calculate quantities such as the mean and variance of the task completion time and the probabilities of the task completing in various ways, such as success or failure at reading a variable message sign (Fisher and Goldstein, 1983; Goldstein and Fisher, 1991, 1992). In this case, if we let $T_i$ be the duration of activity $a_i$, the expected time to complete the task is no longer the expectation of the maximum of the path durations in the OP diagram. Instead, it is now a probability mixture of conditional expectations, where each conditional expectation is the time on average it takes to complete all activities along a path in the OP diagram. For example, if the top path through the OP diagram were taken, and it consisted of only those states listed ($s_1, s_2, s_3, s_5$), we would want to compute the following conditional expectation: $E[T_1 + T_2 + T_3 + T_5|$ path $(s_1, s_2, s_3, s_5)]$. We would need to weight this by the probability that path $(s_1, s_2, s_3, s_5)$ is taken and then do the same thing for all other paths in the OP diagram. Equations for the calculations may be found in Fisher and Goldstein (1983).

### 2.1.3 Identifying Activity Durations

We assumed above that durations of activities are known if calculations or simulations are needed. Durations are ideally found by observation, as in Gray et al. (1993). Durations of certain common activities are available in the literature (see references in Schweickert et al., 2003). Another source is expert opinion. However, it is often difficult for an expert to produce an accurate estimate of the duration of an activity. Schweickert et al. (2003) proposed that it may be more natural for an expert to produce a rank ordering of the differences in durations between pairs of activities. For example, in a telephone operator's task, an expert may judge that the difference in durations between greeting a customer and pressing the enter key is less than the difference in durations between entering a credit card number and listening to a beep. When the judgments are entered into a multidimensional scaling program
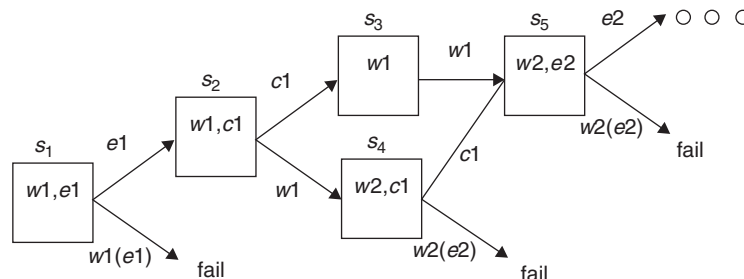


**Figure 3**  OP diagram.

(e.g., Shepard, 1962; Kruskal 1964), scale values for activity durations are produced. If the actual mean and variance of the duration of at least one activity are known for calibration, the scale values can be converted to estimates of mean activity durations. The estimates are only approximations. But Schweickert et al. (2003) found that they can lead to excellent predictions of the criticality and the product of criticality and duration for individual activities in simulations. (Activity durations were assumed to have gamma distributions.)

### 2.1.4  Identifying the Arrangement of Activities

We also assumed above that the arrangement of the activities is known. The activity arrangement in a task is ordinarily found by observation, discussion with operators, and inference. Sometimes an activity network has not been constructed, but a task analysis has been carried out, and the resulting diagram can readily be converted to an activity network (Schweickert et al., 2003; see also Anderson, 1993). In many information-processing tasks, activities are unobservable mental processes such as perceiving and deciding. An analyst may be able to infer their existence or ask about the operator's knowledge of them. Another approach is based on experimentation using the technique of task network inference. With the technique, the activity arrangement is obtained by manipulating experimental factors, such as the number of elements in a display, with the intention of using each factor to influence selectively the duration of a single activity, such as a visual search. The effects of the factors on task completion time provide the information needed to construct a critical-path network or to show that no such network is possible for the factors used (Dzhafarov et al., 2004; Schweickert, 1978; Schweickert and Townsend, 1989; Schweickert et al., 1992; Schweickert, Fisher and Sung, in press). The key idea is that when two activities are influenced selectively, patterns in task completion times differ depending on whether the pair of activities is sequential or concurrent. With this information for pairs of activities, a network can be constructed. After an activity network for the task has been constructed, simulations or calculations can be used to model effects of changes such as aging (Fisher and Glaser, 1996) or equipment modification.

***Application: Variable Message Signs***  Above, using an OP diagram to model a driver reading a variable message sign was discussed (Figure 3). The OP diagram can be used to predict the probability that a driver reads each of the words in the message and therefore understands the signs. In practice, it is often possible to present the message one, two, or three times in the legibility zone and even to vary the duration of each page in a multiple-page message (a message so long that it cannot be presented in its entirety on a single page of a variable message sign). It is by no means clear exactly how long each page should be displayed to maximize the probability that drivers understand the variable message sign when the message is displayed one or more times in the legibility zone. Recent research indicates that displaying the message more than once

increases the likelihood that drivers will understand the message (Dutta et al., 2005). The next step is to use an OP diagram to find the page durations that maximize the probability that drivers understand the message. In this case one would need to add an additional process to the OP diagram, which reflects the time that drivers have to read the message. This will be a function of the driver's speed and the distance from the sign at which the message on the sign first becomes legible. For any given setting of the parameters, one can easily compute the probability that the words on both pages are understood. The optimal page durations can then be estimated by iterating through the space of possible page durations.

## 2.2  Signal Detection Theory

Task analysis is a powerful methodology for understanding what a person does (task description) and interpreting how a person performs (task analysis). A basic distinction in task analysis is often made between resource-limited and data-limited tasks (Norman and Bobrow, 1975). In a *resource-limited task*, such as the ATM menu tasks noted earlier, performance improves as more resources (time in this case) are devoted to the task. If we try to rush the sequence of button pushes, errors are more likely. All of the tasks that were described in Section 2.1 were resource limited. In contrast, *data-limited tasks* do not show increased performance with more resources. These tasks are limited by the quality of the incoming data, so that no matter how many processing resources are employed, the performance (e.g., detecting or recognizing a signal) does not improve. An example of a data-limited task would be trying to hear an important news broadcast on a radio at the limit of reception. If the signal-to-noise ratio is too low, trying to analyze more intensely what was heard will hardly make the signal more recognizable. These data-limited tasks have often been modeled by signal detection theory (Green and Swets, 1966), a discussion to which we now turn.

In a signal detection task, accuracy is the dependent variable. A subject must decide whether or not a weak signal is present. When a signal is presented, it produces neural activation, but the same signal does not always produce the same amount of activation. To make matters worse, an amount of activation usually produced by a signal can sometimes be produced in the absence of a signal by background noise from the environment (or from the nervous system itself). Error-free performance is not possible under these circumstances. According to signal detection theory (SDT), the best that one can do is set a particular amount of activation, call it $x_c$, as a criterion. If the amount of activation present exceeds the criterion, one decides that a signal is present; otherwise, one decides that noise alone was present. The result is a $2 \times 2$ classification of events:

1. If a signal is present and the observer says that a signal is present, the event is called a *hit*.
2. If a signal is present and the observer says that a signal is not present, the event is called a *miss*.

3. If a signal is not present and the observer says that a signal was present, the event is called a *false alarm*.

4. If a signal is not present and the observer says that a signal is not present, the event is called a *correct rejection*.

In the prototypical version of signal detection theory, the activation produced by a signal is normally distributed with mean $\mu_s$ and the activation produced by noise alone is normally distributed with mean $\mu_n$. The variance of the two distributions is assumed to be the same, $\sigma^2$. The more intense the signal, the greater the mean activation produced by it. The greater the difference between the mean of the signal distribution and that of the noise distribution, the more sensitive the observer will be. A measure of sensitivity is $d'$:

$$d' = \frac{\mu_s - \mu_n}{\sigma}$$

The means and variances of the activation distributions, and hence $d'$, are assumed to be influenced by characteristics of the signal and noise but to be beyond the control of the observer.

What is under control of the observer is the location of the criterion. The location of the criterion is frequently specified, not by the value of $x_c$, but by the ratio of the density functions of the signal, $f_s(x_c)$, and noise, $f_n(x_c)$, distributions evaluated at the criterion $x_c$. This ratio is often called the *response bias*. A measure of the response bias is $\beta$, defined as

$$\beta = \frac{f_s(x_c)}{f_n(x_c)}$$

To estimate $d'$ and $\beta$ from data, one ordinarily assumes that the signal and noise distributions are normal, with equal variance, so these quantities are said to be *parametric*.

Signal detection theory is a good example of a mathematical model that can be used as an optimization model, directly influencing practice. The two parameters, $d'$ and $\beta$, describe the actual performance of the task when the distributions of signal and noise are given. But we can go further and use mathematical optimization techniques to find where a person *should* place the criterion (optimum $\beta$) rather than just where a person *does* place it. To do this, we develop an objective function: for example, the long-term expected payoff over many trials. To optimize performance, the observer should adjust the criterion, taking the probability of a signal into account as well as the costs and benefits of the various correct responses and errors (see, e.g., Macmillan and Creelman, 1991). The constraint set is implicit in the model of signal and noise given above.

Specifically, if one can assign values to hits ($V_H$), misses ($V_M$), correct rejections ($V_{CR}$), and false alarms ($V_{FA}$), and if one knows the probability of a signal, $p(s)$ and, by extension, the probability of noise, $p(n)$, the criterion $x_c$ which maximizes the expected gain can easily be found by knowing the optimal $\beta$ where

$$\beta_{op} = \frac{p(n)}{p(s)} \frac{V_{CR} + V_{FA}}{V_H + V_M} \tag{3}$$

Note that this optimum is independent of the actual distributions of signal and noise.

Above, we talked about signal detection theory outside the context of activity networks and more general OP diagrams. We now want to show how one can easily and immediately incorporate the elements of signal detection theory into the framework of OP diagrams. Suppose that a signal is presented and a response obtained. Then one might have only three processes: encoding, decision, and response. However, when a signal is presented, there are two different responses. Correspondingly, when noise is presented, there are two different responses. To model the task, one will need two different OP diagrams, one used when the signal is presented and one used when noise is presented. Consider just the case when the signal is presented. The state in which the decision process completes will now have two transitions associated with that completion, one indicating that the subject responds that the signal is present and one indicating that the subject responds that the signal is absent. The probabilities of these transitions are, respectively, the probability of a hit and the probability of a miss obtained from signal detection theory.

***Application: Inspection Tasks I*** An obvious application of signal detection theory is to data-limited tasks such as matching a paint color to a sample or listening to a car engine to detect a maladjusted tappet. These are inspection tasks and are considered in more detail in Sections 6.1–6.3.

## 2.3 Information Theory

Above, we described tasks in which one uses knowledge about the operation of the encoding and decision processes to assign values to the parameters in an OP diagram, thereby increasing the overall power of these diagrams. Here, we want to describe tasks in which one uses knowledge about the operation of the response selection processes, again to assign values to the parameters in an OP diagram. Specifically, we want to know how response times will vary as a function of the number of different responses that can be made in a given task. As an example, consider an in-vehicle collision-warning system. One could potentially warn drivers not only that a collision was going to occur but also where (in general) that collision was going to occur. As someone concerned with the design of such a system, one would like to know whether drivers will respond most quickly if they are warned that the collision will occur somewhere in front or somewhere behind the middle of the vehicle. Or, instead, should one warn the driver that the collision will occur in front, in back, to the left, or to the right? And, of course, more complex schemes are possible. To understand what needs to be done, some appreciation is needed for the role that information theory can play in this decision.

Generally speaking, one might assume that it is difficult, if not impossible, to decide on a common definition of information, let alone quantify that definition. Yet this is exactly what was done so elegantly by Shannon and Weaver (1949). Briefly, imagine two events, one very likely and one very unlikely. Suppose that the first event is: "The sun will rise tomorrow." The second event is: "The winning lottery number tomorrow will be 978654133." Most people would agree that there is much more information in the second message than in the first. So, let's take as a basic axiom the following: If the probability $p(x_i)$ of a message $x_i$ is greater than the probability $p(x_j)$ of a message $x_j$, the information $I(x_i)$ in message $x_i$ is less than the information $I(x_j)$ in message $x_j$. Rather surprisingly, together with a few other reasonable axioms, it follows necessarily that the information in a message can be written as follows:

$$I(x_i) = -\log p(x_i)$$

Assume that there are $n$ messages in a set $X$ of messages. Then one is frequently interested in the expected information in the message set, what is often called the *uncertainty*:

$$U(X) = E[I(X)] = \sum_{i=1}^{n} I(x_i) \cdot (x_i)$$
$$= \sum_{i=1}^{n} [-\log p(x_i)] \cdot p(x_i)$$

Now, taking this one step further, imagine that we have a set $S$ of $n$ stimuli and a set $R$ of $n$ responses. For the sake of simplicity, assume that $r_i$ is the correct response to stimulus $s_i$. Then we can ask how much information in the stimulus set is transmitted by the responses. In theory, anything between none of the information and all of the information could be transmitted. If a subject always gives the correct response, all of the information in the stimuli is transmitted by the responses. However, suppose that the probability that a subject gives response $r_i$ to stimulus $s_i$ is equal to chance ($1/n$) and, in fact, $p(r_j) = 1/n, j = 1, \ldots, n$. Then, none of the information in the stimuli is contained in the responses since the response that is made is entirely independent of the stimulus that is presented. A measure consistent with these intuitions, defined as the *information transmitted*, is easy to develop and is defined as follows:

$$T(S, R) = U(S) + U(R) - U(S, R)$$

where $U(S, R) = -\sum_{i=1}^{n} \sum_{j=1}^{n} p(s_i, r_j) \log p(s_i, r_j)$.

In light of these developments, Hick (1952) asked the following question: Would the response time in a task be related to the information transmitted? To answer this question, he ran an experiment in which the number $n$ of stimuli shown to participants varied across conditions. Each stimulus was associated with a unique response. The probability $p(s_i)$ that a particular stimulus could occur was simply set to $1/n$. He found a linear relation between the information transmitted and the response time:

$$RT(n) = a + bT_{n+1}(S, R) \qquad (4)$$

The interpretation of this relation is as follows. Suppose that the number $n$ of stimuli were a power of 2 ($n = 2^k, k$ a positive integer) and subjects always responded correctly. Then the information transmitted (assuming no errors) can be shown to be equal to $k$, which is the minimum number of binary decisions it takes to identify one of $n$ stimuli (again, $n = 2^k$). However, the reader will note that the information transmitted is indexed by $n + 1$, not $n$ as one might expect. Hick argued that the respondent making the decision might be choosing among $n + 1$ responses, $n$ of which were associated with a particular stimulus and one of which was associated with the absence of a stimulus. Hyman (1953) observed that the number of responses and information transmitted were well correlated. To determine whether it was the information, not the number of responses, which was controlling response time, he covaried the two and found that the ordering of the response times was consistent with the measure of the information transmitted, not the number of responses.

Predictions of response times from information theory are easily demonstrable in laboratory tasks with a relatively small number of alternatives, perhaps 16 or less (4 bits of information). However, they can be extended successfully to tasks with much higher levels of information per stimulus. For example, Bishu and Drury (1988) showed a good fit of information theory to complex surface wiring tasks in the communications industry, with information per stimulus up to 30 bits.

Finally, and as above with signal detection theory, we want to bring the discussion back to the more general framework of OP diagrams, if only briefly. In a task in which there are no errors (or few) and the number of stimuli in the message set varies across blocks of trials, a simple serial model can be used to represent the latent encoding, response selection, and response execution processes in an OP diagram or other network. The distribution of the duration of the response selection process is one of the parameters in the model. The mean of the distribution can be determined for each number $n$ of stimuli in the message set using information theory. From equation (1) it follows directly that this mean is equal to $bT_{n+1}(S, R)$. The sum of the means of the encoding and response execution processes is $a$.

### Application: In-Vehicle Collision-warning Systems

At the start of this section, brief mention was made of in-vehicle collision-warning systems. The utility of information theory for the design of such systems can now be made more clear. Suppose that participants are asked to indicate as quickly as possible from which direction an alarm has sounded. The first question is whether response times increase as the number of locations from which a warning can sound increases. In a recent experiment this number was varied between two and five (Wallace and Fisher, 1998). Response times

increased linearly as a function of the information transmitted, going up by almost 50% when the uncertainty in the stimulus set was largest. Of course, although times are longer with more alarm locations, subjects know better where to focus their attention when the location of the collision is delineated more clearly. Thus, although with few alarm locations drivers may quickly be able to determine that the collision is in front or in back of them, they will not know where more precisely to look. Additional time will be needed by the driver to find the object with which a collision is imminent. A complete model of the time that it takes a driver to locate the source of a collision will require not only a model of how quickly the driver can locate the general area of concern, but additionally, a model of how quickly within that general area of concern the driver can find the actual object that is creating the collision risk. There is every reason to believe that such a model can be constructed based on related models of visual search (e.g., Arani et al., 1984) and could be used to identify the number of warning locations that will minimize the time it takes drivers to respond appropriately to a threat.

***Application: Mail Sorting***   Mail sorting represents another instantiation of an information-theoretic model. The sorter needs the address of an envelope and sorts it into the correct slot from hundreds of slots in a mail route, each slot representing one address. Hoffmann et al. (1993) used information theory to predict mail-sorting times for Australian Post. In fact, such a model can be manipulated mathematically. Drury (1993) studied a mail-sorting system where part of the incoming mail stream was sorted into the correct order automatically. Ordered mail restricts the choices that are available [e.g., if the previous mail piece went into slot $i$ out of $n$, the information to be processed in the next piece would be a choice between $n - i$ alternatives]. This formed the basis of predicted savings in mail-sorting time from preordering of the mail.

## 2.4   Other Tools

We have described just several of many different tools that might be used to model the latent cognitive processes that govern the performance of participants in both laboratory and field settings. Many of these other tools, including associative networks (Anderson and Bower, 1973), connectionist networks (Rumelhart and McClelland, 1986), and shortest route networks, have their equivalent as OP diagrams and have been discussed elsewhere (Rouse, 1980). Other more general models, ACT-R (Anderson, 1983; Byrne and Anderson, 1998), queuing networks (Liu, 1996; Liu et al., 2006), SOAR (Newell, 1990), and EPIC (Meyer and Kieras, 1997a, b), about which we spoke earlier, can also easily be incorporated in the OP network framework.

## 3   VISUAL AND MEMORY SEARCH

The range of applications of task analyses and activity networks to human factors problems requiring cognition is too large to even begin to catalog, let alone

cover in some detail. However, we feel there is one area that stands centrally in human factors, has an extensive history of mathematical modeling (Townsend and Ashby, 1983), and continues to be crucial, namely search and scanning processes (Wolfe, 2007). Almost every one of us is involved daily in one form of search or another, not the least of which is the search for a particular option on an ATM or PC, or the search through voice mail. Below, models of visual and memory search are discussed and applications of these models are detailed.

### 3.1   Visual Search

The visual search through a display can be as simple as the scanning of an array of symbols presented to a person seated in front of a computer or as complex as the scanning of traffic visible to a driver who may be looking for a particular license plate number in a sea of cars. At the heart of all models of visual scanning are four latent cognitive processes: an encoding $e$ of the information to which attention is being paid; a comparison $c$ of the encoded information with the target; a decision to end the search since the target is present and respond $p$ that such is the case; or a decision to end the search since no target is present and respond $a$ that such is the case.

In theory, there are at least four different ways that one might scan a visual display for a target, the most obvious being a serial scan that terminates when the target is identified (*serial*, *self-terminating*). If there were multiple targets, the scan could not terminate until all stimuli in the display had been identified (*serial*, *exhaustive*). In some cases, users might scan the display in parallel, either stopping when the target was identified (*parallel*, *self-terminating*) or continuing until all stimuli had been scanned (*parallel*, *exhaustive*). It is straightforward to represent the architecture for each of the models and predict the response time when the durations of the latent processes are constant. However, it becomes more difficult to represent the architecture and predict the moments of the response time distributions when the durations of the latent processes are random variables, especially as constraints are added, say, to the number of decision processes in a parallel model that can be ongoing simultaneously.

To give the reader a sense for how the modeling is undertaken, a simple derivation will be made of the expected time that it takes a person to find a target when the search is a serial, self-terminating one and the display consists of an array of symbols (say, letters). Let $E_i$ represent the time to encode the $i$th symbol scanned in the display, whatever it is. Let $C_i$ represent the time to compare the $i$th symbol scanned with the target. The subscript $i$ will be dropped here because it is assumed that the distributions of these times are identical. Let $Y$ represent the time to respond that the target is present and $A$ represent the time to respond that a target is absent. Let $F$ be an indicator random variable that is set to $i$ if the target is identified as the $i$th symbol scanned in the search. Finally, let $P(F = i)$ be the probability that the target is in the $i$th location. Then the expected time, $E[T(\text{present})]$, to find a target when there are $n$

symbols can be written as the weighted sum of the time on average, $E[T(\text{present})| F = i]$, that it takes to find the target in each of the $i$ different positions:

$$E[T(\text{present})] = \sum_{i=1}^{n} E[T(\text{present})|F = i]P(F = i)$$

If the target is in the $i$th position, there are $i$ encoding operations and $i$ comparisons plus one decision to respond that the target is present. Let $e$, $c$, $y$, and $a$ represent, respectively, the expected encoding, comparison, target present, and target absent process durations (these symbols also are used to label the processes; the meaning will always be clear from the context). Assuming that the target is equally likely to be in any one of the $n$ positions [i.e., $P(F = i) = 1/n$], we find

$$\begin{aligned} E[T(\text{present})] &= \sum_{i=1}^{n} \frac{ie + ic + y}{n} \\ &= \frac{(n+1)(e+c)}{2} + y \end{aligned} \qquad (5)$$

The expected time to scan the display when the target is absent is simply $E[T(\text{absent})] = n(e + c) + a$. We can rewrite the expected time to scan the display when the target is present and absent as a linear function of the number of items in the display, as indicated by

$$\begin{aligned} E[T(\text{present})] &= y + \frac{e+c}{2} + \frac{(e+c)n}{2} \\ E[T(\text{absent})] &= a + (e+c)n \end{aligned} \qquad (6)$$

Note that the slope, $(e + c)/2$, of the linear function relating the expected target present response time to the number of stimuli in the display is half the size of the slope, $e + c$, relating the expected target absent response time to the number of stimuli in the display. This is just one of many examples where a mathematical model makes predictions that can easily be tested with actual data. Also note that as long as the assumption that the target is equally likely to be in any one of the $n$ positions on each scan is valid, it matters not in what order or orders the stimuli are scanned.

***Applications: Menu Hierarchies*** In the introduction to this chapter, we explained the role that mathematical models and optimization techniques can play in the design process by referring to the construction of the optimal menu hierarchy for an ATM. We now continue this discussion, specifying here and in detail the exact quantitative procedures that one can use to design the optimal menu hierarchy, not for an ATM, but for a PC. Users today interact with menu hierarchies constantly, whether on their cell phones, at their ATMs, or on their computers at home and at work. Regardless of the technology, it is still the case that the underlying structure of the menu hierarchy has a large impact on the time it takes users to access information at the terminal

nodes. It has been shown that, given some very simple assumptions, one can identify the structure of a particular hierarchy that minimizes the average time it takes unfamiliar users to access the information in that hierarchy (modeling the search behavior of familiar users requires a different set of assumptions). There are two cases. In the first case, the number of menus at each level in the hierarchy is equal to twice the number in the superordinate level, and the number of options in each menu is identical across all menus (Lee and MacGregor, 1985). In the second case, there are no constraints on the initial structure of the hierarchy (Fisher et al., 1990). It is the second case that we consider here.

Suppose that users scan serially through the options in a menu, stopping when they identify the option that leads them to the appropriate next level. Then, it is easy enough to derive expressions for how long on average it will take users to identify a terminal option. What is not so obvious is what alternative structures one should consider. For example, take the menu in Figure 4a. Call this the *seed hierarchy*. It is the menu that a design team might produce, one that is most detailed. Five menus are labeled, beginning at the top, M(1,1), and so on. There are two options in each of the five menus. For example, in menu M(2,1) there are two options, options 3 and 4. There are three *terminal menus*, M(3,1), M(3,2), and M(2,2), and six *terminal options* (7, 8, 9, 10, 5, and 6) in these menus. Ideally, one would like to examine the complete space of semantically well-defined hierarchies that can lead to the retrieval of information from the six terminal options. However, there is currently no way automatically of generating all such semantically well-defined hierarchies. Still, it is possible to identify a large subset of the semantically well-defined hierarchies. Specifically, suppose that a *nested hierarchy* is defined as one that is formed from the seed hierarchy by replacing one or more options in a menu with all of the terminal options that come beneath it. Examples include the nested hierarchies in Figures 4b and c. There are six nested hierarchies that can be formed from the one seed hierarchy. In a slightly more complex example, assume that there are 64 terminal options, 2 options in the top-level menu, 2 options in each of the second-level menus, and so on, down to the 64 terminal options in each of the 32 sixth-level menus. Then it can easily be shown that there are over 1 million different nested hierarchies, each semantically well defined. Experimentally, it would be impossible to search this space exhaustively. However, a recursive algorithm can easily be implemented which identifies the hierarchy that minimizes the expected terminal option access time (Fisher et al., 1990). It is an example of the application of dynamic programming, one of the optimization techniques to which reference was made earlier.

Very briefly, the time on average that it takes to find a terminal option from a nonterminal menu can be written recursively as the time on average that it takes to find the option in the current menu that leads to the terminal option plus a probability mixture of the time on average it takes to find the terminal option from each of the menus that can be reached from the current menu.
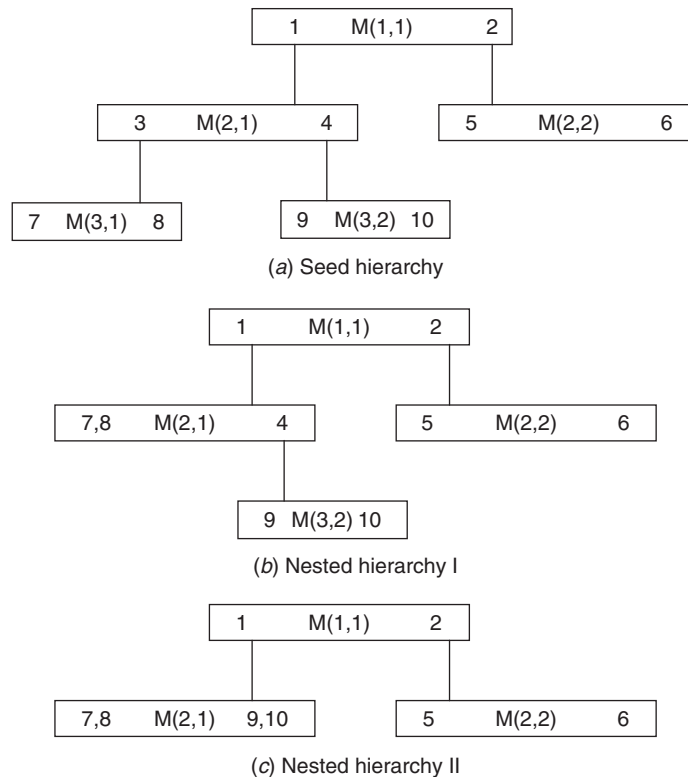
**Figure 4** Menu hierarchy.

This recursive formula is then implemented in computer code. Each menu in the hierarchy is represented on the computer as a node in a linked list, with the link from each option in a menu pointing to the menu which can be reached from that option. To search the entire space of nested hierarchies, one compares the time on average it takes to reach a terminal option from the current menu in the seed hierarchy with the time on average that it takes to reach a terminal option when all terminal options are nested in the current menu. If the time on average it takes to reach the terminal option from the current menu with the nested terminal options is shorter than the time on average that it takes to reach the terminal option with the options left unnested, one can replace the seed hierarchy with the nested hierarchy so constructed. In this way, the search space is reduced dramatically. Here is an example where optimization techniques, not just mathematical models, play a critical role in the design process.

### 3.2 Memory Search

In a memory search task, a participant is given a list of stimuli to memorize (say, four digits). The number of digits in the memory set is referred to as the *memory set size*. After memorizing the digits, the experimental trial begins. A probe digit is then displayed. The participant must indicate whether the probe is in the memory

set. Response time is graphed as a function of the memory set size. The best fitting lines relating the response times to memory set size for the case where the target is and is not present are often roughly parallel. The serial, self-terminating model cannot easily explain such parallelism, as can easily be seen from equation (6), if, among other things, the assumption that the distributions of the process durations associated with each item in the memory set do not depend on the identity of the item is generalized across memory sets of different sizes. However, a serial, exhaustive model can easily explain the parallelism (Sternberg, 1966, 1975; Townsend, 1972). Note that in memory search, unlike visual search, only the probe digit needs to be encoded since all of the items in the memory set have already been encoded. Thus, the formula for memory search will include only one encoding.

This might be seen as the rather tidy end to the puzzle of how it is that items in memory are scanned. However, the resolution depends on a number of critical assumptions, one of which, as we just stated, is that the distributions of the first and second comparison times in the serial, self-terminating model are identical. To see that this assumption is required, we need to refer to OP diagrams again and show that a parallel exhaustive model can mimic a serial, self-terminating one. The parallel exhaustive model is represented in Figure 5. To begin, imagine that there are two items in the memory
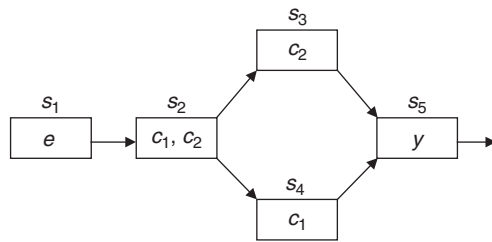
**Figure 5** OP diagram: memory search.

set and one target on the screen that matches one of the items in the memory set. Then the target is first encoded (state $s_1$ in Figure 5). If the search were a parallel exhaustive one, the two items in the memory set could be compared in parallel with the target ($s_2$). In this example, let $c_i$ represent the comparison time of the $i$th memory set item with the target. If $c_1$ finished before $c_2$, execution of $c_2$ would continue by itself ($s_3$), and finally, the participant would respond ($s_5$). Alternatively, if $c_2$ finished before $c_1$, execution of $c_1$ would continue by itself ($s_4$) and again the participant would respond ($s_5$). Note that because the search is exhaustive, both items in the memory set need to be compared before the response is executed. The rather surprising finding here (Townsend, 1972) is that the expected response time with this parallel, exhaustive model can mimic the expected response time with a serial, self-terminating model. To see this, imagine a serial, self-terminating model where the first comparison is relatively fast (the system is not fatigued) and the second is relatively slow. In particular, the distribution of the first comparison time in the serial, self-terminating model is equal to the distribution of the time spent in state $s_2$ and the distribution of the comparison time of the second (first) stimulus in memory with the target when it is processed last is equal to the distribution of the time spent in state $s_3(s_4)$. Finally, imagine that in the serial, self-terminating model the probability that the comparison of the target with the second item in memory occurs after the comparison of the target with the first item in memory is equal to this same probability in the parallel, exhaustive model (i.e., the probability of the path represented by the transition from $s_2$ to $s_3$). Then the expected response times of the parallel, exhaustive and serial, self-terminating models will be identical. Fortunately, more detailed tests exist which can resolve the mimicking (Schweickert, 1978; Townsend and Wenger, 2004). This example very nicely points out the importance of quantifying models wherever possible, thereby reducing needless testing or exploration of alternatives that turn out not to be identifiably different.

***Application: Toxicology*** One of the more elucidating applications of memory search models was developed by Smith and Langolf (1981). They asked a simple question: Could low levels of mercury exposure produce effects on the speed of processing in people who were otherwise asymptomatic? To test this hypothesis,

they had chlor-alkali workers who had been exposed to different levels of mercury perform a simple memory search task, estimating for each person the comparison time (among other quantities). They then regressed the estimated comparison time for each person on the corresponding level of exposure of that person to mercury. There were significant effects of the mercury level on comparison times. From a practical standpoint, the most heavily exposed workers had an increase of 100% in their scanning times, suggesting a serious reduction in short-term memory capacity (Baddeley, 1992). Interestingly, this model developed for purely theoretical purposes has become some 30 years later a useful tool for uncovering neurobehavioral toxicity (Fiedler et al., 1996).

## 4 VIGILANCE

When signals occur rarely, a common finding is that observers tend to miss more signals after some time on task. This is called a *vigilance decrement*. Often, there is a decrease in false alarms as well; that is, there is a decline in the total number of reports of a signal, correct or incorrect [see Davies and Parasuraman (1982) and See et al. (1995) for reviews]. It is natural to use signal detection theory to determine whether, as time goes by, observers become less sensitive to signals, or less prone to say "signal," or both. Results are complicated, but as a rough guide, Parasuraman and Davies (1977) found that sensitivity (measured, e.g., by $d'$) tends to decline when one stimulus occurs at a time (so identification relies on memory) and stimuli occur relatively frequently. The criterion (measured, e.g., by $x_c$ or $\beta$) tends to increase when stimuli are presented simultaneously (in a same or different task) or stimuli occur rarely.

As described earlier, one reason that signal detection theory is useful is that it divides the observer's processing into encoding and decision and provides separate measures characterizing each, $d'$ and $\beta$. Variables such as the probability of a signal are predicted, and often found in data, to significantly change the value of the decision parameter $\beta$, which represents the observer's choice of where to put the criterion, but not to significantly change the value of the encoding parameter $d'$, which represents the observer's sensitivity to the signal. However, it is readily acknowledged that the assumptions underlying the calculation and interpretation of $d'$ and $\beta$ are not met exactly.

To investigate alternatives, See et al. (1997) compared the performance in vigilance tasks of two measures of sensitivity and five measures of response bias, with an emphasis on the latter. [For the formulas, see See et al. (1997).] The two measures of sensitivity were the parametric $d'$, which was discussed above, and a widely used nonparametric analog, $A'$. The authors report that the two measures were functionally equivalent. They were highly correlated over subjects, each declined with time on task, and neither was influenced by signal probability or payoff scheme, factors thought to influence the response bias rather than the sensitivity. Thus, $d'$ seemed to function as expected.

However, such was not the case for response bias. For response bias, two parametric measures $\beta$ and $c$, and three nonparametric measures, $B'$, $B'_H$, and $B''_D$, were compared. The authors report that $\beta$ did not perform well; in particular, it was the least sensitive measure to signal probability and payoff scheme. The best performing measure was $c$ (Ingham, 1970),

$$c = \frac{x_c - (\mu_s + \mu_c)/2}{\sigma}$$

(The measure $c$ is the $z$ score for the criterion, $x_c$, but calculated with respect to a mean halfway between the means of the signal and noise distributions.) Of the nonparametric measures, the best performing was $B''_D$ (Donaldson, 1992),

$$B''_D = \frac{(1 - H)(1 - F) - HF}{(1 - H)(1 - F) + HF}$$

where $H$ is the probability of a hit and $F$ is the probability of a false alarm. Both $c$ and $B''_D$ were sensitive to signal probability and payoff scheme, both indicated a predicted change in bias over time, and both functioned well even when performance was at chance level. The data of See et al. (1997) satisfied a test for normal distributions with equal variance, and it would be useful to know whether their conclusions hold in other situations.

Balakrishnan (1998a,b) has argued that violations of assumptions underlying use of parameters $d'$ and $\beta$ could be causing misleading interpretations of data, but these pass unnoticed because of the apparent robustness of the measures under signal probability or payoff manipulations. Signal detection theory assumes that the distributions of activation produced by a signal or by noise are not influenced by the probability of a signal. This may be true for the distribution of a single sampled amount of neural activity, such as when signals are rare. But neural activation may be extended over locations and over time when signals are frequent. In this case, the observer may sample several amounts of activation. The probability distribution of a sample depends on the sample size. Hence, a change in the amount of activation sampled would be a change in encoding produced by a variable (the probability of a signal) thought to influence only the decision. (Note that if several samples are taken and the results combined to produce a better decision, we have moved from a data-limited to a resource-limited task. How the data are combined determines how the discriminability changes with the number of samples, or more generally with the time over which samples are taken. This would be another example of a speed–accuracy trade-off.) Whether this change is registered as a change in $d'$ or $\beta$ or both would depend on the forms of the underlying distributions and the way the size of the sample of observations is determined. To avoid such problems, Balakrishnan (1998a) proposed that distribution-free measures of response bias be used, and he developed new measures based on confidence ratings. Analysts using signal detection theory often ask observers to give a number indicating their confidence that a particular stimulus was a signal or noise. For example, response 1 would indicate very low confidence that the stimulus was a signal, and response 8 would indicate very high confidence that the stimulus was a signal. Note that with this procedure the analyst does not know what the observer would actually say if asked whether the stimulus was signal or noise. Balakrishnan proposed modifying the procedure slightly, so the observer produces a judgment about signal or noise together with a confidence rating that the judgment is correct. For example, response 1 would indicate a judgment of noise with very high confidence, and response 5 would indicate a judgment of noise with very low confidence. Similarly, response 6 would indicate a judgment of signal with very low confidence, and response 10 would indicate a judgment of signal with very high confidence.

The modified procedure leads to Balakrishnan's (1998a) distribution-free measures of response bias. An overall measure of bias is $\Omega$, the total of the amount of bias at each rating having a bias (not all ratings have a bias). For example, suppose that the rating 5 stands for "with low confidence, the stimulus is noise" and the rating 6 stands for "with low confidence, the stimulus is signal." For simplicity, suppose that the probability of signal equals the probability of noise equals $\frac{1}{2}$. Suppose that when 5 is used, three-fourths of the time it is used for noise and one-fourth of the time it is used for signal. Then there is no bias at rating 5. It is intended to indicate noise, and it usually does. It contributes nothing to $\Omega$. But suppose that when 5 is used, one-fourth of the time it is used for noise and three-fourths of the time it is used for signal. Then there is a bias at rating 5. It contributes to $\Omega$. Again, $\Omega$ is summed over only those ratings that have bias.

The amount of bias at rating 5, when it has a bias, is the total proportion of trials for which the rating 5 was used. Suppose that there were 200 noise trials and 200 signal trials, 400 total. Suppose that rating 5 was used 6 times for noise stimuli and 30 times for signal stimuli. The amount of bias contributed by rating 5 is $36/400 = 0.09$. An overall measure of bias is $\Omega$, the total of the amount of bias at each rating having a bias.

Using the measure of bias $\Omega$ in a vigilance task, Balakrishnan (1998b) found no evidence of bias in the decision rule for either relatively frequent or relatively rare signals (i.e., values of $\Omega$ were close to zero for probability of a signal 0.5 and 0.1). On the other hand, there was evidence for increased sensitivity for more frequent signals, indicated by increased $A'$. In other words, signal frequency appears to influence the encoding rather than the decision, the opposite of the usual signal detection theory interpretation. The performance measures themselves do not indicate how the processing was done, but Balakrishnan points out that the results are consistent with models in which the subject samples repeatedly from the stimulus, rather than just once. For rare signals the hypothesis of equal variances for signal and noise distributions can be rejected for the data of Balakrishnan (1998b). Instead, the variance is large for rare signals. For a critique of the

approach of Balakrishnan, see Mueller and Weidemann (2008).

***Application: Inspection Tasks II*** In any particular application, the payoffs are maximized when the criterion is set as indicated in equation (3) provided that the assumptions of signal detection theory are met. If the assumptions are met, then to improve performance, one would train observers to locate their criteria appropriately. However, if, contrary to assumption, signal frequency influences encoding by changing the size of the sample the observer takes from the stimulus, then to improve performance, training would emphasize taking time for adequate observation of the stimuli.

## 5 INSPECTION

Whether unaided manual or partially or fully automated, test and inspection tasks abound. The most obvious examples are from manufacturing quality control with products ranging in complexity from apples to Apple computers. Other examples come from checkout procedures in spacecraft, from aircraft structural inspection, from airport security, and even from inspection of restaurants for compliance with hygiene regulations. What these inspection examples have in common is that their input is an item whose state is unknown (apple, restaurant) and their output is the same item whose state has been determined. Unfortunately, not all state determinations are perfect: there are inspection errors, as noted in Section 2.2. In the current section, our interest is in mathematical models of inspection and their use in performance optimization. For a chapter length treatment of test and inspection, see Drury (2001).

As with any modeling activity the starting point should be a task analysis (see Section 2.1). Many detailed task analyses of inspection have been performed, resulting in a generic function-level model (e.g., Drury, 2001): initiate, present, search, decision, and response. Task analyses of specific inspection tasks go much deeper than this to provide insights and best practices (e.g., Drury, 1999). Here, however, we restrict ourselves to what are usually seen as the most difficult functions: search and decision. These are often the functions having the lowest reliability (Drury et al., 1997) and taking the greatest time to complete. Each has several useful mathematical models (e.g., visual search theory, Section 3.1) and opportunities for optimization making them appropriate instances for the current chapter. In addition, the models can be combined to give a more integrated view of the entire inspection task and can be used where parts of the task are automated. Both of these extensions are considered after models of search and decision are given individually in the context of inspection tasks.

### 5.1 Visual and Memory Search: Alternative Model

The visual search networks in Section 3.1 were developed to model rapid search processes, where each distracter is compared with the target (or target set) until a match is found. In contrast, inspection tasks often involve large and complex objects (e.g., circuit boards or aircraft internal structures) when the search process takes place much more slowly (many seconds or even minutes) and there are no distracters as such, just other elements that may or may not have a defect (e.g., an IC chip placed backward or a crack in aircraft structure). Here a model of search as a sequence of eye fixations appears more appropriate. Such models have existed for many years and are based on the following facts:

1. Visual information is available only when the eye is stationary or tracking a moving object. These fixations typically take between 0.2 and 1.0 s and the rapid saccadic movements between fixations preclude visual information intake.

2. In a single fixation, the probability of detecting a target falls off with the angle between the target and the optic axis. This means that a target is detectable only (with a given probability) in an area around the optic axis known as the *visual lobe*. Note that the visual lobe is not the fovea: Lobe size depends on target–background contrast and can range from subfoveal size for very difficult targets to almost the entire visual field of view for extremely easy targets.

A major preoccupation of visual search modelers has been how successive fixations are chosen from the visual field to perform the search task. The general consensus is that the fixation sequence arises partly from top-down factors (e.g., a predetermined search sequence based on the inspector's experience) and partly from bottom-up factors (e.g., a potential target at the periphery of vision, leading the next saccade to fixate that point). Models of this type are available (e.g., Wolfe, 1994).

If we treat the bottom-up information as essentially an "end game" to confirm a target, the top-down aspect has typically been modeled as either a random process or a systematic process (Morawski et al., 1980). A random process is characterized as having no memory for previous fixations, while a systematic process assumes perfect memory. In fact, a more general model of partial memory was devised by Arani et al. (1984), with the random and systematic models as special cases. They showed that memory has to be almost perfect to invalidate a random model, collaborating many studies that have fitted both models to the data and found adequate fits for the random model.

The random model assumes that each fixation $i$ is chosen randomly from a set of possible fixations (i.e., sampling with replacement). From this model it is easy to derive the cumulative search time distribution [i.e., the probability $P(t)$ that the target will be located at or before time $t$] as

$$P(t) = 1 - e^{-\lambda t}$$

Here, the parameter $\lambda$ is a constant incorporating lobe size and fixation duration information. In fact, both the mean and the standard deviation of this exponential
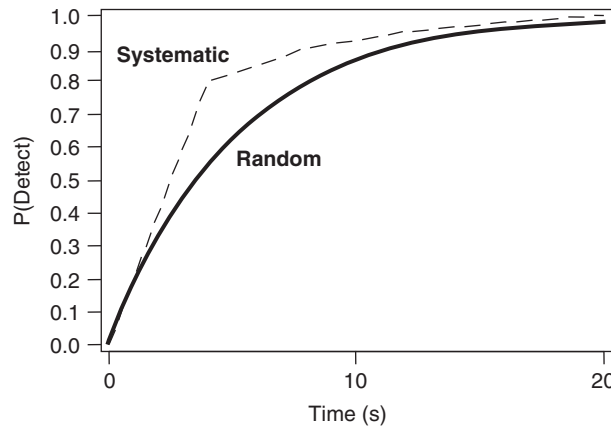
**Figure 6**  Cumulative probability distribution of a random model.

distribution are equal to $1/\lambda$. If the model is valid (and it typically is), a useful deduction from the model is that the time to detect a target will be extremely variable. A typical cumulative probability distribution of a random model is shown in Figure 6.

Note that the process has diminishing returns in that the expected gain from continued search decreases. This is the basis for an optimization model of stopping time in search. The model was developed first in Tsao et al. (1979) and developed much more fully in Chi and Drury (1998). If the value of detecting a target is $v$, the probability of an item having a target is $p$, and the cost of a unit of inspection time is $k$, we can find the three outcomes of a search task with the probability and value of each: (1) the target is present and found, (2) the target is present and not found, and (3) the target is not present. No false alarm is logically possible for a search task: Either the target is found or it is not. That is not to say that false alarms do not occur occasionally in searches, rather that the stopping models assume that the search terminates with a target detection or a failure to detect. If we multiply probability by value and sum over all outcomes, we have the expected value of the search task. This gives, after simplification,

$$\text{value expected} = -kt + vp(1 - e^{-\lambda t})$$

Setting the first derivative with respect to $t$ at zero to maximize expected value gives

$$k = vp\lambda e^{-\lambda t_{\text{opt}}} \qquad t_{\text{opt}} = \frac{1}{\lambda} \ln \frac{\lambda vp}{k}$$

Note that $t_{\text{opt}}$ increases when $p$ is high, $v$ is high, and $k$ is low. Thus, a longer time should be spent inspecting each area where (1) there is a greater prior probability of a defect, (2) there is a greater value to finding a defect, and (3) there is a lower cost of the inspection.

This describes a possible optimum behavior in a search task where a single target can occur. It has been verified by Drury and Chi (1995) as a reasonable model of actual inspection behavior. For a slightly different model, where the operator covers the search field in overlapping fixations, an equivalent optimization model was devised by Bavejo et al. (1996). That model described human field-of-view movements rather than the (untested) eye movements very well.

Extensions have been made to search models to cover multiple instances of a target (Drury and Hong, 2000) and multiple instances of multiple target types (Hong and Drury, 2002). Our optimization model of search can also be applied to multiple targets of the same or different types. Hong (2002) derived and validated such a model against human search time data with good results.

### 5.2  Decision Model: SDT Revisited

After completing the search function, the inspector will now have either (1) not found a target, in which case the item is by definition good, or (2) have found one (or even more than one) target that needs to be assessed for acceptance or rejection against a standard. The decision process (2) thus arises only when a potential target (an *indication* in nondestructive inspection terminology) has been found. At this point, any model that has two states of the item (defect, no defect) and two decision outcomes (accept, reject) is appropriate. SDT (Section 2.2) certainly meets this criterion and has a long history of application to inspection tasks. For example, Drury and Addison (1973) found that increased feedback in a glass inspection task raised the discriminability $d'$, and the benefit persisted over many months of measurement. In the 1970s and 1980s it became quite fashionable to apply SDT to inspection: for example, the studies reported in the book by Drury and Fox (1975). Legitimate warnings were raised about the use of a parametric form of SDT: for example, the assumption of normal distributions of equal variance (Megaw, 1979).

Optimization aspects of the SDT model, for example, the choice of optimum criterion $\beta_{\text{opt}}$, are directly testable. In noninspection decision tasks, the general finding is that $\beta_{\text{opt}}$ does not change as rapidly as it should with changes in the cost−value structure or changes in the a

priori probability of a defect. This has become known as the *sluggish beta* phenomenon. It arises if the decision maker does not take all of the available information into account in reaching a decision (i.e., the human as a degraded optimizer). This really calls into question whether any model of the human as a maximizer of expected value has any validity and eventually led to much more general models of the human as a satisficer rather than an optimizer, most famously the model of Newell and Simon (1972). However, if we treat the human as a degraded optimizer in the decision aspects of an inspection task (Chi and Drury, 2001), reasonable agreement with performance data is found. The warning is raised, however, that people may not be optimizers in the mathematically strict sense in inspection decisions, perhaps because they do not deal well with low-probability events (e.g., finding a bomb in an airline passenger's bag) or large costs (e.g., the tragedy of losing an airline to terrorism).

### 5.3 Reintegration of Search and Decision

Inspection tasks typically include both search and decision components. For example, Drury (2002) showed that a single unified model (essentially the generic functions, initiate, present, search, decision, and response, listed above) described all of the various airport security inspection processes. Some processes have no search (e.g., listening to a car engine for a loose tappet), whereas some have no decision (e.g., inspection of jet engine hubs for cracks where any crack must cause rejection). These are, however, the exception. The integration of models such as visual search and SDT involves issues beyond each individual model.

An early integration example was by Drury (1973), who collected many studies of the "speed effect" in inspection [i.e., the speed–accuracy trade-off (SATO)]. With visual search following a random cumulative model [e.g., equation (5)] and SDT providing the ultimate levels of type 1 and type 2 errors when search is complete, the SATO model shows the probability of a hit increasing over time with diminishing returns to reach some ultimate value less than 1.0. The probability of a false alarm for this model starts at zero for very short inspection times and gradually increases with inspection time, leveling at a different value. This model fitted much of the available SATO data and was interpreted as evidence for a search-plus-decision model. Later, Chi and Drury (2001) tested optimization aspects of this model against human performance in a task of inspecting circuit boards, again finding good agreement.

The second use of an integrated model is in diagnosis of inspection error. At its simplest the search-plus-decision model shows that search alone cannot produce false alarms. Hence, if there is only a search process, false alarms are logically excluded, and in practice will be extremely rare, as was shown to be the case by Drury and Forsman (1996). Where both search and decision occur, it is possible to separate the errors from the two functions with some additional effort. In a task of inspecting jet engine bearings, Drury and Sinclair (1983) used the fact that search was visual whereas decision was tactile to differentiate between search errors and decision errors. They found that search performance was poor but consistent across inspectors, whereas decision performance varied widely among inspectors. Their analysis led to the development of a successful training–retraining program for these inspectors (Drury and Kleiner, 1990). In another study of aircraft structural inspection, Drury et al. (1997) were able to use videotape analysis to separate the two functions, with findings very similar to those of Drury and Sinclair.

Finally, an integration of the two models has implications for analysis of the results of all inspection tasks. When SDT was first used for vigilance tasks, search was not recognized. However, many of the misses in inspection arise from the search process, not from a bias toward acceptance in the decision process. Hence, interpreting overall inspection results in terms of SDT is erroneous unless no search is involved. This is what Wiener (1975) called *jumping off the d-prime end*, and it is a legitimate criticism of early inspection modeling work (e.g., Drury and Addison, 1973). Unfortunately, this misinterpretation still happens. Again, the moral is that there is nothing as valuable as a good model or as misleading as an inappropriate model.

***Application: Automated Inspection*** If we can model the human inspector with reasonable success, how can we extend this modeling to situations where human and automation perform inspection tasks jointly? There are now excellent automated alternatives to some aspects of inspection; see, for example, the detailed review in Drury (2000). How can we incorporate human and automation models into overall inspection models to derive appropriate levels of automation (cf. Parasuraman et al., 2000)? Part of the problem is that most papers on automated inspection denigrate human roles, emphasize the wonders of algorithms, and often provide data only on probability of detection, ignoring false alarms.

The most obvious first step to explain the integration of human and automated inspection is to compare their relative merits directly. Drury and Sinclair (1983) examined an automated system for jet engine bearing inspection using the same measures of performance as were used for human inspectors, in this case ROC curves. Their conclusion was that neither human nor automation was particularly effective, leading to recommendations to improve the automated system and to the subsequent human training program (Drury and Kleiner, 1990).

A more comprehensive step was taken by Hou et al. (1993), who examined search and decision separately for human and automation. They use an SDT measure of discriminability (*A*) as well as inspection speed to compare human and algorithmic alternatives for each function. Their conclusion was that both purely automated systems, as well as the unaided manual system, were inferior to hybrid human–algorithm systems for circuit board inspection. We may be able to compare different human and automation hybrids by direct measurement, but true a priori allocation of function will come only when we can predict from models of the alternatives which hybrid systems to build and test.

## 6 DUAL TASKS

We have been describing tasks in which a person makes just one response in a given situation, that response depending on the stimulus ensemble presented to the person. However, the real world sometimes provides much more challenging situations. For example, suppose that a sign tells a driver to change lanes and then a car cuts in front. The responses are to turn and to brake. Typically, the time to respond to the second stimulus is greater than it would be if it had been presented alone. This and many other findings can be explained by a model of *dual-task* performance originally proposed by Davis (1957) and studied extensively since (e.g., McCann and Johnston, 1992; Pashler, 1998).

The model is in Figure 7. Stimulus $s_1$ is presented, followed after a brief interval by stimulus $s_2$, with required responses $r_1$ and $r_2$, respectively. Each stimulus requires perceptual processing, denoted $a_1$ or $a_2$ as appropriate, cognitive processing, denoted $b_1$ or $b_2$, and motor preparation processing, denoted $c_1$ or $c_2$. The interval between the stimuli is denoted SOA (stimulus-onset asynchrony). Perceptual and motor processing of either stimulus can go on concurrently with any other processing. But cognitive processing (response selection) for the two stimuli is sequential, so process $b_1$ must be completed before process $b_2$ can start. The delay in responding to the second stimulus is due to $b_2$ waiting for $b_1$ to finish.

Clearly from the bar chart in Figure 7 the processes can be represented as activities in a critical-path network. Knowledge of the activity arrangement can be obtained by selectively influencing activities (Schweickert, 1978; Schweickert, Fisher and Sung, in press) using the method of task network inference discussed earlier. Analysis of selective influence for this particular model is sometimes called *locus-of-slack logic* (e.g., McCann and Johnston, 1992). Much is known about where factors such as stimulus quality, display size, arithmetic difficulty, and so on, have their effects. For example, in a dual task by Johnston and McCann (2006), the first stimulus was a tone to be judged as higher or lower than a reference tone. The second stimulus was a trapezoid representing a runway. Its angle was to be judged as higher or lower than that of a reference trapezoid presented shortly earlier (corresponding to a judgment about approach to the "runway"). If the tone judgment was difficult, the time to respond to the trapezoid increased. Data indicated that tone judgment and trapezoid judgment were at locations $b_1$ and $b_2$, respectively, in the model. For reviews and critiques
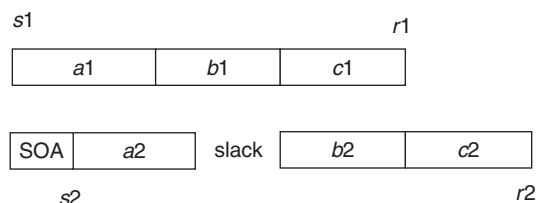
of the model, see Pashler (1998), Logan (2002), and Townsend and Wenger (2004). The model assumes only one response selection process goes on at a time, and it is worth considering why. The two major models of response selection are accumulators and random walks. With each, a stimulus is presumed to transmit information over time. Incoming information is classified as favoring one of the possible responses. In an accumulator, information favoring a response increases the activation of that response. In a random walk, information favoring a response increases the net activation of that response and decreases the net activation of every other response. In both models, for each response a criterion has been set. As soon as the activation for some response reaches its criterion, that response is made.

For a given stimulus, every possible response is made with some probability and there is a probability that no response is made. The behavioral data consist of those probabilities, together with the probability distributions of the response times. It is helpful to realize that an accumulator model can always be constructed to account perfectly for the behavioral data (Dzhafarov, 1993). Hence, more import than goodness of fit is whether experimental factors can be explained coherently. On the whole, results are as one would expect; for example, stimulus quality ordinarily influences the rate at which activation increases and payoffs ordinarily influence criterion values. For a review, see Luce (1986).

A single neuron begins firing when the algebraic sum of the excitation and inhibition reaching it exceeds a threshold, so neural resources needed for an accumulator or a random walk in itself seem small. But with either an accumulator or a random walk, the system must be set by (1) selecting possible responses, (2) forming temporary associations between anticipated incoming bits of information and the responses they favor for the situation, and (3) setting the values of the response criteria. The resources for assembling and maintaining the settings may be considerable. Although evidence is indirect, there is growing opinion that the settings constrain only one decision to be made at a time

Many have considered whether two response selection processes can go on concurrently, at least sometimes. In the EPIC model (Meyer and Kieras, 1997a,b) a person has the option of executing two response selection processes concurrently and without interference. Most models allowing concurrent response selection assume that it is slower when concurrent (e.g., Navon and Miller, 2002; Tombu and Jolicoeur, 2003).

### Application: Scheduling Stimulus Presentations

For designing displays, one useful finding is that a delay in the onset of a second stimulus need not delay the response to the second stimulus or might only delay it by a small amount (e.g., Smith, 1969). In other words, the perceptual processing of the second stimulus may not be on the critical path to the second response and so may have slack. It is also useful to note that there is evidence that humans are able to control the order of the cognitive processes; that is, they can control whether $b_1$ is executed before $b_2$ or vice versa (Ehrenstein et al., 1997).



**Figure 7** Dual-processing network.

There is a relevant finding from scheduling theory. Suppose an activity is faster when executed alone than when executed concurrently with another activity, and the goal is to minimize the average of the times at which each task is finished with respect to the same time-zero starting point. The optimal schedule is usually to allocate all the capacity to one activity and then allocate it all to the other activity (i.e., to schedule the activities sequentially rather than concurrently) (Conway et al., 1967). To see this, suppose that activities *a* and *b* each take one unit of time if executed alone and two units of time if executed concurrently. Suppose that they are both ready to start at time zero. The average of their completion times with respect to a time-zero starting point is 1.5 if they are sequential and 2 if they are concurrent. In other words, humans may schedule response selection processes sequentially not because they must but because it is optimal.

If there is a choice as to when to display stimuli, having the second stimulus appear early would seem to do no harm and may be helpful. There are two potential problems with such a procedure. First, as soon as a stimulus is presented, irrelevant information from it may be sent to processes for the other stimulus (crosstalk) (see, e.g., Hommel, 1998). This may lead to increased response times and errors. Second, presenting the stimuli close together in time may lead to parallel processing, which can be more inefficient if, as noted above, the goal is to minimize the average completion time of two processes measured from the same starting point.

## 7 PSYCHOMOTOR PROCESSES

As noted above, processing time in a task can roughly be categorized as perceptual, cognitive, and motor, and the largest of these is often motor time. The study of movement is interdisciplinary, but as with sensation, the physics of the system is an integral part of the modeling. For an introduction to models, see Jagacinski and Flach (2003), and for a discussion of controversies, see *Controversies in Neuroscience, I: Movement Control* (Editors, 1992).

One of the simplest models to lead to a reasonable approximation of human movement is the mass, spring, and damper in Figure 8 [see Crossman and Goodeve (1963) for an early presentation]. A force $F(t)$ is applied over time by an agonist muscle to a limb of mass $m$. A restoring force is produced by an antagonist muscle represented by a spring. (The agonist could be represented by a spring as well.) Another force is damping due, say, to friction from sliding across a mouse pad or to an internal source such as the antagonist muscle itself or a joint. (The damper is illustrated as a piston in a cylinder filled with oil.) We consider a horizontal movement rather than a rotation through an angle; an analysis of a rotation would not be very different.

Let the position of the limb at time $t$ be $x$, with position zero at time zero. By Hooke's law, the spring produces a force proportional to its displacement from its equilibrium position, $q$. The direction of the force
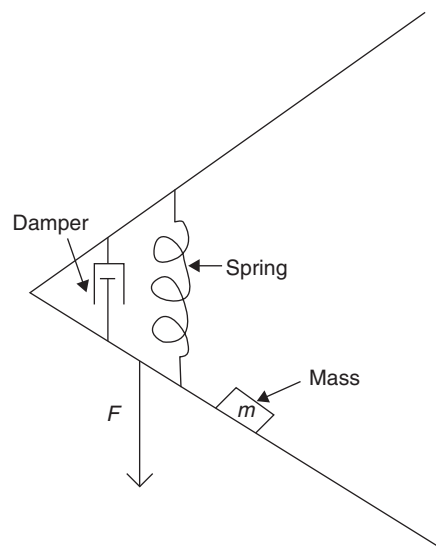


**Figure 8** Mass, spring, and damper.

depends on whether the spring is stretched or contracted. The direction is opposite to the direction of the spring's displacement from its equilibrium position, so the force due to the spring is $-k(x - q)$. Note that if the agonist is represented by a spring-producing force $-k_1(x - q_1)$ and the antagonist is represented by a spring-producing force $-k_2(x - q_2)$, the sum of the two forces is $-(k_1 + k_2)[x - (k_1 q_1 + k_2 q_2)/(k_1 + k_2)]$. That is, the model with two springs can be replaced by an equivalent model with a single spring. The force due to the damper is proportional to the velocity but in direction opposite to it, that is, $-bd\,x(t)/dt$. By Newton's second law, the sum of all the forces equals the mass times the acceleration. That is,

$$m\frac{d^2x}{dt^2} = F(t) - k(x - q) - \frac{bdx}{dt} \qquad (7)$$

where $F(t)$ is the force applied by the agonist muscle and $-k(x - q)$ is the force applied by the antagonist muscle. Most movements are more complicated, and the model would include components such as gravity and multiple joints.

A model of motion is useful not only for describing the motion but also for considering how the motion is controlled. For optimal control (i.e., producing an input that maximizes some objective function), general principles are found, for example, in Kirk (1970) and Hogan (1988). For biological systems, two main ways of controlling a movement have been proposed. Suppose that the goal is to produce a position $A$ of the limb. The first way is for the system to estimate and then apply the force $F(t)$ needed to produce position $A$. The second way is not ordinarily available in a physical system, where the characteristics of the spring and damper are fixed. But in a biological system the stiffness

and other features of the muscles can be changed. Hence, to produce movement, the equilibrium position of the spring can be set directly to that needed to produce the limb position goal (Feldman, 1966). With both methods, later corrections can be made based on feedback, although these are more often considered with the first method.

To be useful, a model must produce known findings about human movement. One of the main results is Fitts's law. Suppose that a person moves a limb to a target of width $W$ centered at a distance $A$ away from the starting position. The time to make the movement is well approximated by

$$MT = c + d \log_2 \frac{2A}{W} \tag{8}$$

where $c$ and $d$ are free parameters (Fitts, 1954). The quantity $\log_2(2A/W)$ is called the *index of difficulty*. The relation was first fit to data for people moving a stylus back and forth continually between two targets, each of width $W$, with the centers of the targets separated by distance $A$. But it fits well for many other situations, for example, for moving a finger to a calculator button or a mouse pointer to a target (Card et al., 1983).

The simple mass, spring, and damper model leads to Fitts's law approximately when a force is applied to the limb (Langolf et al., 1976; for review, see Jagacinski and Flach, 2003). We illustrate the method of presetting the equilibrium position by giving a similar derivation leading to Fitts's law as an approximation for short movement times. Suppose the goal is to move the limb from starting position zero to position $A$. With this method of control, the force $F(t)$ in equation (7) is zero. Then the solution to equation (7) is (e.g., Resnick and Halliday, 1963)

$$x - q = Ce^{-bt/2m} \cos(\omega' t + \delta)$$

where $\omega' = \sqrt{k/m - (b/2m)^2}$, and $C$ and $\delta$ are parameters depending on the initial conditions. It is straightforward to check that this is a solution by taking first and second derivatives. When the motion is bounded, all solutions have this form for the underdamped case (when $b$ is small). In that case, the movement is a damped oscillation. Through trigonometric identities, the solution can be expressed in various equivalent ways.

Suppose that at time zero the initial position $x$ is zero and the initial velocity $dx/dt$ is zero. Then it is straightforward to see that $C = -q/\cos \delta$, $\delta$ is the angle whose tangent is $-b/2m\omega'$, and $\cos \delta = \sqrt{1 - b^2/4mk}$.

To set the limb position goal to $A$, set the equilibrium position $q$ to $A$. With these parameters, the position $x$ of the limb at time $t$ is

$$x = A - Ae^{-bt/2m} \frac{\cos(\omega' t + \delta)}{\cos \delta}$$

According to the model, the limb will oscillate, passing back and forth over the target position $A$ and

stopping at it at time infinity. (Oscillations in a movement can often be viewed when moving a mouse pointer to a target on a computer screen.) We are interested in the time at which the limb moves into the target interval and does not leave it. As an approximation, let us say that this happens when the envelope of the oscillations is within the target interval. Consider the lower boundary of the envelope (the result is the same if we consider the upper boundary). It reaches the target interval when

$$A - \frac{W}{2} = A - \frac{Ae - bt/2m}{\cos \delta}$$

Then

$$\frac{W}{2} = \frac{Ae - bt/2m}{\cos \delta}$$

$$\ln \frac{W}{2A} = -\frac{bt}{2m} + \ln(\cos \delta)$$

$$t = \frac{2m}{b} \ln \frac{2A}{W} + \frac{2m}{b} \ln(\cos \delta)$$

The log can be changed from base $e$ to base 2 by multiplying by $\ln 2$, and the result is in the form of Fitts's law.

We mention one more regularity for checking a working model. In the task for which Fitts's law applies, a person is given a target's center position and width and produces a movement time. In a slightly different task, a person is given a target position $A$ and a movement time goal $MT$ and produces a movement to a position. The standard deviation of the movement positions produced, $W_e$, is well approximated by Schmidt's law,

$$W_e = \frac{cMT}{A}$$

where $c$ is a free parameter (Schmidt et al., 1979). The two tasks are similar, as are the two laws, and a model leading to each as a special case has been proposed by Meyer et al. (1990).

The basic mass, spring, and damper model described above fails to predict some aspects of movement accurately (see, e.g., Langolf et al., 1976; Jagacinski et al., 1980), so there are many variations of the basic model. Controlling a movement with a single brief initial step function force predicts movements more asymmetrical than are found in data, and better fits are produced by assuming an initial accelerating force and a final decelerating force (the bang-bang model; see Jagacinski and Flach, 2003). The equilibrium point hypothesis has trouble explaining fast movements. De Lussanet et al. (2002) propose as an improvement controlling the movement by moving the equilibrium point to its goal position at a constant velocity rather than in a jump. Some evidence that the equilibrium point moves is provided in an experiment by Bizzi et al. (1992).

It is difficult to differentiate between the hypothesized methods of control (i.e., control by directly producing forces and control by setting the equilibrium
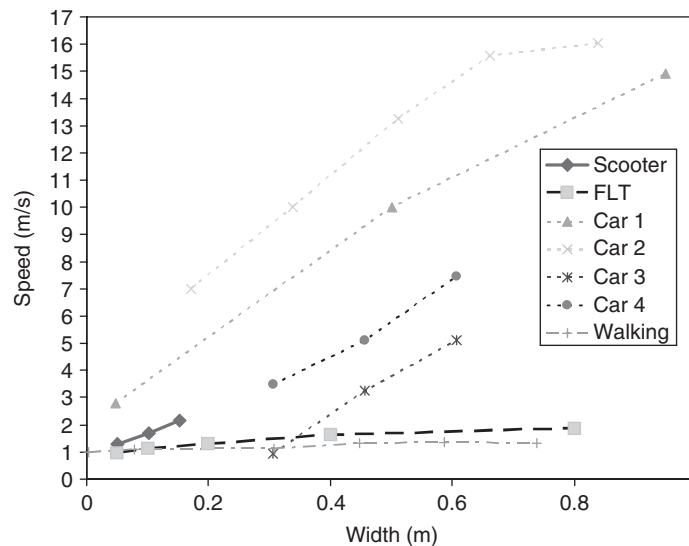
**Figure 9**  Relation between speed and path width.

point). One difficulty is that when changes in the system are observed, it is difficult to establish that they occurred for the purpose of control. Naturally, for a complicated system, different methods of control are probably used in different situations, as proposed by Schmidt and McGown (1980).

Fitts's law describes terminal aiming tasks, where the path to the target is unimportant but hitting the target at the terminal point of the aiming task is vital. This task is self-paced, and in fact, Fitts's law describes a speed–accuracy trade-off. A rather different form of self-paced movement is a path control task where the operator must move along a path without exceeding lateral boundaries. Examples are walking along a narrow corridor, driving along a narrow road, or even sewing along a seam of fixed width. An early study of line drawing along fixed-width paths (Drury, 1971) derived a model based on the operator as an intermittently acting servomechanism. At any instant, the operator finds himself or herself at some point across the allowed width and must choose how to make an open-loop movement during the next sampling interval. Drury et al. (1987) and Montazer et al. (1989) modeled this task as one of choosing a direction (angle $\theta$ to centerline) and a distance ($R$ in direction $\theta$) for the next movement. They assumed that the objective function was to maximize the distance traveled along the path ($R \cos \theta$) while minimizing the probability of going outside the path boundaries. The constraint set was derived from models of the buildup of lateral error in blind movements (e.g., Beggs and Howarth, 1970). The optimization model gave the same formulation as Drury's original (1971) model, namely

$$\text{speed} = \text{constant} \times \text{path width}$$

The model was derived and validated for movements on straight and circular courses. For very large widths, the speed is limited by other factors, such as a speed limit on a highway, so the linear relationship will eventually flatten out at high widths. Figure 9 shows speed–width relationships found by a number of authors for many different vehicles, including unpublished data on a personal "scooter" with side-by-side wheels. As in other examples, the optimization model provides a good description of performance.

***Application: Manual Assembly***  Many workers perform manual assembly tasks when their hands are located above their shoulders or far out from the sides of their body. Although this is not the preferred location (Konz, 1967), it is still a situation common in the workforce (Wiker et al., 1989). Some such tasks resemble a repetitive Fitts's tapping task. In fact, it has been estimated that anywhere between 40 and 80% of the cycle times in typical manual assembly tasks are due to the move and positioning elements required to perform the task (Arberg, 1963). To study such elements by themselves, Wiker et al. asked subjects to move a tool back and forth, repeatedly, between two holes. The tool was shaped like a small hand drill and contained a stylus that the subject had to position in the center of the hole. In this task it was possible to adjust, among other things, the movement amplitude (the distance between the two holes, let this be labeled $A$), direction (horizontal or vertical), target hole diameter ($2W$), positioning tolerance (the distance $T$ between the outer perimeter of a pin being placed in the center of a round hole and the edge of the hole), task duration, tool mass ($m$, in kilograms), duty cycle (number of repetitions per second, $N$), and hand elevation (height of arm above shoulder height, $E$). Wiker et al. (1989)

found that there was a lawful relation between these factors and the movement and positioning time (MT) [an elaboration of Fitts's law first proposed by Hoffman (1981; cited in Chung, 1983)]:

$$MT(ms)$$
$$= 106 + (68 + 0.015E + 0.00034E \times m \times N)\text{IDM}$$
$$+ (24 + 0.001E \times N)\text{IDP} \qquad (9)$$

where the index of move, IDM, and index of position, IDP, are defined as follows:

$$\text{IDM} = \log_2 \frac{A}{2W} \qquad \text{IDP} = \log_2 \frac{2W}{T}$$

(The relation between equations (9) and (8) is immediately apparent given the definition of IDM, which was referred to above in a different context as the index of difficulty.)

As an example of the changes in the movement and positioning times observed as a function of changes in the independent variables, Wiker et al. (1989) find that the move and positioning times increased, respectively, by 15.3 and 26.5% when the arm went from $15°$ below shoulder height to $60°$ above shoulder height. The equation above is particularly useful if, say, one wants to identify the optimal elevation, assuming that workers of many different heights will be performing the job.

## 8  TRAINING, EDUCATION, AND INSTRUCTIONAL SYSTEMS

Training continues to be of central importance in human factors in both the private and public sectors. However, quantitative models that could be used in training, and more broadly in education, have remained elusive for the most part. This is changing radically. Below we discuss some of the very earliest work and then segue to a discussion of some more current research.

### 8.1  Paired-Associate Models

Much was done in the 1950s and 1960s with paired-associate learning (Estes, 1959; Bush and Mosteller, 1951). In that work, and in much that followed (Bower, 1961), an attempt was made to predict the rate at which the associations between stimuli and responses were learned. In a typical paradigm, participants would be given a stimulus, say a letter, and be asked to produce the correct response, say a particular digit. A number of factors were varied, including the total number of stimuli in the list, the time between repetitions of the same stimulus, and the time between the last trial and an evaluation of performance.

In the simplest case, the one we describe here, the stimulus–response pair is assumed to be in one of two states, either learned ($C$, or conditioned) or not learned ($\overline{C}$, not conditioned). Let $c$ be the probability that it is learned on any one trial. Let $g$ be the probability $P(\text{correct})$ that the participant guesses the correct answer, assuming that the stimulus–response association is not learned; assume that this probability is 1 if the association is learned. Then the model can be presented formally as a two-state Markov chain where:

| State on Trial $n$ | State on Trial $n + 1$ | | Response when in state | $P(\text{correct})$ |
|---|---|---|---|---|
| | $C$ | $\overline{C}$ | | |
| $C$ | 1 | 0 | $C$ | 1 |
| $\overline{C}$ | $c$ | $1 - c$ | $\overline{C}$ | $g$ |

It can easily be shown that the probability that a participant responds correctly after $n$ trials is equal to

$$P(\text{correct}) = 1 - (1 - g)(1 - c)^n$$

Now, suppose that one wanted to maximize the joint probability that each of the stimuli had been learned after $n$ training trials, where $n$ is greater than the number of paired associates. Then Karush and Dear (1966) have derived the optimal training strategy. It is so simple that it can easily be described in a sentence or two. One must first train exactly once all stimulus–response pairs. One then needs to keep track of an index for each stimulus–response pair: That index is set to the zero if the response on the preceding trial was incorrect; it is set to the number of successive correct responses if the response on the preceding trial was correct. One then selects to train on the next trial the stimulus with the smallest such index. Unfortunately, simple two-state models cannot explain some of the results. In fact, Katsikopoulos and Fisher (2001) have shown that if one is going to explain several of the most critical results, one will need two Markov chains, one applied on each trial in which an association is trained and one applied on each trial in which an association is not trained. Moreover, the chains will need, at a minimum, four states. This has made the analytic derivation of the optimal global training strategy all but impossible. Instead, optimization is performed over a relatively small horizon or simulations are used to approximate the best training schedule.

***Application: Morse Code***   The applications of the work in paired-associate learning are relatively few but still potentially significant. In the military, some soldiers continue to be trained in the use of Morse code as a backup. The associations between the stimuli, "dots" and "dashes," and the responses, letters and digits, must be learned. The training takes a very long time, so the military was interested in learning whether something could be done to reduce the training time (Fisher and Townsend, 1993). For an application such as this, the assumptions required to identify the optimal order in which to train the stimuli as set forth by Karush and Dear (1966) are reasonably well satisfied, and so their method can be applied with very little additional cost (since the training was already being done on a computer and the stimuli and responses recorded). Other applications might well include the learning of simple multiplication

facts (the multiplicands are the stimulus and the product is the response) or the vocabulary in a foreign language (the foreign word is the stimulus and the English word equivalent is the response).

## 8.2 Complex Skill Acquisition

A more complex model of learning is needed than the paired-associate model described above if one is going to explain the majority of learning that goes on in the classroom and elsewhere. Anderson (1983) has developed just such a model, one that is widely used and constantly under development, its most recent version being referred to as ACT-R (Anderson et al., 2004). Very briefly, the process starts with a task analysis, often using the GOMS procedures mentioned above (Card et al., 1983). Once a task is decomposed into increasingly better defined goals, a working model is built from two general types of knowledge: declarative and procedural knowledge.

*Declarative knowledge* is based on the instructions that the participant receives when performing a task and on the task analysis. These are stored in a semantic network. For example, a student will learn that the sum of three angles in a triangle must equal $180°$. *Procedural knowledge* is stored as production rules, rules that take actions consistent with the declarative component, including retrieving information from memory, focusing attention on a certain area of the display, pressing a key in response to a stimulus, and so on. For example, the three rules in Table 1, described in more detail in Taatgen and Lee (2003), are used to add three numbers (see rules 1, 2, and 3). There are various hard constraints on both the general form that productions can take and the execution of those productions. Most relevant here

are two such constraints. First, only one item at a time can be retrieved from declarative memory. Second, the production rules must fire in sequence.

From the standpoint of learning, the question that must be addressed is how participants over time become faster when performing complex tasks. In general terms, skill acquisition has been broken down into three stages: the cognitive, associative, and autonomous stages (Fitts, 1964). The answer within the framework of ACT-R has varied over time. Most recently, Taatgen and Anderson (2002) have suggested that with enough use, two rules that occur one after one another can be combined in such a way that the retrieval request from declarative memory in the first rule is no longer assumed to be necessary. Thus, only one production rule is necessary, a rule consisting of the IF part of the first rule and the THEN part of the second rule. An example is given in Table 1 in rule 1&2. A relatively time-consuming step has now been saved, in particular the retrieval from declarative memory of the sum of the first two numbers. This is identified by them as *production compilation* and is consistent with their results in several different experiments.

### Application: Training Air Traffic Controllers

Taatgen and Lee (2003) used ACT-R to model the improvement in performance over time of participants (college undergraduates) asked to perform the Kanfer–Ackerman air traffic controller task (Ackerman, 1988; Ackerman and Kanfer, 1994). The task is a complex one and can be decomposed hierarchically into the unit task level (e.g., land a plane on the runway), the functional level (e.g., find a runway to land), and the keystroke level (e.g., press a particular key). Taatgen and Lee identified the declarative knowledge and task-independent procedural rules needed to perform each task at each level. In addition, they identified formally under what conditions declarative knowledge and task-independent production rules could be combined into task-specific production rules through the mechanism that was identified above as production compilation. Using values for parameters obtained outside the Kanfer–Ackerman air traffic controller task, they predicted for each of the first 10 trials how performance at the unit task level, functional level, and keystroke level would vary. They found that qualitatively their model fits the results very nicely. This is an example where the mathematical model could be used to limit the selection of interfaces to examine more completely, perhaps in an experiment, assuming that too large a number were available initially to evaluate fully.

## 9 WARNINGS

A warning device can be characterized in terms of signal detection theory. Consider a smoke alarm that goes off when the concentration of certain particles in the air exceeds a critical value. The critical value corresponds to the response bias. The standardized difference between the mean concentration of particles when there is a fire and when there is not corresponds to

**Table 1 Production Rules and Compilation**

| Rule 1 | IF | the goal is to add three numbers, $n_1, n_2,$ and $n_3,$ |
|---|---|---|
| | THEN | send a retrieval request to declarative memory for the sum $s_{1+2} = n_1 + n_2$ of the first two numbers. |
| Rule 2 | IF | the goal is to add three numbers AND the sum $s_{1+2}$ of the first two numbers is retrieved, |
| | THEN | send a retrieval request to declarative memory for the sum of (a) the first two numbers, $s_{1+2}$, and (b) the third number $n_3$; label this last sum $s_{1+2+3}$. |
| Rule 3 | IF | the goal is to add three numbers AND the sum $s_{1+2+3}$ of the first three numbers is retrieved, |
| | THEN | the answer is the retrieved sum, $s_{1+2+3}$. |
| Rule 1&2 | IF | the goal is to add the two numbers 3 and 5 together with a third number, $n_3$, |
| | THEN | send a request to declarative memory for the sum of 8 and $n_3$. |

the sensitivity of the device. In most environments the operator is not in a position to change the frequency of hits (there is a fire and the warning device indicates such), misses, false alarms (there is no fire and the warning device indicates that there is such), and correct rejections of the device itself. However, in some environments the operator can, and probably will, actually want to alter the frequency of false alarms (without tinkering with the warning system mechanics).

For example, consider nurses working in an intensive care unit (ICU). Such nurses will want to reduce the number of times that the warning sounds. This will reduce the number of times that a true emergency is present, given that the warning sounded, as well as the number of times that no emergency is present, given that the warning sounded. Meyer and Bitan (2002) realized that this will in turn reduce the information in the warning. At the extreme, if the operator is always able to take actions that prevent the warning from sounding in a true emergency, only false alarms will be generated. It is known that the operator's response time is influenced by the frequency of false alarms produced by the warning system; most important in this context, response time decreases when the frequency of false alarms increases (Getty et al., 1995). This raises the question of whether a warning device is more valuable or less valuable for better operators (i.e., operators who have reduced the number of times that an emergency situation occurs).

Because the answer depends on the combined effect of several quantities, this question would be difficult to answer without a model. Using signal detection theory, Meyer and Bitan (2002) calculated the predictive value of warnings as a function of the probability of a system failure. The positive predictive value is the probability that there actually is a failure given that the device produces an alarm. Let $F$ denote an actual failure of the system and $f$ denote that an alarm is given. Then the positive predictive value, PPV, is $P(F|f) =$ hits/(hits + false alarms). Analogously, the negative predictive value, NPV, is the probability the system is normal given that the device does not produce an alarm. Let $N$ denote the normal state of the system, and let $n$ denote that no warning is given. Then the negative predictive value is $P(N|n)$. It is clear from the information in Table 2 that the positive predictive value decreases as the operator reduces the number of instances of system failure from 100 (left side) to 10 (right side). The negative predictive value can easily be shown to increase here.

Given the data in the table, Meyer and Bitan (2002) asked how one might index the overall effect on the operator of changes in these two indices, the positive and negative predictive values, that were heading in opposite directions. They suggested several ways that one might combine this information, one of which is to use information theory and, in particular, to compute the information transmitted by the warning system, something we have already discussed above. Meyer and Bitan (2002) considered three hypothetical warning devices with different values of $d'$ and $\beta = 1$ (i.e., neutral bias). For each, they found that the negative predictive value increased slightly as the probability of actual failure decreased from 0.200 to 0.001. However, the positive predictive value decreased and was considerably lower at the low end of this range than at the high end, and so was the information transmitted. In short, the diagnostic value of an alarm is worse for better operators.

***Application: Intensive Care Units***   The conclusion above was borne out in an experiment on a simulated intensive care nurse's workstation, with an imperfect device warning that a patient needed attention (Meyer and Bitan, 2002). Performance of participants improved over time in the experiment. But the positive predictive value of an alarm decreased and the negative predictive value increased in such a way that the information transmitted by the alarm decreased as they became better operators. Unfortunately, the behavioral implications of these findings are not immediately clear. On the one hand, it is known that as the informative value of a warning goes down, operators are less likely to take action. On the other hand, it is clearly beneficial to reduce the number of situations in which a warning is required. Here is a situation where a mathematical model of a system has uncovered a problem that would probably not have been recognized. However, by itself it cannot be used to solve the problem. More information is needed about the performance of operators in such complex situations.

## 10   SUMMARY

We set out to show the broad range of uses to which mathematical models can be put in the design of the interface between the human user and the environment. Most centrally, we set out to show that such models not only can be used to design an interface, but can actually be used to optimize the interface. We gave several such examples, including (1) the design of optimal menu hierarchies, (2) the design of optimal inspection schedules, and (3) the design of optimal training sequences. Of course, as we made clear throughout the chapter, mathematical models have a broader use

**Table 2  Positive Predictive Value**

| Device's Response | System State | | Device's Response | System State | |
|---|---|---|---|---|---|
| | *F* | *N* | | *F* | *N* |
| *f* | 90 | 10 | *f* | 9 | 10 |
| *n* | 10 | 90 | *n* | 1 | 90 |
| | $PPV = \dfrac{90}{90 + 10} = 0.9$ | | | $PPV = \dfrac{9}{9 + 10} = 0.47$ | |

than simply optimizing the interface. We gave several examples of other uses as well, including (1) the prediction of performance with a particular interface to determine whether it will function as desired and therefore should be considered for implementation, (2) the identification of the effects of neurotoxic agents on the speed of latent processes, and (3) the determination of whether two apparently different models actually make different predictions.

We are hopeful that mathematical modeling will continue to play an important and increasing role in the design of the interface. There are several indications that such will be the case, including a recent special issue of *Human Factors* devoted entirely to mathematical models as well as the formation several years ago within the Human Factors and Ergonomics Society of a technical group whose interests focus on modeling. And, of course, we hope that this chapter will motivate others in the research community to think more broadly about how they too might apply one or more of the many modeling techniques described herein to the design of an interface.

## ACKNOWLEDGMENTS

## REFERENCES

Ackerman, P. L. (1988), "Determinants of Individual Differences during Skill Acquisition: Cognitive Abilities and Information Processing," *Journal of Experimental Psychology: General*, Vol. 117, pp. 288–318.

Ackerman, P. L., and Kanfer, R. (1994), *Kanfer-Ackerman Air Traffic Controller Task© CD_ROM Database, Data Collection Program and Playback Program*, Office of Naval Research, Cognitive Science Program, Arlington, VA.

Anderson, J. A. (1983), *The Architecture of Cognition*, Harvard University Press, Cambridge, MA.

Anderson, J. R. (1993), *Rules of the Mind*, Lawrence Erlbaum Associates, Mahwah, NJ.

Anderson, J. R., and Bower, G. H. (1973), *Human Associative Memory*, V.H. Winston & Sons, Washington DC.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004), "An Integrated Theory of Mind," *Psychological Review*, Vol. 111, No. 4, pp. 1036–1060.

Arani, T., Karwan, M. H., and Drury, C. G. (1984), "A Variable-Memory Model of Visual Search," *Human Factors*, Vol. 26, pp. 631–639.

Arberg, U. (1963), *Frequency of Occurrence of Basic MTM Motions, Research Report 1*, Svenska MTM-Foreningen, Stockholm, Sweden.

Baddeley, A. D. (1992), "Is Working Memory Working?" *Quarterly Journal of Experimental Psychology*, Vol. 44A, No. 1, pp. 1–31.

Balakrishnan, J. D. (1998a), "Some More Sensitive Measures of Sensitivity and Response Bias," *Psychological Methods*, Vol. 3, pp. 68–90.

Balakrishnan, J. D. (1998b), "Measures and Interpretations of Vigilance Performance: Evidence against the Detection Criterion," *Human Factors*, Vol. 40, pp. 601–623.

Bavejo, A., Drury, C. G., Karwan, M., and Malone, D. M. (1996), "Derivation and Test of an Optimum Overlapping-Lobe Model of Visual Search," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 28, pp. 161–168.

Beggs, W. D. A., and Howarth, C. I. (1970), "Movement Control in a Repetitive Motor Task," *Nature*, Vol. 225, pp. 752–753.

Bishu, R. R., and Drury, C. G. (1988), "Information Processing in Assembly Tasks: A Case Study," *Applied Ergonomics*, Vol. 19, No. 2, pp. 90–98.

Bizzi, E., Hogan, N., and Mussa-Ivaldi, F. A. (1992), "Does the Nervous System Use Equilibrium-Point Control to Guide Single and Multiple Joint Movements?" *Behavioral and Brain Sciences*, Vol. 15, pp. 603–611.

Bower, G. H. (1961), "Application of a Model to Paired-Associate Learning," *Psychometrika*, Vol. 26, pp. 255–280.

Bush, R. R., and Mosteller, F. A. (1951), "A Mathematical Model for Simple Learning," *Psychological Review*, Vol. 58, pp. 505–512.

Byrne, M. D., and Anderson, J. R. (1998), "Perception and Action," in *The Atomic Components of Thought*, J. R. Anderson and C. Lebiere, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 167–200.

Byrne, M. D., and Gray, W. D. (2003), "Returning Human Factors to an Engineering Discipline: Expanding the Science Base through a New Generation of Quantitative Methods" (Preface to the Special Section), *Human Factors*, Vol. 45, pp. 1–4.

Card, S. K., Moran, T. P., and Newell, A. (1983), *The Psychology of Human-Computer Interaction*, Lawrence Erlbaum Associates, Mahwah, NJ.

Chaffin, D. B., Andersson, G. B. J., and Martin, B. J. (1999), *Occupational Biomechanics*, 3rd ed., Wiley, New York.

Chi, C.-F., and Drury, C. G. (1998), "Do People Choose an Optimal Response Criterion in an Inspection Task?" *IIE Transactions*, Vol. 30, pp. 257–266.

Chi, C.-F., and Drury, C. (2001), "Limits to Human Optimization in Inspection Performance," *International Journal of Systems Science*, Vol. 32, No. 6, pp. 689–701.

Chung, C. S. (1983), "A Study of Human Arm Movements with Applications to Predetermined Motion Time Systems," unpublished master's thesis, University of Melbourne, Australia.

Conway, R. W., Maxwell, W. L., and Miller, L. W. (1967), *Theory of Scheduling*, Addison-Wesley, Reading, MA.

Cordo, P., and Gurfinkel, V., Eds. (1992), "Controversies in Neuroscience I: Movement Control" (Special Issue), *Behavioral and Brain Sciences*, Vol. 15, No. 4.

Crossman, E. R. F. W., and Goodeve, P. J. (1963), "Feedback Control of Hand-Movement and Fitts' Law," paper presented at the Meeting of the Experimental Psychology Society, Oxford, July; *Quarterly Journal of Experimental Psychology* (1983), Vol. 35A, pp. 251–278.

Davies, D. R., and Parasuraman, R. (1982), *The Psychology of Vigilance*, Academic, London.

Davis, R. (1957), "The Human Operator as a Single Channel Information System," *Quarterly Journal of Experimental Psychology*, Vol. 9, pp. 119–129.

de Lussanet, M. H. E., Smeets, J. B. J., and Brenner, E. (2002), "Relative Damping Improves Linear Mass-Spring Models of Goal-Directed Movements," *Human Movement Science*, Vol. 21, pp. 85–100.

Donaldson, W. (1992), "Measuring Recognition Memory," *Journal of Experimental Psychology: General*, Vol. 121, pp. 275–277.

Drury, C. G. (1971), "Movements with Lateral Constraint," *Ergonomics*, Vol. 14, pp. 293–305.

Drury, C. G. (1973), "The Effect of Speed of Working on Industrial Inspection Accuracy," *Applied Ergonomics*, Vol. 4, pp. 2–7.

Drury, C. G. (1993), "A Note on Fitts' Law and Assembly Order," *Ergonomics*, Vol. 36, No. 7, pp. 801–806.

Drury, C. G. (1999), "Human Factors Good Practices in Fluorescent Penetrant Inspection," in *Human Factors in Aviation Maintenance: Phase Nine, Progress Report*, DOT/FAA/AM-99/xx, National Technical Information Service, Springfield, VA.

Drury, C. G. (2000), "Global Quality: Linking Ergonomics and Production," *International Journal of Production Research*, Vol. 38, No. 17, pp. 4007–4018.

Drury, C. G. (2001), "Human Factors and Automation in Test and Inspection," in *Handbook of Industrial Engineering*, 3rd ed., G. Salvendy, Ed., Wiley, New York, pp. 1887–1920.

Drury, C. G. (2002), "A Unified Model of Security Inspection," in *Proceedings of the Aviation Security Technical Symposium*, Federal Aviation Administration, Atlantic City, NJ.

Drury, C. G., and Addison, J. L. (1973), "An Industrial Study of the Effects of Feedback and Fault Density on Inspection Performance," *Ergonomics*, Vol. 16, pp. 159–169.

Drury, C. G., and Chi, C.-F. (1995), "A Test of Economic Models of Stopping Policy in Visual Search," *IIE Transactions*, Vol. 27, pp. 392–393.

Drury, C. G., and Forsman, D. R. (1996), "Measurement of the Speed Accuracy Operating Characteristic for Visual Search," *Ergonomics*, Vol. 39, No. 1, pp. 41–45.

Drury, C. G., and Fox, J. G., Eds. (1975), *Human Reliability in Quality Control*, Taylor & Francis, London.

Drury, C. G., and Hong, S.-K. (2000), "Generalizing from Single Target Search to Multiple Target Search," *Theoretical Issues in Ergonomics Science*, Vol. 1, No. 4, pp. 303–314.

Drury, C. G., and Kleiner, B. M. (1990), "Training in Industrial Environments," in *Proceedings of the Human Factors Association of Canada Meeting*, Ottawa, Ontario, Canada, pp. 99–108.

Drury, C. G., and Sinclair, M. A. (1983), "Human and Machine Performance in an Inspection Task," *Human Factors*, Vol. 25, pp. 391–400.

Drury, C. G., Montazer, M. A., and Karwan, M. H. (1987), "Self-Paced Path Control as an Optimization Task," *IEEE Transactions*, Vol. SMC-17.3, pp. 455–464.

Drury, C. G., Spencer, F. W., and Schurman, D. (1997), "Measuring Human Detection Performance in Aircraft Visual Inspection," in *Proceedings of the 41st Annual Human Factors and Ergonomics Society Meeting*, Albuquerque, NM.

Dutta, A., Fisher, D. L., and Noyce, D. A. (2005), "Use of a Driving Simulator to Evaluate and Optimize Factors Affecting Understandability of Variable Message Signs," *Transportation Research F*, Vol. 7, pp. 209–227.

Dzhafarov, E. N. (1993), "Grice-Representability of Response Time Distribution Families," *Psychometrika*, Vol. 58, pp. 281–314.

Dzhafarov, E. N., Schweickert, R., and Sung, K. (2004), "Mental Architectures with Selectively Influenced but Stochastically Interdependent Components," *Journal of Mathematical Psychology*, Vol. 48, pp. 51–64.

Ehrenstein, A., Schweickert, R., Choi, S., and Proctor, R. W. (1997), "Scheduling Processes in Working Memory: Instructions Control the Order of Memory Search and Mental Arithmetic," *Quarterly Journal of Experimental Psychology*, Vol. 50A, pp. 766–802.

Elmaghraby, S. E. (1977), *Activity Networks: Project Planning and Control by Network Models*, Wiley, New York.

Estes, W. K. (1959), "Toward a Statistical Theory of Learning," *Psychological Review*, Vol. 57, pp. 94–107.

Feldman, A. G. (1966), "Functional Tuning of the Nervous System during Control of Movement or Maintenance of a Steady Posture, III: Mechanographic Analysis of the Execution by Man of the Simplest Motor Task," *Biophysics*, Vol. 11, pp. 766–775.

Fiedler, N., Feldman, R. G., Jacobson, J., Rahill, A., and Wetherell, A. (1996), "The Assessment of Neurobehavioral Toxicity: SGOMSEC Joint Report," *Environmental Health Perspectives*, Vol. 104, Suppl. 2, pp. 179–191.

Fific, M., Little, D. R., and Nosofsky, R. M. (2010), "Logical-Rule Models of Classification Response Times: A Synthesis of Mental-Architecture, Random-Walk, and Decision-Bound Approaches," *Psychological Review*, Vol. 117, pp. 309–348.

Fisher, D. L. (1993), "Optimal Performance Engineering," *Human Factors*, Vol. 35, pp. 115–140.

Fisher, D. L., and Glaser, R. (1996), "Molar and Latent Models of Cognitive Slowing: Implications for Aging, Dementia, Depression, Development and Intelligence," *Psychonomic Bulletin and Review*, Vol. 3, pp. 458–480.

Fisher, D. L., and Goldstein, W. M. (1983), "Stochastic PERT Networks as Models of Cognition: Derivation of the Mean, Variance, and Distribution of Reaction Time Using Order-of-Processing (OP) Diagrams," *Journal of Mathematical Psychology*, Vol. 27, pp. 121–115.

Fisher, D. L., and Townsend, J. T. (1993), *Models of Morse Code Skill Acquisition: Simulation and Analysis*, U.S. Army Research Institute, Training Systems Research Division, Automated Instructional Systems Technical Area, Alexandria, VA, February.

Fisher, D. L., Yungkurth, E., and Moss, S. (1990), "Optimal Menu Hierarchy Design: Syntax and Semantics," *Human Factors*, Vol. 32, No. 6, pp. 665–683.

Fitts, P. M. (1954), "The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement," *Journal of Experimental Psychology*, Vol. 47, pp. 381–391.

Fitts, P. M. (1964), "Perceptual-Motor Skill Learning," in *Categories of Human Learning*, A. W. Melton, Ed., Academic, New York, pp. 243–285.

Getty, D. J., Swets, J. A., Pickett, R. M., and Gonthier, D. (1995), "System Operator Response to Warnings of Danger: A Laboratory Investigation of the Predictive Value of a Warning on Human Response Time," *Journal of Experimental Psychology: Applied*, Vol. 1, pp. 19–33.

Goldstein, W. M., and Fisher, D. L. (1991), "Stochastic Networks as Models of Cognition: Derivation

of Response Time Distributions Using the Order-of-Processing Method," *Journal of Mathematical Psychology*, Vol. 35, pp. 214–241.

Goldstein, W. M., and Fisher, D. L. (1992), "Stochastic Networks as Models of Cognition: Deriving Predictions for Resource-Constrained Mental Processing," *Journal of Mathematical Psychology*, Vol. 36, pp. 129–145.

Gray, W. D. (2008), "Cognitive Architectures: Choreographing the Dance of Mental Operations with the Task Environment," *Human Factors*, Vol. 50, pp. 497–505.

Gray, W. D., John, B. E., and Atwood, M. E. (1993), "Project Ernestine: Validating a GOMS Analysis for Predicting and Explaining Real-World Task Performance," *Human-Computer Interaction*, Vol. 8, pp. 237–309.

Green, D. M., and Swets, J. A. (1966), *Signal Detection Theory and Psychophysics*, Wiley, New York.

Hick, W. E. (1952), "On the Rate of Information Gain," *Quarterly Journal of Experimental Psychology*, Vol. 4, pp. 11–26.

Hogan, N. (1988), "Physical Systems Theory and Controlled Manipulation," in *Natural Computation*, W. Richards, Ed., MIT Press, Cambridge, MA, pp. 430–442.

Hoffmann, E. R., Macdonald, W. A., and Almond, G. A. (1993, April), "Quantification of the Cognitive Difficulty of Mail Sorting," *International Journal of Industrial Ergonomics*, Vol. 11, No. 2, pp. 83–98.

Hommel, B. (1998), "Automatic Stimulus-Response Translation in Dual-Task Performance," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 24, pp. 1368–1394.

Hong, S. K. (2002), "Human Performance in Visual Search for Multiple Targets," unpublished doctoral dissertation, State University of New York at Buffalo, Buffalo, NY.

Hong, S. K., and Drury, C. G. (2002), "Sensitivity and Validity of Visual Search Models for Multiple Targets," *Theoretical Issues in Ergonomic Science*, Vol. 3.1, pp. 85–110.

Hou, T. S., Lin, L., and Drury, C. G. (1993), "An Empirical Study of Hybrid Inspection Systems and Allocation of Inspection Function," *International Journal of Human Factors in Manufacturing*, Vol. 3, pp. 351–367.

Hyman, R. (1953), "Stimulus Information as a Determinant of Reaction Time," *Journal of Experimental Psychology*, Vol. 45, pp. 188–196.

Ingham, J. G. (1970), "Individual Differences in Signal Detection," *Acta Psychologica*, Vol. 34, pp. 39–50.

Jagacinski, R. J., and Flach, J. M. (2003), *Control Theory for Humans: Quantitative Approaches to Modeling Performance*, Lawrence Erlbaum Associates, Mahwah, NJ.

Jagacinski, R. J., Repperger, D. W., Moran, M. S., Ward, S. L., and Glass, B. (1980), "Fitts' Law and the Microstructure of Rapid Discrete Movements," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 6, pp. 309–320.

Jax, S. A., Rosenbaum, D. A., Vaughan, J., and Meulenbroek, R. G. J. (2003), "Computational Motor Control and Human Factors: Modeling Movements in Real and Possible Environments," *Human Factors*, Vol. 45, pp. 5–27.

John, B. E. (1990), "Extensions of GOMS Analyses to Expert Performance Requiring Perception of Dynamic Visual and Auditory Information," in *Proceedings of the CHI '90 Conference on Human Factors in Computing Systems*, J. Carrasco and J. Whiteside, Eds., Association for Computing Machinery, New York, pp. 107–115.

John, B. E., and Newell, A. (1989), "Toward an Engineering Model of Stimulus-Response Compatibility," in *Stimulus-Response Compatibility: An Integrated Perspective*, R. W. Proctor and T. G. Reeve, Eds., Elsevier, New York, pp. 427–479.

Johnston, J. C., and McCann, R. S. (2006), "On the Locus of Dual-Task Interference: Is There a Bottleneck at the Stimulus Classification Stage?" Quarterly Journal of Experimental Psychology, Vol. 59, pp. 694–719.

Karush, W., and Dear, R. E. (1966), "Optimal Stimulus Presentation Strategy for a Stimulus Sampling Model of Learning," *Journal of Experimental Psychology*, Vol. 3, pp. 19–47.

Katsikopoulos, K., and Fisher, D. L. (2001), "Formal Requirements of Markov State Models for Paired Associate Learning," *Journal of Mathematical Psychology*, Vol. 45, No. 2, pp. 324–333.

Kirk, D. E. (1970), *Optimal Control Theory: An Introduction*, Prentice-Hall, Englewood Cliffs, NJ.

Konz, S. (1967), "Design of Work Stations," *Journal of Industrial Engineering*, Vol. 18, pp. 413–423.

Kruskal, J. B. (1964), "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," *Psychometrika*, Vol. 29, pp. 1–27.

Langolf, G. D., Chaffin, D. B., and Foulke, J. A. (1976), "An Investigation of Fitts' Law Using a Wide Range of Movement Amplitudes," *Journal of Motor Behavior*, Vol. 8, pp. 113–128.

Laughery, K. R. (1985), "Modeling Human Operators on a Microcomputer–A Micro Version of SAINT," *Simulation*, Vol. 44, pp. 10–16.

Lee, E., and MacGregor, J. (1985), "Minimizing User Search Time in Menu Retrieval Systems," *Human Factors*, Vol. 27, pp. 157–162.

Liu, Y. (1996), "Queueing Network Modeling of Elementary Mental Processes," *Psychological Review*, Vol. 103, pp. 116–136.

Liu, Y., Feyen, R., and Tsimhoni, O. (2006), "Queueing Network-Model Human Processor QN-MHP): A Computational Architecture for Multitask Performance in Human-Machine Systems," *ACM Transactions on Computer-Human Interaction*, Vol. 13, pp. 37–70.

Logan, G. D. (2002), "Parallel and Serial Processing," in *Stevens' Handbook of Experimental Psychology: Methodology in Experimental Psychology*, 3rd ed., Vol. 4, J. Wixted, Ed., Wiley, New York, pp. 271–300.

Luce, R. D. (1986), *Response Times: Their Role in Inferring Elementary Mental Organization*, Oxford University Press, New York.

Macmillan, N. A., and Creelman, C. D. (1991), *Detection Theory: A User's Guide*, Cambridge University Press, Cambridge.

McCann, R. S., and Johnston, J. C. (1992), "Locus of the Single-Channel Bottleneck in Dual-Task Interference," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 18, pp. 471–484.

Megaw, E. D. (1979), "Factors Affecting Visual Inspection Accuracy," *Applied Ergonomics*, Vol. 10, No. 1, pp. 27–32.

Meyer, D. E., and Kieras, D. E. (1997a), "A Computational Theory of Executive Cognitive Processes and Multiple-Task Performance, 1: Basic Mechanisms," *Psychological Review*, Vol. 104, pp. 3–65.

Meyer, D. E., and Kieras, D. E. (1997b), "A Computational Theory of Executive Cognitive Processes and Multiple-Task Performance, 2: Accounts of Psychological Refractory-Period Phenomena," *Psychological Review*, Vol. 104, pp. 749–791.

Meyer, D. E., Smith, J. E. K., Kornblum, S., Abrams, R. A., and Wright, C. E. (1990), "Speed-Accuracy Tradeoffs in Aimed Movements: Toward a Theory of Rapid Voluntary Action," in *Attention and Performance XIII*, M. Jeannerod, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 173–226.

Meyer, J., and Bitan, Y. (2002), "Why Better Operators Receive Worse Warnings," *Human Factors*, Vol. 44, pp. 343–353.

Montazer, M. A., Drury, C. G., and Karwan, M. H. (1989), "An Optimization Model of Self-Paced Tracking on Circular Courses," *IEE Transactions*, Vol. SMC 18.6, pp. 908–916.

Morawski, T., Drury, C. G., and Karwan, M. H. (1980), "Predicting Search Performance for Multiple Targets," *Human Factors*, Vol. 22, No. 6, pp. 707–718.

Mueller, S. T., and Weidemann, C. T. (2008), "Decision Noise: An Explanation for Observed Violations of Signal Detection Theory," *Psychonomic Bulletin & Review*, Vol. 15, pp. 464–494.

Navon, D., and Miller, J. (2002), "Queuing or Sharing? A Critical Evaluation of the Single-Bottleneck Notion," *Cognitive Psychology*, Vol. 44, pp. 193–251.

Newell, A. (1990), *Unified Theories of Cognition*, Harvard University Press, Cambridge, MA.

Newell, A., and Simon, H. A. (1972), *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, NJ.

Norman, D. A., and Bobrow, D. G. (1975), "On Data-Limited and Resource Limited Processes," *Cognitive Psychology*, Vol. 7, pp. 44–64.

Parasuraman, R., and Davies, D. R. (1977), "A Taxonomic Analysis of Vigilance Performance," in *Vigilance: Theory, Operational Performance, and Physiological Correlates*, R. R. Mackie, Ed., Plenum, New York, pp. 559–574.

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000), "A Model for Types and Levels of Human Interaction with Automation," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, Vol. 30, pp. 286–297.

Pashler, H. (1998), *The Psychology of Attention*, MIT Press, Cambridge, MA.

Patton, E. W., and Gray, W. D. (2010), SANLab-CM: A Tool for Incorporating Stochastic Operations into Activity Network Modeling. *Behavior Research Methods*, Vol. 42, No. 3, pp. 877–883.

Pew, R. W. (2008), "More Than 50 Years of History and Accomplishments in Human Performance Model Development," *Human Factors*, Vol. 50, pp. 489–496.

Pritsker, A. A. B. (1979), *Modeling and Analysis Using Q-GERT Networks*, 2nd ed., Wiley, New York.

Resnick, R., and Halliday, D. (1963), *Physics: For Students of Science and Engineering*, Wiley, New York.

Rouse, W. B. (1980), *Systems Engineering Models of Human-Machine Interaction*, North-Holland, New York.

Rumelhart, D. E., and McClelland, J. L. (1986), *Parallel Distributed Processing*, Vol. 1, *Explorations in the Microstructure of Cognition*, MIT Press, Cambridge, MA.

Salthouse, T. A., and Somberg, B. L. (1982), "Isolating the Age Deficit in Speeded Performance," *Journal of Gerontology*, Vol. 37, pp. 349–357.

Schmidt, R. A., and McGown, C. (1980), "Terminal Accuracy of Unexpectedly Loaded Rapid Movements: Evidence for a Mass-Spring Mechanism in Programming," *Journal of Motor Behavior*, Vol. 12, pp. 149–161.

Schmidt, R. A., Zelaznik, N. H., Hawkins, B., Frank, J. S., and Quinn, J. T., Jr. (1979), "Motor Output Variability: A Theory for the Accuracy of Rapid Motor Acts," *Psychological Review*, Vol. 86, pp. 415–451.

Schweickert, R. (1978), "A Critical Path Generalization of the Additive Factor Method: Analysis of a Stroop Task," *Journal of Mathematical Psychology*, Vol. 18, pp. 105–139.

Schweickert, R., and Townsend, J. T. (1989), "A Trichotomy: Interactions of Factors Prolonging Sequential and Concurrent Mental Processes in Stochastic Discrete Mental (PERT) Networks," *Journal of Mathematical Psychology*, Vol. 33, pp. 328–347.

Schweickert, R., Fisher, D. L., and Goldstein, W. (1992), *General Latent Network Theory: Structural and Quantitative Analysis of Networks of Cognitive Processes*, Technical Report 92-1, Mathematical Psychology Program, Purdue University, West Lafayette, IN.

Schweickert, R., Fisher, D. L., and Proctor, R. (2003), "Steps towards Building Mathematical and Computer Models from Cognitive Task Analysis," *Human Factors*, Vol. 45, pp. 77–103.

Schweickert, R., Fisher, D. L., and Sung, K. (in press). *Discovering Cognitive Architecture by Selectively Influencing Mental Processes*. World Scientific, Singapore.

Schweickert, R., Goldstein, W., and Fisher, D. L. (2010). Additive factors and stages of mental processes, *Journal of Mathematical Psychology*, Vol. 54, pp. 405–414.

See, I. E., Howe, S. R., Warm, J. S., and Dember, W. N. (1995), "Meta-Analysis of the Sensitivity Decrement in Vigilance," *Psychological Bulletin*, Vol. 117, pp. 230–249.

See, J. E., Warm, J. S., Dember, W. N., and Howe, S. R. (1997), "Vigilance and Signal Detection Theory: An Empirical Evaluation of Five Measures of Response Bias," *Human Factors*, Vol. 39, pp. 14–29.

Shannon, C. E., and Weaver, W. (1949), *The Mathematical Theory of Communications*, University of Illinois Press, Urbana, IL.

Sharit, J., Czaja, S. J., Nair, S., and Lee, C. C. (2003), "Effects of Age, Speech Rate, and Environmental Support in Using Telephone Voice Menu Systems," *Human Factors*, Vol. 45, pp. 234–251.

Shepard, R. N. (1962), "The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function, I," *Psychometrika*, Vol. 27, pp. 125–140.

Smith, M. C. (1969), "The Effect of Varying Information on the Psychological Refractory Period", in *Attention and Performance II*, W. G. Koster, Ed., North-Holland, Amsterdam, *Acta Psychologica*, Vol. 30, pp. 220–231.

Smith, P. J., and Langolf, G. D. (1981), "The Use of Sternberg's Memory Scanning Paradigm in Assessing Effects of Chemical Exposure," *Human Factors*, Vol. 23, pp. 701–708.

Sternberg, S. (1966), "High-Speed Scanning in Human Memory," *Science*, Vol. 153, pp. 652–654.

Sternberg, S. (1975), "Memory-Scanning: New Findings and Current Controversies," *Quarterly Journal of Experimental Psychology*, Vol. 27, pp. 1–32.

Taatgen, N. A., and Anderson, J. R. (2002), "Why Do Children Learn to Say 'Broke'? A Model of the Past Tense without Feedback," *Cognition*, Vol. 86, pp. 123–155.

Taatgen, N. A., and Lee, F. J. (2003), "Production Compilation: A Simple Mechanism to Model Complex Skill Acquisition," *Human Factors*, Vol. 45, pp. 61–76.

Tang, K. E. (2001), "Menu Design with Visual Momentum for Compact Smart Products," *Human Factors*, Vol. 43, pp. 267–277.

Tombu, M., and Jolicoeur, P. (2003), "A Central Capacity Sharing Model of Dual-Task Performance," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 29, pp. 3–18.

Townsend, J. T. (1972), "Some Results Concerning the Identifiability of Parallel and Serial Processes," *British Journal of Mathematical and Statistical Psychology*, Vol. 25, pp. 168–199.

Townsend, J. T., and Ashby, F. G. (1983), *The Stochastic Modeling of Elementary Psychological Processes*, Cambridge University Press, Cambridge.

Townsend, J. T., and Wenger, M. J. (2004), "The Serial-Parallel Dilemma: A Case Study in a Linkage of Theory and Method," *Psychonomic Bulletin and Review*, Vol. 11, pp. 391–418.

Tsao, Y. C., Drury, C. G., and Morawski, T. B. (1979), "Human Performance in Sampling Inspection," *Human Factors*, Vol. 21, No. 1, pp. 99–105.

Wallace, J. S., and Fisher, D. L. (1998), "Sound Localization: Theory and Practice," *Human Factors*, Vol. 40, pp. 50–68.

Wiener, E. L. (1975), "Individual and Group Differences in Inspection," in *Human Reliability in Quality Control*, C. G. Drury and J. G. Fox, Eds, Taylor & Francis, London, pp. 19–30.

Wiker, S. F., Langolf, G. D., and Chaffin, D. B. (1989), "Arm Posture and Human Movement Capability," *Human Factors*, Vol. 31, pp. 421–441.

Wolfe, J. M. (1994), "Guided Search 2.0: A Revised Model of Visual Search," *Psychonomic Bulletin and Review*, Vol. 1, No. 2, pp. 202–238.

Wolfe, J. J. (2007), "Guided Search 4.0: Current Progress with a Model of Visual Search," in *Integrated Models of Cognitive Systems*, W. Gray, Ed., Oxford, New York, pp. 99–119.

# CHAPTER 34

# HUMAN SUPERVISORY CONTROL

**Thomas B. Sheridan**
**Massachusetts Institute of Technology**
**Cambridge, Massachusetts**

## 1 DEFINITIONS OF HUMAN SUPERVISORY CONTROL

Human supervisory control is construct, formally something constructed by the mind: a theoretical entity, a working hypothesis or concept pertaining to a relatonaship between a human and a machine or physical system. It is not by itself a normative or predictive model, though it is descriptive of relationships between system elements where both human and computer actively interact. The word "human" is added because the term "supervisory control" is sometimes used by control engineers to describe software agents that aid in system measurement.

This chapter is not a comprehensive or even-handed review of the literature in human–robot interaction, monitoring, diagnosis of failures, human error, mental workload, or other closely related topics. Sheridan (1992, 2002), Sarter and Amalberti (2000), Degani (2004), and Sheridan and Parasuraman (2006) cover these aspects more fully.

The term human supervisory control is derived from the close analogy between the characteristics of a human supervisor's interaction with subordinate human staff members and a person's interaction with "intelligent" automated subsystems. A supervisor of people gives directives that are understood and translated into detailed actions by staff members. In turn, staff members aggregate and transform detailed information about process results into summary form for the supervisor. The degree of intelligence of staff members determines the supervisor's willingness to delegate. Automated subsystems permit the same sort of interaction to occur between a human supervisor and the process (Ferrell and Sheridan, 2010). Supervisory control behavior is interpreted to apply broadly to include vehicle control (aircraft and spacecraft, ships, highway and undersea vehicles), continuous process control (oil, chemicals, power generation), robots and discrete tasks (manufacturing, space, undersea, mining), and medical and other human–machine systems.

In a strictest definition, the term human supervisory control (or just supervisory control as often used in the present context) indicates that one or more human operators are setting initial conditions for intermittently adjusting and receiving information from a computer that itself closes an inner control loop through electromechanical sensors, effectors, and the task environment. In a broader sense, supervisory control means interaction with a computer to transform data or to produce control actions. Figure 1 compares supervisory control with direct manual control (Figure 1a) and full automatic control (Figure 1e). Figures 1c and 1d characterize supervisory control in the strict formal sense; Figure 1b characterizes supervisory control in the latter (broader) sense.

The essential difference between these two characterizations of supervisory control is that in the first and stricter definition the computer can act on new information independent of and with only blanket authorization and adjustment from the supervisor; that is, the computer implements discrete sets of instructions by itself,
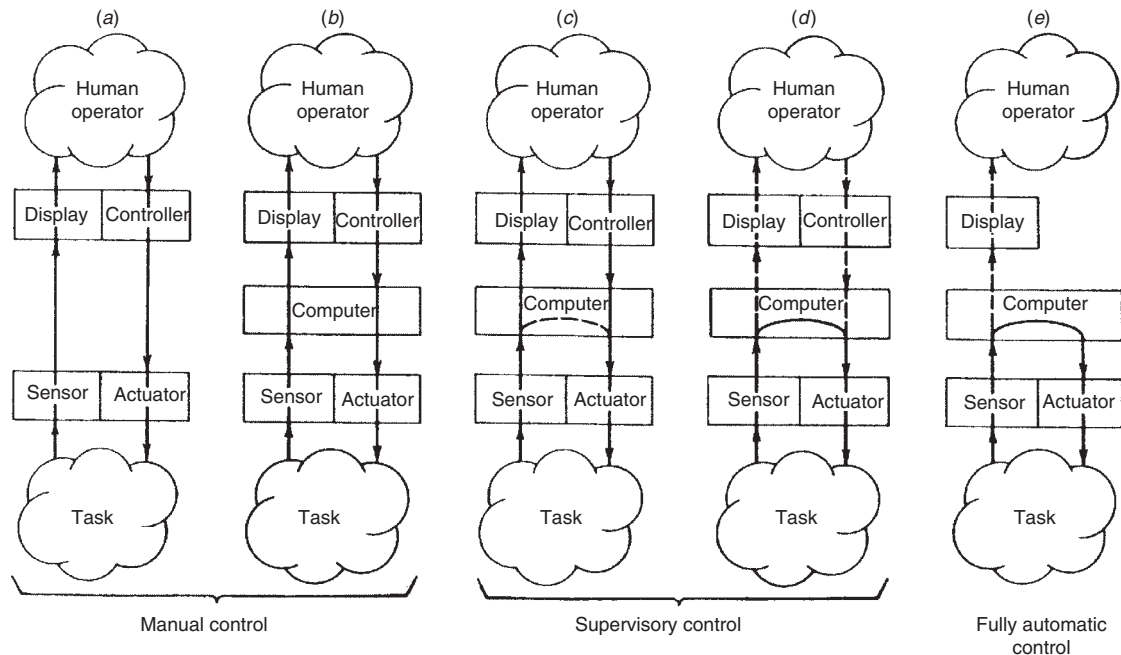
**Figure 1** Supervisory control as related to direct manual control and full automation.

closing the loop through the environment. In the second definition the computer's detailed implementations are open loop; that is, feedback from the task has no effect on computer control of the task except through the human operator. The two situations may appear similar to the supervisor, since he or she always sees and acts through the computer (analogous to a staff) and therefore may not know whether it is acting open loop or closed loop in its fine behavior. In either case the computer may function principally on the efferent or motor side to implement the supervisor's commands (e.g., do some part of the task entirely and leave other parts to the human or provide some control compensation to ease the task for the human). Alternatively, the computer may function principally on the display side (e.g., to integrate and interpret incoming information from below or to give advice to the supervisor as to what to do next, as with an "expert system"). Or it may work on both the efferent and afferent sides.

## 2 SOME HISTORY

The 1940s saw human factors engineering come into being to ensure that soldiers could operate machines in World War II. In the 1950s human factors emerged as a professional field, first in essentially empirical "knobs and dials" form, concentrating on the human–machine interface, accompanied by ergonomics, which focused on the physical properties of the workplace. This was

supported over the next decade by the theoretical underpinnings of human–machine systems theory and modeling (Sheridan and Ferrell, 1974). Such theories included control, information, signal detection, and decision theories originally developed for application to physical systems but now applied explicitly to the human operator. As contrasted with human factors engineering at the interface, human–machine systems analysis considers characteristics of the entire causal "loop" of decision, communication, control, and feedback—through the operator's physical environment and back again to the human.

From the late 1950s the computer began to intervene in the causal loop: electronic compensation and stability augmentation for control of aircraft and similar systems, electronic filtering of signal patterns in noise, and electronic generation of simple displays. It was obvious that if vehicular or industrial systems were equipped with sensors that could be read by computers and, by motors that could be driven by computers, then, even though the overall system was still very much human controlled, control loops between those sensors and motors could be closed automatically. Thus, the chemical plant operator was relieved of keeping the tank at a given level or the temperature at a reference; he or she needed only to set in that desired level or temperature signal from time to time. So, too, after the autopilot was developed for the aircraft, the human pilot needed only to set in the desired altitude to heading; an automatic system would strive to achieve this reference, with the pilot monitoring to ensure that the aircraft did in fact go where desired.

The automatic building elevator, of course, has been in place for many years and is certainly one of the first implementations of supervisory control. Recently, developers of new systems for word processing and handling of business information (i.e., without the need to control any mechanical processes) have begun thinking along supervisory control lines.

The full generality of the idea of supervisory control came to the author and his colleagues (Sheridan, 1960; Ferrell and Sheridan, 1967) as part of research on how people on Earth might control vehicles on the moon through round-trip communication time delays (imposed by the speed of light). Under such constraint, remote control of lunar roving vehicles or manipulators was shown to be possible only by performing in "move-and-wait" fashion. This means that the operator can commit only to a small incremental movement open loop, that is, without feedback (which actually is as large a movement as is reasonable without risking collision or other error), then stopping and waiting one delay period for feedback to "catch up," then repeating the process in steps until the task is completed.

Experimental attempts to drive or manipulate continuously this way only produced instability, as simple control theory predicts (i.e., where loop gains exceed unity at a frequency such that the loop time delay is one half-cycle, instead of errors being nulled out, they are only reinforced). Performing remote manipulation with delayed force feedback was later shown by Ferrell (1967) to be essentially impossible since forces at unexpected times act as significant disturbances to produce instability. At least the visual feedback can be ignored by the operator.

It was shown that if, instead of the human operator remaining within the control loop, he or she communicates a goal state relative to the remote environment, and if the remote system incorporates the capability to measure proximity to this goal state, the achievement of this goal state can be turned over to the remote subordinate control system for implementation. In this case there is no delay in the control loop implementing the task, and thus there is no instability.

There necessarily remains, of course, a delay in the supervisory loop. This delay in the supervisor's confirmation of desired results is acceptable as long as (1) the subgoal is a sufficiently large "bite" of the task, (2) the unpredictable aspects of the remote environment are not changing too rapidly (i.e., disturbance bandwidth is low), and (3) the subordinate automatic system is trustworthy.

Under these conditions and as computers gradually become more capable both in hardware and software (and as "machine intelligence" finally makes its real if modest appearance), it is evident that telemetry transmission delay is in no way a prerequisite to the usefulness of supervisory control. The incremental goal specified by the human operator need not be simply a new steady-state reference for a servomechanism (as in resetting a thermostat) in one or even several dimensions (e.g., resetting both temperature and humidity or commanding a manipulator endpoint to move to a new position, including three translations and three rotations

relative to its initial position). Each new goal statement can be the specification of an entire trajectory of movements (as the performance of a dance or a symphony) together with programmed branching conditions (what to do in case of a fall or other unexpected event).

In other words, the incremental goal statement can be a program of instructions in the full sense of a computer program which make the human supervisor an intermittent real-time computer programmer, acting relative to the subordinate computer much the same as a teacher or parent or boss behaves relative to a student or a child or subordinate worker. The size and complexity of each new program are necessarily a function of how much the computer can (be trusted to) cope with in one bite, which in turn depends on the computer's own sophistication (knowledge base) and the complexity (uncertainty) of the task.

## 3   EXAMPLES OF HUMAN SUPERVISORY CONTROL IN CURRENT TECHNOLOGICAL SYSTEMS

While supervisory control first evolved in delayed feedback situations such as controlling robots on the moon from Earth, it has grown to encompass a wide variety of other systems and for different reasons that have mostly to do with what humans do best (setting goals) and what computers do best (routine execution of conrol actions based on sensed feedback).

Supervisory control is now found in various forms in many industrial, military, medical, and other contexts. However, this form of human interaction with technology is still relatively little recognized or understood in a formal way by system designers who want to take maximum advantage of automation yet want to benefit from the intelligence of the human agent.

Aircraft autopilots are now "layered," meaning that the pilot can select among various forms and levels of control. At the lowest level the pilot can set in a new heading or rate of climb. Or he or she can program a sequence of heading changes at various waypoints or a sequence of climb rates initiated at various altitudes or program the inertial guidance system to take the aircraft to a given runway at a distant city. Given the existence of certain ground-based equipment, the pilot can program an automatic landing on a given runway, and so on. The pilot not only can set commands for different control models but also can also modify different modes of display: how information is presented. Sheridan (2002) reviews how such automation is creeping into the aircraft flight deck. Sarter and Amalberti (2000) describe the modern flight management system in some detail.

Efforts now underway by governments in both the United States and Europe are major technological upgrades of the air traffic control systems. In the United States it is called NextGen (for NextGeneration Air Transportation System), and in the European Community it is called Single European Sky, or SESAR. The two efforts are being coordinated, and in both cases involve the introduction of much new automation and

**Table 1 Some NextGen Flight Operations Using Supervisory Control**

Negotiating four-dimensional (4D) (three in space, one in time) flight trajectories shortly before pushback

Dealing with off-nominal aircraft on the airport surface

Controller/pilot use of digital data communication (Datalink)

Traffic flow manager use of new capacity/flow/weather models

Aircraft operation conflict and resolution responsibilities

Responding to aircraft deviation from their assigned 4D trajectories

Weather conflict and decision to reroute around weather patterns

Effecting a new ''best equipped–best served'' policy

Dynamic reconfiguration of en route or terminal airspace

Merging and spacing in terminal airspace

Setting up for continuous curved rather than step-down descent

Pairing for descent to parallel runways

supervisory control, for example in the flight operations listed in Table 1 (Sheridan, 2010).

The unmanned aeronautical vehicle (UAV) is now subsuming an ever greater role in military operations and soon will do the same in domestic airspace to monitor national borders, inspect crops, and possibly eventually carry freight. UAVs are typically flown by setting successive waypoints in 3D space, a supervisory function.

Supervisory control of a simpler sort is now evident in the cruise control system of current automobiles and trucks and is being upgraded in the form of "advanced" or "intelligent" cruise control, wherein a radar detector controls speed to maintain a safe distance behind a leading vehicle.

In modern hospital operating rooms, intensive care units, and ordinary patient wards there are numerous supervisory control systems at work. The modern anesthesiology workstation is a good example. Drugs in liquid or gaseous form are pumped into the patient at rates programmed by the anesthesiologist and by sensors monitoring patient respiration heart rate and other variables.

Modern chemical and nuclear plants can be programmed to perform heating, mixing, and various other processes according to a time line and including various sensor-based conditions for shutting down or otherwise aborting the operation. Nandi and Ruhe (2002) describe the use of supervisory control in sintering furnaces. Seiji et al. (2001) provide an extensive review of modern supervisory control in nuclear power plants.

Robots of all kinds are being developed: for industrial manufacturing (e.g., for both inspection and assembly of products on assembly lines); for space (e.g., planetary rovers); for undersea applications (e.g., in the British Petroleum oil spill and oceanographic research); for security applications (e.g., inspecting threatening packages in airports and other public places);

military applications (e.g., detonating improvised explosive devices); home cleaning applications (e.g., cleaning swimming pools, vacuuming carpets); offices (e.g., delivering mail); and hospitals (e.g., for minimally invasive surgery). Most of these robots have mobility capability; some have arms for manipulation. Almost all embody supervisory control, at least in primitive form.

Many of the examples cited above characterize the first or stricter definition of supervisory control previously given (Figures 1c and d), where the computer, once programmed, makes use of its own artificial sensors to ensure completion of the tasks assigned. Many familiar systems, such as automatic washing machines, dryers, dishwashers, or stoves, once programmed, perform their operations open loop; that is, there is no measurement or knowledge of results. If the task can be performed in such open-loop fashion, and if the human supervisor can anticipate the task conditions and is good at selecting the right open-loop program, there is no reason not to employ this approach. To the human supervisor, whether the lower level implementation is open or closed loop is often opaque and/or of no concern; the only concern is whether the goal is achieved satisfactorily. For example, a programmable microwave oven without the temperature sensor in place operates open loop, whereas the same oven with the temperature sensor operates closed loop. To the human supervisor or programmer, they look the same.

A very important aspect of supervisory control is the ability of the computer to "package" information for visual display to the human supervisor, including data from many sources; from the past, present, or even predicted future; and presented in words, graphs, symbols, pictures, or some combination. Ubiquitous examples of such integrated displays are so-called *decision support tools* in aircraft and air traffic control, chemical and power plants, and various other industrial or military settings too numerous to review here. General interest in supervisory displays became evident in the mid-1970s (Edwards and Lees, 1981; Sheridan and Johannsen, 1976; Wiener and Curry, 1980; Sheridan and Hennessy, 1984).

## 4 SUPERVISORY ROLES AND HIERARCHY

The human supervisor's roles are (1) planning offline what task to do and how to do it; (2) teaching (or programming) the computer what was planned; (3) monitoring the automatic action online to make sure that all is going as planned and to detect failures; (4) intervening, which means the supervisor takes over control after the desired goal state has been reached satisfactorily or interrupts the automatic control in emergencies to specify a new goal state and reprogram a new procedure; and (5) learning from experience so as to do better in the future. These are usually time-sequential steps in task performance.

We may view these steps as being within three nested loops, as shown in Figure 2. The innermost loop, monitoring, closes on itself; that is, evidence of something interesting or completion of one part of the cycle of

| SUPERVISORY STEP | ASSOCIATED MENTAL MODEL | ASSOCIATED COMPUTER AID |
|---|---|---|
| 1. PLAN | | |
| a) Understand controlled process | Physical variables: transfer relations | Physical process training aid |
| b) Satisfice objectives | Aspirations: preferences and indifferences | Satisficing aid |
| c) Set general strategy | General operating procedures and guidelines | Procedures training and optimization aid |
| 2. TEACH | | |
| a) Decide and test control actions | Decision options: state-procedure-action implications; expected results of control actions | Procedures library; action decision aid (in-situ simulation) |
| b) Decide, test, and communicate commands | Command language (symbols, syntax, semantics) | Aid for editing commands |
| 3. MONITOR AUTOMATION | | |
| a) Acquire, calibrate, and combine measures of process state | State information sources and their relevance | Aid for calibration and combination of measures |
| b) Estimate process state from current measure and past control actions | Expected results of past actions | Estimation aid |
| c) Evaluate process state: detect and diagnose failure or halt | Likely modes and causes of failure or halt | Detection and diagnosis aid for failure or halt |
| 4. INTERVENE | | |
| a) If failure: execute planned abort | Criteria and options for abort | Abort execution aid |
| b) If error benign: act to rectify | Criteria for error and options to rectify | Error rectification aid |
| c) If normal end of task: complete | Options and criteria for task completion | Normal completion execution aid |
| 5. LEARN | | |
| a) Record immediate events | Immediate memory of salient events | Immediate record and memory jogger |
| b) Analyze cumulative experience; update model | Cumulative memory of salient events | Cumulative record and analysis |

**Figure 2**  Functional and temporal nesting of supervisory roles.

the monitoring strategy leads to more investigation and monitoring. We might include minor online tuning of the process as part of monitoring. The middle loop closes from intervening back to teaching; that is, human intervention usually leads to programming of a new goal state in the process. The outer loop closes from learning back to planning; intelligent planning for the next subtask is usually not possible without learning from the last one.

The three supervisory loops operate at different time scales relative to one another. Revisions in fine-scale monitoring behavior take place at brief intervals. New programs are generated at somewhat longer intervals. Revisions in significant task planning occur only at still longer intervals. These differences in time scale further justify Figure 2.

More and more a multiplicity of computers are used in a supervisory control system, as shown in Figure 3. One typically large computer is in the control room to generate displays and interpret commands. This can be called a human-interactive computer (HIC), part of a human-interactive system (HIS). It in turn forwards that command to various microprocessors that actually close individual control loops through their own associated sensors and effectors. The latter can be called task-interactive computers (TICs), each part of its own task-interactive system (TIS).

The HIC is conceived to be a large enough computer to communicate in a human-friendly way using near-natural language, good graphics, and so on. This includes being able to accept and interpret commands and to give the supervisor useful feedback. The HIC should be able to recognize patterns in data sent up to it from below and decide on appropriate algorithms for response, which it sends down as instructions. Eventually, the HIC should be able to run "what would happen if . . ." simulations and be able to give useful advice from a knowledge base, that is, include an expert system.

The HIC, located near the supervisor in a control room or cockpit, may communicate across a barrier of time or space with a multiplicity of TICs, which probably are microprocessors distributed throughout the plant or vehicle. The latter are usually coupled intimately with artificial sensors and actuators in order to deal in low-level language and to close relatively tight control loops with objects and events in the physical world.

The human supervisor can be expected to communicate with the HIC intermittently in information "chunks" (alphanumeric sentences, icons, etc.) while the task communicates with the TIC continuously in computer language at the highest possible bit rates. The availability of these computer aids means that the human supervisor, while retraining the knowledge-based behavior function, is likely to download some of the rule-based programs and almost all of the skill-based programs into the HIC. The HIC, in turn, should download a few of the rule-based programs, and most of the skill-based programs, to the appropriate TICs.

Figure 4 presents the functions of Figure 2 in the form of a flowchart. Each supervisory function is shown above, and the (usually multiple) automated subsystems of the TIC are shown below. Normally, for any given task, the planning and learning roles are performed offline relative to the online human-mediated and automatic operations of the other parts of system and therefore are shown at the top with light lines connecting them to the rest of the system. Teaching precedes monitoring on the first cycle but thereafter follows monitoring and intervening (as necessary) within the intermediate loop. The inner loop monitoring role is carried out within the "estimate state" and "allocate attention" boxes.

Allocation of functions between the human and the machine need not be fixed. There have been numerous papers discussing the potential for dynamic allocation—where the allocation changes as a function of the flow of demands and the workload of the two entities (see, e.g., Sheridan, 1997). In the sections that follow the various supervisory roles are discussed in more detail, bringing
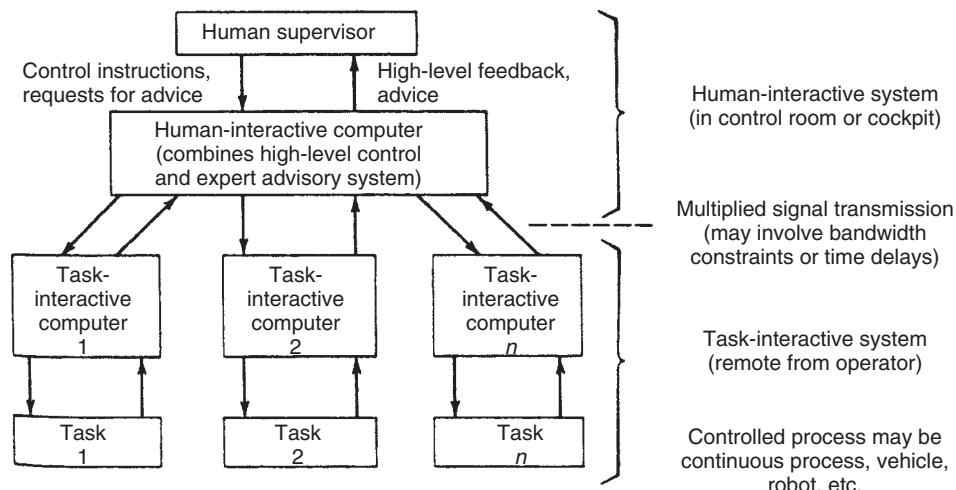


**Figure 3** Hierarchical nature of supervisory control.
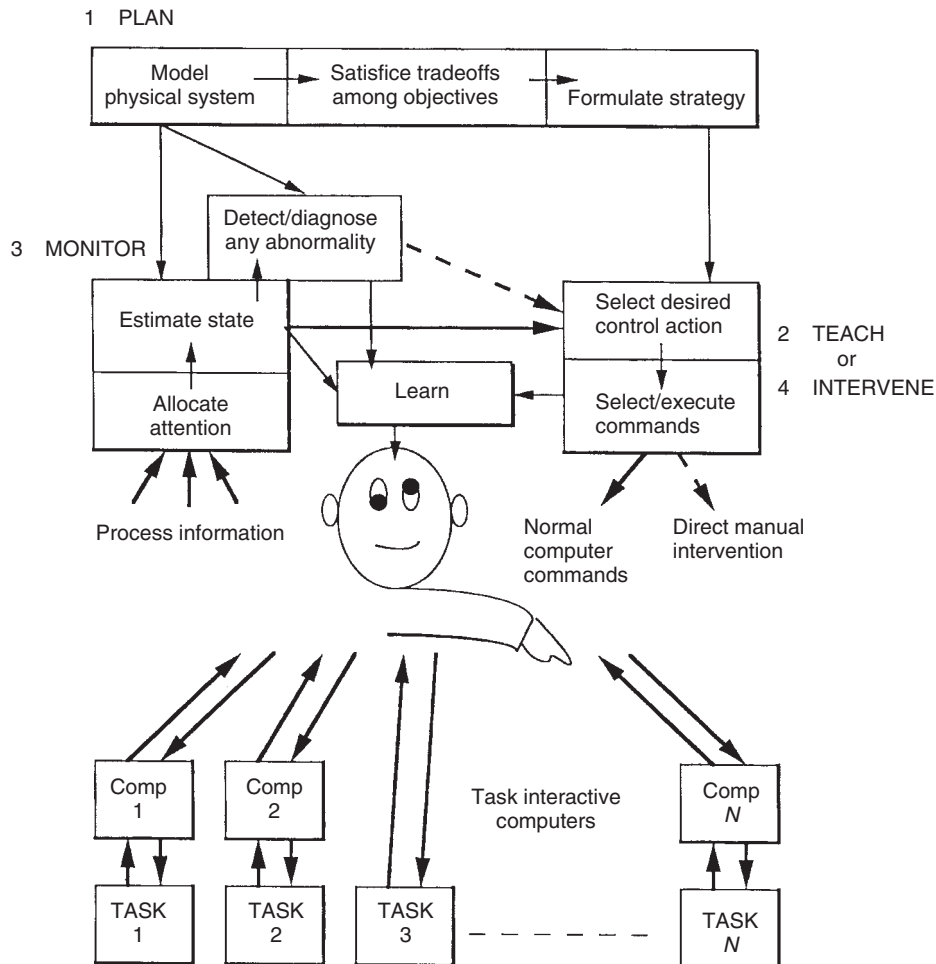
**Figure 4**  Flowchart of supervisor functions (including both mental models and decision aids). (From Sheridan, 1992.)

in examples of research problems and prototype systems to aid the supervisor in these roles.

## 5  SUPERVISORY LEVELS AND STAGES

Supervisory control may involve varying degrees of computer aiding in acquiring information and executing control, as in Table 2). This "level of automation" idea, originally presented in Sheridan and Verplank (1978) with 10 rather than 8 levels, has been picked up and used by others in various ways. Parasuraman et al. (2000) added the idea that the successive stages of information acquisition, information analysis, action decision, and action implementation are usually automated to different degrees. The best degree of automation is seldom the same at the various stages.

Figure 5 is an example of how, in the writer's opinion, the Federal Aviation Administrations's NextGen

**Table 2  Scale of Degrees of Automation**

1. The computer offers no assistance; the human must do it all.

2. The computer suggests alternative ways to do the task.

3. The computer selects one way to do the task and

4. Executes that suggestion if the human approves or

5. Allows the human a restricted time to veto before automatic execution or

6. Executes the suggestion automatically, then necessarily informs the human, or

7. Executes the suggestion automatically, then informs the human only if asked.

8. The computer selects the method, executes the task, and ignores the human.

project will automate operations at the four stages in the midterm and far term as compared to the current levels.

Experience with military systems calls attention to the appropriateness of various levels of automation (Cummings, 2005):

The Patriot missile system has a history of friendly fire incidents that can at least be partially attributed to a lack of understanding of human limitations in supervisory control. . . .

The Patriot missile has two modes: semiautomatic (management by consent, (level 4 above—an operator must approve a launch) and automatic (management by exception —the operator is given a period of time to veto the computer's decision (level 5 above). However, in practice the Patriot is typically left in the automatic mode and the friendly fire incidents are believed to be a result of problems in the automatic mode. There are known "ghosting" problems with the Patriot radar: because operations are in close proximity to other Patriot missile batteries, false targets will appear on a Patriot operator's screen. Under the automatic mode (management by exception), operators are given approximately 15 seconds to reject the computer's decision, which is insufficient to solve both false targeting problems as well as adequately address friend or foe concerns through any other means of communication. After the accident investigations, the US Army admitted that there is no standard for Patriot training, autonomous operations procedures (automatic mode) are not clear, and that operators commonly lose situational awareness of air tracks).

## 6 PLANNING AND LEARNING: COMPUTER REPRESENTATION OF RELEVANT KNOWLEDGE

The first and fifth supervisory roles described previously, planning and learning, may be considered together since they are similar activities in many ways. Essentially, in the planning role the supervisor asks "What would happen if . . . ?" questions of the accumulated knowledge base and considers what the implications are for hypothetical control decisions. In learning, the supervisor asks "What did happen?" questions of the database for the more recent subtasks and considers whether the initial assumptions and final control decisions were appropriate.

The designer of an automatic control system or manual control system must ask: "What variables do I wish to make do what, subject to what constraints and what criteria?" The planning role in supervisory control requires that the same kinds of questions be answered, because, in a sense, the supervisor is redesigning an automatic control system each time that he or she programs a new task and goal state. Absolute constraints

on time, tools, and other resources available need to be clear, as do the criteria of trade-off among time, dollars and resources spent, accuracy, and risk of failure.

Just as computer simulation figures into planning, it also figures into supervisory control—the difference being that such simulation may more likely be subjected to time stress in supervisory control. Simulation requires acquiring some idea of how the process (system to be controlled) works, that is, a set of equations relating the various controllable variables, the various uncontrollable but measurable variables (disturbances), and the degree of unpredictability (noise) on measured system response variables. This is a common representation of knowledge. Given measured inputs and outputs, there are well-established means to infer the equations if the processes are approximately linear and differentiable.

Once such a model is in place, the supervisor can posit hypothetical inputs and observe what the outputs would be. Also, one may use such a process model as an "observer" (in the sense of modern control theory). Namely, when control signals are put into both the model and actual processes and the model parameters are then trimmed to force certain model outputs to conform to corresponding actual process outputs that can be measured (Figure 6), other process outputs that are inconvenient to measure may be estimated ("observed") from the model. Just as this is a theoretical prerequisite to optimal automatic control of physical systems, so it is likely to be a useful practice to aid humans in supervisory control (Sheridan, 1984).

A different type of knowledge representation is that used by the artificial intelligence (AI) community. Here knowledge is usually couched in the form of if–then logical statements called production rules, semantic association networks, and similar forms. The input to a simulated program usually represents in cardinal numbers a hypothetical physical input to a simulated physical system. In contrast, the input to the AI knowledge base can be a question about relationships for given data or a question about data for given relationships. This can be in less restrictive ordinal form (e.g., networks of diadic relations) or in nominal form (e.g., lists).

Currently, there is great interest in how best to transfer expertise from the human brain (knowledge representation, mental model) into the corresponding representation or model within the computer, how best to transfer it back, and when to depend on each of those sources of information. This research on mental models has a lively life of its own (Falzon, 1982; Gentner and Stevens, 1983; Rouse and Morris, 1984; Sheridan, 1984; Moray, 1997) quite independent of supervisory control.

An important aspect of planning is visualization. The now rather sophisticated tool of computer simulation, when augmented by computer graphics, enables remarkable visualization possibilities. When further augmented by human interactive devices such as head-mounted visual and auditory displays and high-bandwidth force-reflecting haptics (mechanical arms), the operator can be made to feel present in a virtual world, as has been popularized by the oxymoron *virtual reality*. Of course, the idea of virtual reality is not new. The original idea
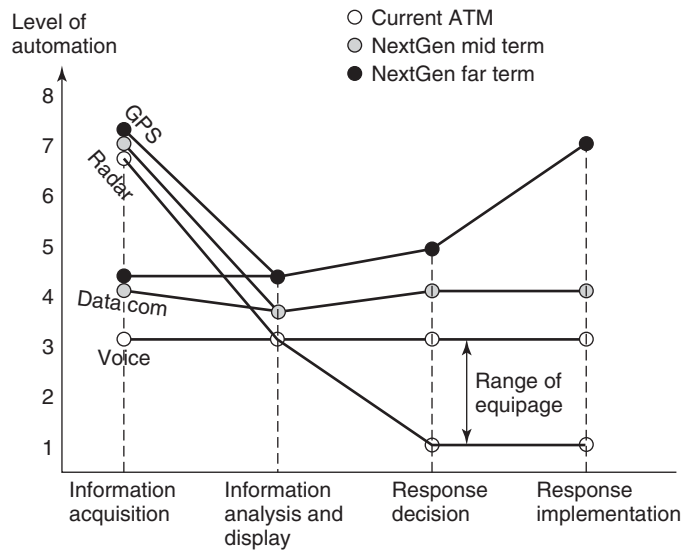
**Figure 5**  Estimated levels of automation of NextGen automation.
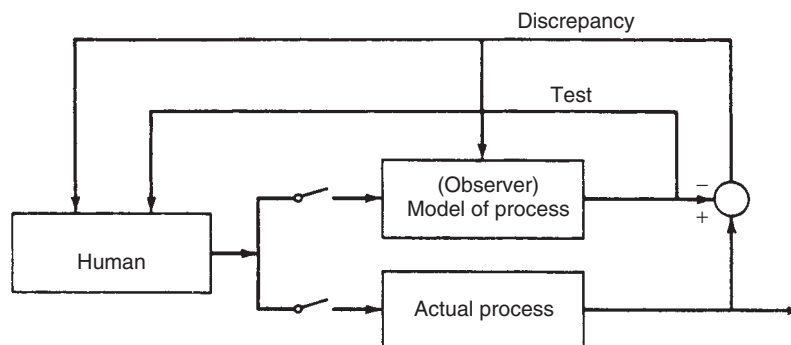


**Figure 6**  Use of computer-based observer as an aid to supervisor.

of Edwin Link's first flight simulators (developed early in the 1940s) was to make the pilot trainee feel as if he or she were flying a real aircraft. First they were instrument panels only, then a realistic out-the-window view was created by flying a servo-driven video camera over a scale model, and finally, computer graphics were used to create the out-the-window images. Now all commercial airlines and military services routinely train with computer display, full-instrument, moving-platform flight simulators. Similar technology has been applied to ship, automobile, and spacecraft control. The salient point for the present discussion is that the new simulation capabilities now permit visualization of alternative plans as well as better understanding of complex state information in situ, during monitoring. That same technology, of course, can be used to convey a sense of presence in an environment that is not simulated but is quite real and

merely remote—communicated via closed-circuit video with cameras slaved to the observer's head.

Supervisory aiding in planning of the moves of a telerobot is illustrated by the work of Park (1991). His computer graphic simulation let a supervisor try out moves of a telerobot arm before committing to the actual move. He assumed that for some obstacles the positions and orientation were already known and represented in a computer model. The user commanded each straight-line move to a subgoal point in three-dimensional space by designating a point on the floor or the lowest horizontal surface (such as a tabletop) by moving a cursor to that point (say, A in Figure 7a) and clicking, then lifting the cursor by an amount corresponding to the desired height of the subgoal point (say, A) above that floor point and observing on the graphic model a blue vertical line being generated from the floor point to the subgoal point in space. This process was repeated
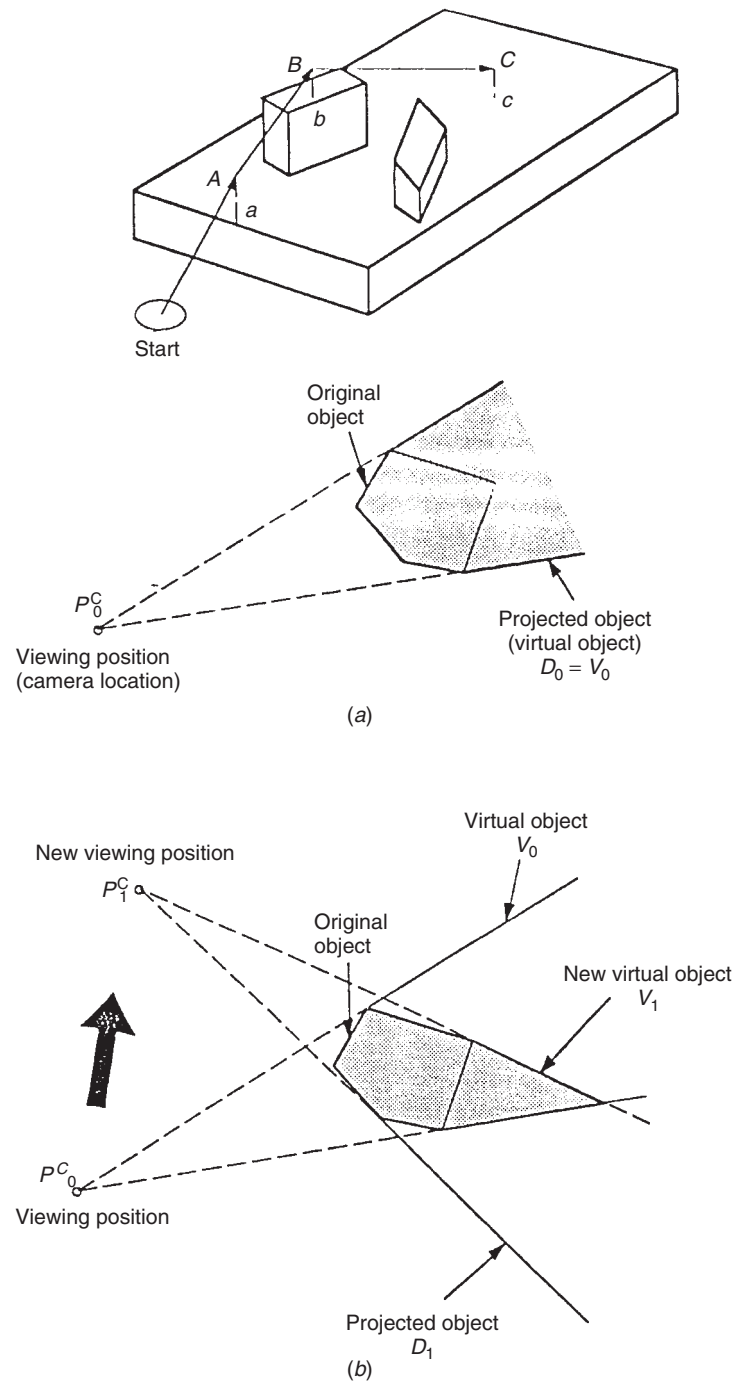
**Figure 7** Park's display of computer aid for obstacle avoidance: (*a*) human specification of subgoal points on graphic model; (*b*) generation of virtual obstacles for a single viewing position (above) and a pair of viewing positions (below). (From Park, 1991.)

for successful subgoal points (say, B and C). Using the computer display, the user could view the resulting trajectory model from any desired perspective (although the "real" environment could be viewed only from the perspective provided by the video camera's location). Either of two collision avoidance algorithms could be invoked: a detection algorithm that indicated where on some object a collision occurred as the arm was moved from one point to another or an automatic avoidance algorithm that found (and drew on the computer screen) a minimum-length, no-collision trajectory from the starting point to the new subgoal point. Park's aiding scheme also allowed new observed objects to be added to the model by graphically "flying" them into geometric correspondence with the model display. Another aid was to generate virtual objects for any portion of the environment in the umbral region (not visible) after two video views (Figure 7b). In this case the virtual objects were treated in the same way in the model and in the collision avoidance algorithms as the visible objects. Experiments with this technique showed that it was easy to use and that it avoided collisions.

At the extreme of time desynchronization is recording an entire task on a simulator, then sending it to the telerobot for reproduction. This might be workable when one is confident that the simulation matches the reality of the telerobot and its environment or when small differences would not matter (e.g., in programming telerobots for entertainment). Doing this would certainly make it possible to edit the robot's maneuvers until one was satisfied before committing them to the actual operation. Machida et al. (1988) demonstrated such a technique by which commands from a master–slave manipulator could be edited much as one edits material on a video-tape recorder or a word processor. Once a continuous sequence of movements had been recorded, it could be played back either forward or in reverse at any time rate. It could be interrupted for overwrite or insert operations. Their experimental system also incorporated computer-based checks for mechanical interference between the robot arm and the environment.

A number of planning aids are manifest in modern air traffic control. Computers are used to show the expected arrival of aircraft at airports and the gaps between them. This helps the human controller to command minor changes in aircraft speed or flight path to smooth the flow. The center TRACON automation system (CTAS) assists in providing an optimal schedule and three-dimensional spacing. Other systems use radar data to project ahead and alert the controller to potential conflicts (violation of aircraft separation standards) (Wickens et al., 1997). NextGen promises a number of additional decision aiding displays, such as those listed in Table 1.

One aspect of supervisory control that is often not planned and is taken for granted (and where learning can be painful) is team coordination in distributed decision making. NextGen has thankfully recognized the problem and has established research efforts into what in that context is called "cooperative air traffic management." In the military operations context Cummings (2005) provides an example:

On April 14, 1994, two US Army Black Hawk helicopters were transporting U.S., French, British, and Turkish commanders, as well as Kurdish para-military personnel across this zone when two US F-15 fighters shot them down, killing all 26 on board. The Black Hawks had previously contacted and received permission from the AWACs to enter the no-fly zone. Yet despite this, AWACs confirmed that there should be no flights in the area when the F-15s misidentified the US helicopters as Iraqi Hind helicopters. The teamwork displayed in this situation was a significant contributing factor to the friendly fire incident, as the F-15s never learned from AWACs that a friendly mission was supposed to be in the area. It was later determined that the F-15 wingman backed up the other F-15's decision that the targets were Iraqi forces despite being unsure, which was yet another breakdown in communication. Each team member did not share information effectively, resulting in the distributed decision making of the AWACs and F-15s pilots to come to incorrect and fatal conclusions

## 7  TEACHING THE COMPUTER

Teaching or programming a task, including a goal state and a procedure for achieving it and including constraints and criteria, can be formidable or quite easy, depending on the command hardware and software. By command hardware is meant the way in which human response (hand, foot, or voice) is converted to physical signals to the computer. Command hardware can be either analogic or symbolic. Analogic means that there is a spatial or temporal isomorphism among human response, semantic meaning, and/or feedback display. For example, moving a control up rapidly to increase the magnitude of a variable quickly, which causes a display indicator to move up quickly, would be a proper analogic correspondence.

Symbolic command, by contrast, is accomplished by depressing one or a unique series of keys (as typing words on a typewriter) or uttering one or a series of sounds (as in speaking a sentence), each of which has a distinguishable meaning. For symbolic commands a particular series or concatenation of such responses has a different meaning from other concatenations. Spatial or temporal correspondence to the meaning or desired result is not a requisite. Sometimes analogic and symbolic can be combined: for example, where up–down keys are both labeled and positioned accordingly.

It is natural for people to intermix analogic and symbolic commands or even to use them simultaneously. Typical industrial robots are taught by a combination of grabbing hold and leading the endpoint of the manipulator around in space relative to the workpiece, at the same time using a switch box on a cable (a teach pendant) to key in codes for start, stop, speed, and so on, between various reference positions. This happens, for example, when a person talks and points at the same time or plays the piano and conducts a choir with his or her head or free hand.

In regard to teaching the computer, Ferris et al (2010) use the term *directability*, which they define as ability to direct efficiently and safely the activities of the automation. They point out that the interface must be designed to avoid what Norman 1986 has called the gulf of execution, "where an operator struggles with identifying and operating the proper controls and commands to translate an intended action into the machine's language." They point out that problems occur when, for example, different aircraft employ automation controls with similar shape, feel, and/or location that activate different systems or require different manipulations (Abbott et al., 1996). Such inconsistencies can leave pilots who transition between aircraft or airlines highly vulnerable to errors, especially when under stress.

Supervisory command systems have been developed for mechanical manipulators that utilize both analogic and symbolic interfaces with the supervisor and that enable teaching to be both rapid and available in terms of high-level language. Brooks (1979) developed such a system, which he called SUPERMAN, which allows the supervisor to use a master arm to identify objects and demonstrate elemental motions. He showed that even without time delay for certain commands, which refer to predefined location, supervisory control that included both teaching and execution took less time and had fewer errors than manual control.

Yoerger (1982) developed a more extensive and robust supervisory command system that enables a variety of arm–hand motions to be demonstrated, defined, called on, and combined under other commands. In one set of experiments, Yoerger compared three different procedures for teaching a robot arm to perform a continuous seam weld along a complex curved workpiece. The end effector (welding tool) had to be kept 1 in. away and retain an orientation perpendicular to the curved surface to be welded and move at constant speed. Yoerger tested his subjects in three command (teaching) modes. The first mode was for the human teacher to move the master (with slave following in master–slave correspondence) relative to the workpiece in the desired trajectory. The computer would memorize the trajectory and then cause the slave end effector to repeat the trajectory exactly. The second mode was for the human teacher to move the master (and slave) to each of a series of positions, pressing a key to identify each. The human would then key in additional information specifying the parameters of a curve to be fit through these points and the speed at which it was to be executed, and the computer would then be called upon for execution. The third mode was to use the master–slave manipulator to contact and trace along the workpiece itself, to provide the computer with the knowledge of the location and orientation of the surfaces to be welded. Then, using the typewriter keyboard, the human teacher would specify the positions and orientations of the end effector relative to the workpiece. The computer could then execute the task instructions relative to the geometric references given.

Identifying the geometry of the workpiece analogically and then giving symbolic instructions relative to it proved the constant winner. The reasons for this advantage apparently are the same as for Brooks's results

described previously, provided of course that the time spent in the teaching loop is sufficiently short.

There are many programming languages for industrial robots, but a lack of standardization of programming methods for robots poses challenges. For example, there are over 30 different manufacturers of industrial robots, so there are also 30 different robot programming languages required.

Some robot programming languages are essentially visual. The software system for the Lego Mindstorms NXT robots is worthy of mention. It is based on and written by Labview. The approach is to start with the program rather than the data. The program is constructed by dragging icons into the program area and adding or inserting them into a sequence. For each icon you then specify the parameters (data). For example, for the motor drive icon you specify which motors and by how much they move.

A *scripting language* is a high-level programming language that is used to control the software application and is interpreted in real time, or "translated on the fly," instead of being compiled in advance. A scripting language may be a general-purpose programming language or it may be limited to specific functions used to augment the running of an application or system program. Some scripting languages, such as RoboLogix, have data objects residing in registers, and the program flow represents the list of instructions, or instruction set, that is used to program the robot. The RoboLogix instruction set is shown in Figure 8.

Programming languages are generally designed for building data structures and algorithms from scratch, while scripting languages are intended more for connecting, or "gluing," components and instructions together. Consequently, the scripting language instruction set is usually a streamlined list of program commands that are used to simplify the programming process and provide rapid application development.

Teaching airplane autopilots is a good example of the teaching role in supervisory control. Modern airplanes can now adjust their throttle, pitch, and yaw damping characteristics automatically. They can take off and climb to altitude autonomously or fly to a given latitude and longitude and can maintain altitude and direction despite wind disturbances. They can approach and land automatically in zero-visibility conditions. To do these tasks, airplanes make use of artificial sensors, motors, and computers programmed in supervisory fashion by pilots and ground controllers. In this sense airplanes are telerobots in the hands of their pilot teachers. In the aviation world the supervising pilot is called a flight manager.

The flight management system (FMS) is the aircraft embodiment of the HIC discussed previously and currently is where the supervisory teaching is done. The typical FMS has a cathode ray tube (CRT) display and both generic and dedicated keysets. More than 1000 modules provide maps for terrain and navigational aids, procedures, and synoptic diagrams of various electrical and hydraulic subsystems. Proposed electronic maps show planned flight route, weather, and other navigational aids. When the pilot enters a certain flight
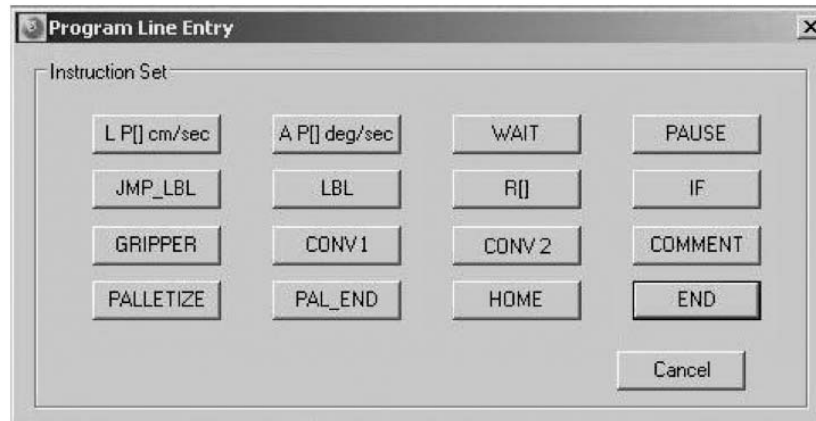
**Figure 8**   RoboLogix instruction set (from Wikipedia).

plan, the FMS can visualize the trajectory automatically and call attention to any waypoints that appear to be erroneous on the basis of a set of reasonable assumptions. Conflict probe displays call the ground controller's attention to incipient separation violations, while cockpit traffic displays do the same for the pilot.

The problem of authority is one of the most difficult (Boehm-Davis et al., 1983). Popular mythology is that the pilot is (or should be) in charge at all times. But when a human turns control over to an automatic system, it is the exception that she or he can do something else for a while (as in the case of setting one's alarm clock and going to sleep). It is also recognized that there are limited windows of opportunity for escaping from the automation (once you get on an elevator you can get off only at discrete floor levels). People are seldom inclined to "pull the plug" unless they receive clear signals indicating that such action must be taken and unless circumstances make it convenient for them to do so. Examples of some current debates follow:

1. Should there be certain states or a certain envelope of conditions for which the automation will simply seize control from the pilot?

2. Should the computer deviate from a programmed flight plan automatically if critical unanticipated circumstances arise?

3. If the pilot programs certain maneuvers ahead of time, should the aircraft execute these automatically at the designated time or location, or should the pilot be called upon to provide further concurrence or approval?

4. In the case of a subsystem abnormality, should the affected subsystem be reconfigured automatically, with after-the-fact display of what has failed and what has been done about it? Or should the automation wait to reconfigure until after the pilot has learned about the abnormality, perhaps been given some advice on the options, and had a chance to take initiative?

It is important to emphasize that simple and ideal command-and-feedback patterns are not to be expected as systems get more complex. In interactions between a human supervisor and his or her subordinates, or a teacher and the students, it can be expected that the teaching process will not be a one-way communication. Some feedback will be necessary to indicate whether the message is understood or to convey a request for clarification on some aspect of the instructions. Further, when the subordinate or student does finally act on the instruction, the supervisor may not understand from the immediate feedback what the subordinate has done and may ask for further details. This is illustrated in Figure 9 by the light arrows, where the bold arrows characterize the conventional direction of information in feedback control.

Teaching a computer for supervisory control actually goes beyond what can be thought of as providing if–then–else instructions for mechanical actions (as with a robot or vehicle), usually called the *control law*. It also includes setting or changing parameters of how properties of the system (states) are measured, how such information is displayed, how the interaction of the system with its environment is modeled or simulated for planning future actions, as well as properties of the control interface. These many options for system parameter change are illustrated in Figure 10.

## 8   MONITORING OF DISPLAYS AND DETECTION OF FAILURES

The human supervisor monitors the automated execution of the task to ensure proper control (Parasuraman, 1987). This includes intermittent adjustment or trimming if the process performance remains within satisfactory limits. It also includes detection of if and when it goes outside limits and the ability to diagnose failures or other abnormalities. The subject of failure detection in human–machine systems has received considerable attention (Rasmussen and Rouse, 1981). Moray (1986) regards such failure detection and diagnosis as the most

FUNCTION OF HUMAN          FORM OF              FUNCTION OF COMPUTER
                          COMMUNICATION
                       (analogic or symbolic)

Command by supervisor

(1)  Generation of command  ————————————▶  (2)  Understanding of command
                            Principal direction

(4)  Clarification of command  ◀————————  (3)  Display of understanding
                            Secondary direction

feedback of state

(2)  Understanding of state  ◀————————  (1)  Display of system state
                            Principal direction

(3)  Query about state  ————————————▶  (4)  Display of clarification
                            Secondary direction

**Figure 9**  Intermediate feedback in command and display. Heavy arrows indicate the conventional understanding of functions. Light arrows indicate critical additional functions that tend to be neglected. (From Sheridan, 1984.)
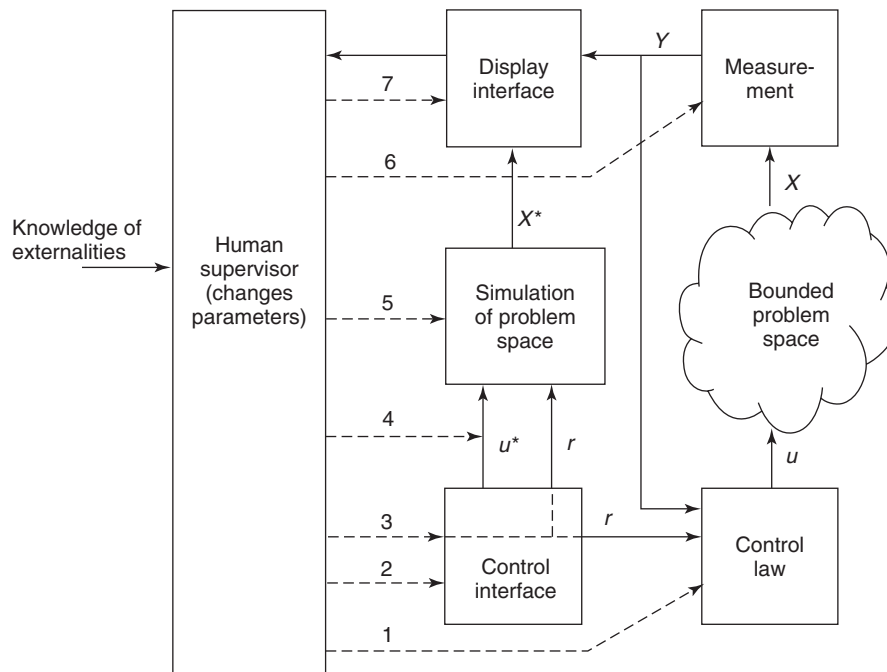


**Figure 10**  Set of options for changing parameters in supervisory control.

important human supervisory role. I prefer the view that all five supervisory roles are essential and that no one can be placed above the others.

The supervisory controller tends to be removed from full and immediate knowledge about the controlled process. The physical processes that he or she must monitor tend to be large in number and distributed widely in space (e.g., around a ship or plant). The physical variables may not be immediately sensible by him or

her (e.g., steam flow and pressure) and may be computed from remote measurements on other variables. Sitting in the control room or cockpit, the supervisor is dependent on various artificial displays to give feedback of results as well as knowledge of new reference inputs or disturbances. These factors greatly affect how he or she detects and diagnoses abnormalities in the process, but whether removal from active participation in the control loop makes it harder (Ephrath and

Young, 1981) or easier (Curry and Ephrath, 1977) remains an open question. Gai and Curry (1978) and Wickens and Kessell (1979, 1981) have studied various psychophysical aspects of this problem.

Ferris et al. (2010) call attention to what they call *observability* (i.e., lack of adequate feedback about targets, actions, decision logic, or operational limits). The concept comports with the same term in control theory, which has to do with how well the internal states of a system can be inferred by knowledge of its external outputs:

> Low observability has been shown to lead to a lack or loss of mode awareness, that is, a lack of knowledge and understanding of the current and future automation configuration and behavior . . . One manifestation of reduced mode awareness are automation surprises, in which pilots detect a discrepancy between actual and expected or assumed automation behavior . . . often resulting from uncommanded or indirect (i.e., without an explicit instruction by the pilot) mode transitions.

These authors also discuss the serious problem of mode awareness and errors. Mode errors of *commission* occur when a pilot executes an action that is appropriate for the assumed but not the actual current mode of the system. Mode errors of *omission* take place when the pilot fails to take an action that would be required given the currently active automation configuration and behavior (Abbott et al., 1996).

Sarter et al. (2007) describe a study in which airline pilots participated in a full-mission 747-400 simulation that included a variety of challenging automation events. Using eye motion instrumentation, they found that pilots monitor basic flight parameters to a much greater extent than visual indications of the automation configuration. More specifically, pilots frequently fail to verify manual mode selections or notice automatic mode changes. In other cases, they do not process mode annunciations in sufficient depth to understand their implications for aircraft behavior.

In traditional control rooms and cockpits the tendency has been to provide the human supervisor with an individual and independent display of each variable and for a large fraction of these to provide a separate additional alarm display that lights up when the corresponding variable reaches or exceeds some value. Thus, modern aircraft may easily have over 1000 displays and modern chemical or power plants 5000 displays. In the writer's experience, in one nuclear plant training simulator, during the first minute of a "loss of coolant accident," 500 displays were shown to have changed in a significant way, with 800 more in the second minute.

Clearly, no human being can cope with so much information coming simultaneously from so many seemingly disconnected sources. Just as clearly, such signals in any real operating system actually are highly correlated. In real-life situations in which we move among people, animals, plants, or buildings, our eyes, ears, and other senses easily take in and comprehend vast amounts of information just as much as in the power plant. Our

genetic makeup and experience enable us to integrate the bits of information from different parts of the retina and from different senses from one instant to the next, presumably because the information is correlated. We say we "perceive patterns" but do not pretend to understand how. In any case the challenge is to design displays in technological systems to somehow integrate the information to enable the human operator to perceive patterns in time and space and across the senses. As with teaching (command), the forms of display may be either analogic (e.g., diagrams, plots) or symbolic (e.g., alphanumerics) or some combination.

In the nuclear power industry the safety parameter display system (SPDS) is now required of all plants in some form. The idea of the SPDS is to select a small number (e.g., 6–10) of variables that tell the most about plant safety status and to display them in integrated fashion such that by a glance the human operator can see whether something is abnormal and, if so, what and to what relative degree. Figure 11 shows an example of an SPDS. It gives the high-level or overview display (a single computer "page"). If the operator wishes more detailed information about one variable or subsystem, he or she can page down (select lower levels). These can be diagrams having lines or symbols that change color or flash to indicate changed status and alphanumeric to give quantitative or more detailed status. These can also be bar graphs or cross plots or integrated in other forms. One novel technique is the Chernoff face (Figure 11*c*), in which the shapes of eyes, ears, nose, and mouth differ systematically to indicate different values of variables, the idea being that facial patterns are easily perceived. Allegedly, the Nuclear Regulatory Commission, fearful that some enterprising designer might employ this technique before it was proven, formally forbade it as an acceptable SPDS.

As noted previously (Figure 3), an important potential of the HIC is for modeling the controlled process. Such a model may then be used to generate a display of observed state variables that cannot be seen or measured directly. Another use is to run the model in fast time to predict the future, given of course that the model is calibrated to reality at the beginning of each such predictive run. A third use, now being developed for application to remote control of manipulators and vehicles in space, helps the human operator cope with telemetry time delays (as shown in Figure 12, wherein video feedback is necessarily delayed by at least several seconds). By sending control signals to a computer model as a basis for superposing the corresponding graphic model on the video, the graphic model will "lead" the video picture and indicate what the video will do several seconds hence. This has been shown to speed up the execution of simple manipulation tasks by 70–80% (Noyes and Sheridan, 1984).

Advances in computer graphics, as driven by the computer game industry, film animation and special effects, and other simulations (virtual reality), have meant that computer displays are theoretically limitless in what they can display: dynamically at high resolution, in color, and on a head-mounted display if that is called for. The challenge for the display designer is then:
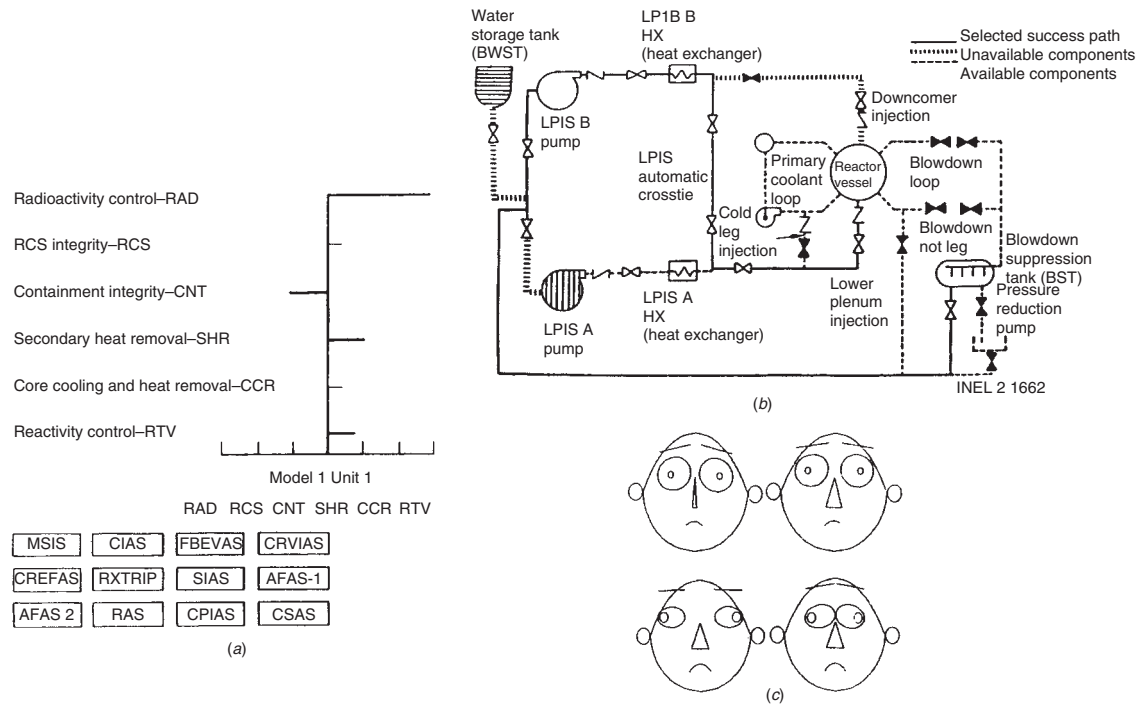
**Figure 11** Safety parameter display system for a nuclear power plant.

What will provide the most effective interaction with the human supervisor?

A final aspect of supervisory monitoring and display concerns format adaptivity—the ability to change the format and/or the logic of the displays as a function of the situation. Displays in aerospace and industrial systems now have fixed formats (e.g., the labels, scales, and ranges are designed into the display). Alarms have fixed set points. However, future computer-generated displays even for the same variables may be different at various mission stages or in various conditions. Thus, formats may differ for aircraft takeoff, landing, and on-route travel and be different for plant startup, full-capacity operation, and emergency shutdown. Some alarms have no meaning or may be expected to go off when certain equipment is being tested or taken out of service. In such a case adaptive formatted alarms may be suppressed or the set points changed automatically to correspond to the operating mode. Future displays and alarms could also be formatted or adjusted to the personal desires of the supervisor to provide any time scale, degree of resolution, and so on, necessary at the time. Ideally, some future displays could adapt based on a running model of how the human supervisor's perception was being enhanced.

A currently popular research challenge is to measure highway vehicle driver task workload and whether driver's use of the potential in-vehicle information distracters, such as cell phone, radio, navigation system, and so on, should be prohibited during busy demands of traffic (Boer, 2000; Llaneras, 2000; Lee et al., 2002). This would make the driver interfaces adaptive.

There are hazards, of course, in allowing emergency displays to be too flexible, to the point where they cause errors rather than preventing them. Mode errors, where the operator believes that he or she is operating in one mode but actually is operating in a different mode, can be dangerous. An example of where flexibility in monitoring displays went awry was in an aircraft accident that occurred in Europe several years ago. In this instance the pilot could ask to have either descent rate (thousands of feet per minute) or descent angle (degrees) presented, and depending on how the model control panel had been set, the number was indicated by two digits displayed at the same location. In this case the pilot forgot which mode he had requested (although that information was also displayed but at a different location). The result was a misreading and a tragic crash.

## 9 INTERVENING AND HUMAN RELIABILITY

Sarter and Woods (2000) and Wiener (1988) write about automation surprises, the tendency of automatic systems to catch the human supervisor off-guard such that the human thinks: What is the automation doing now? What will it do next? How did I get into this mode? Why did it do this? How do I stop the machine from doing this? Why won't it do what I want?

The challenge of surprise is a great one, and there are no easy answers. Computers do what they have been
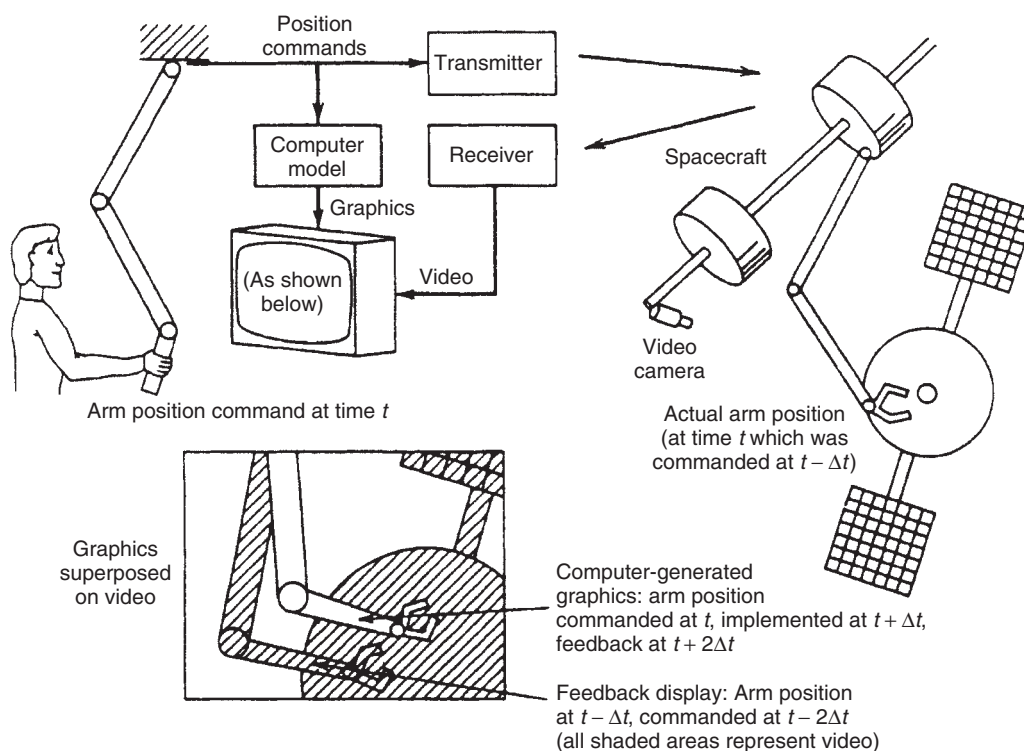
**Figure 12**   Predictor display for delayed telemanipulator.

programmed to do, which is not always what the user intended. User education—toward better understanding of how the system works—is one remedy. Another is to provide error messages that are couched in a language understandable to the operator (not in the jargon of the computer programmer, a problem so familiar to all users of computers). Generally, the solution lies in some form of feedback—to lead the human in making a mild or radical intervention, as appropriate.

The supervisor decides to intervene when the computer has completed its task and must be retaught for the next task, when the computer has run into difficulty and requests of the supervisor a decision as to which way to go, or when the supervisor decides to stop automatic action because he or she judges that system performance is not satisfactory. Intervention is a problem that really has not received as much attention as teaching and monitoring. Yet systems are being planned in which the supervisory operator is expected to receive advice from a computer-based system about remote events and within seconds decide whether to accept the computer's advice (in which case the response is commanded automatically) or reject the advice and generate his or her own commands (in effect, intervene in an otherwise automatic chain of events).

It is at the intervention stage that human error most reveals itself. Errors in learning from past experience, planning, teaching, and monitoring will surely exist. Many of these are likely to be corrected as the supervisor

notes them "during the doing." It is after the automatic system is functioning and the supervisor is monitoring intermittently that those human errors make a difference and where it is therefore critical that the human supervisor intervene in time and take appropriate action when something goes wrong. Thus, the intervention stage is where human error is most manifest.

If human error is not caught by the supervisor, it is perpetuated slavishly by the computer, much as happened to the Sorcerer's Apprentice. For this reason supervisory control may be said to be especially sensitive to human error. Several factors affect the supervisor's decision to intervene and/or his or her success in doing so.

1. *Trade-Off between Collecting More Data and Taking Action in Time*. The more data collected from the more sources, the more reliable is the decision of what, if anything, is wrong and what to do about it. Weighed against this is that if the supervisor waits too long, the situation will probably get worse, and corrective action may be too late. Formally, the optimization of this decision is called the *optional stopping problem*.

2. *Risk Taking*. The supervisor may operate from either risk-averse criteria such as minimax (minimize the worst outcome that could happen) or more risk-neutral criteria such as expected value (maximize the subjectively expected gain).

Depending on the criterion, the design of a supervisory control system may be very different in complexity and cost.

3. *Mental Workload*. This problem is aggravated by supervisory control. When a supervisory control system is operating well in the automatic mode, the supervisor may have little concern. When there is a failure and sudden intervention is required, the mental workload may be considerably higher than in direct manual control, where in the latter case the operator is already participating actively in the control loop. In the former case the supervisor may have to undergo a sudden change from initial inattention, moving physically and mentally to acquire information and learn what is going on, then making a decision on how to cope. Quite likely this will be a rapid transient from very little to very high mental workload.

Although the subject of human error is currently of great interest, there is no consensus on either a taxonomy or a theory of causality of errors. One common error taxonomy relates to locus of behavior: sensory, memory, decision, or motor. Another useful distinction is between errors of omission and those of commission. A third is between slips (correct intentions that inadvertently are not executed) and mistakes (intentions that are executed but that lead to failure).

In supervisory control there are several problems of human error worth particular mention. One is the type of slip called capture. This occurs when the intended task requires a deviation from a well-rehearsed (behaviorally) and well-programmed (in the computer) procedure. Somehow habit, augmented by other cues from the computer, seems to capture behavior and drive it on to the next (unintended) step in the well-rehearsed and computer-reinforced routine.

A second supervisory error, important in both planning and failure diagnosis, results from the human tendency to seek confirmatory evidence for a single hypothesis currently being entertained (Gaines, 1976). It would be better if the supervisor could keep in mind a number of alternative hypotheses and let both positive and negative evidence contribute symmetrically in accordance with the theory of Bayesian updating (Sheridan and Ferrell (1974). Norman (1981), Reason and Mycielska (1982), Rasmussen (1982), and Rouse and Rouse (1983) provide reviews of human error research from their different perspectives.

Theoretically, anything that can be specified in an algorithm can be given over to the computer. However, the reason the human supervisor is present is to add novelty and creativity: precisely those ingredients that cannot be prespecified. This means, in effect, that the best or most correct human behavior cannot be prespecified and that variation from precise procedure must not always be viewed as errant noise. The human supervisor, by the nature of his or her function, must be allowed room by the system design for what may be called trial and error (Sheridan, 1983).

What training should the human supervisory controller receive to do a good job at detecting failures and intervening to avoid errors? As the supervisor's task becomes more cognitive, is the answer to provide training in theory and general principles? Curiously, the literature seems to provide a negative answer (Duncan, 1981). In fact, Moray (1986), in his review, concludes that there seems to be no case in the literature where training in the theory underlying a complex system has produced a dramatic change in fault detection or diagnosis. Rouse (1985) similarly concludes "that the evidence [e.g., Morris and Rouse (1985)] does not support a conclusion...that diagnosis of the unfamiliar requires theory and understanding of system principles." Apparently, frequent hands-on experience in a simulator (i.e., with simulated failures) is the best way to enable a supervisor to retain an accurate mental model of a process.

A final issue to be mentioned in conjunction with human reliability is that of trust. It comes in two forms: overtrust, also called automation bias, and under trust. Ferris et al. (2010) provide illustrations of both. An example of overtrust occurred in the fatal 1995 accident over Cali, Columbia, where pilots got confused over waypoint indications and were far off course but neverthess trusted the FMS to take care of them as they flew into a mountain. An example of untertrust was a survey of fighter pilots who opinied that UAVs could never replace human piloted aircraft in various search and other missions.

## 10 MODELING SUPERVISORY CONTROL

For 35 years various models of supervisory control have been proposed. Most of these have been models of particular aspects of supervisory control, not apparently claiming to model all or even very many aspects of it. The simplest model of supervisory control might be that of nested control loops (Figure 13), where one or more
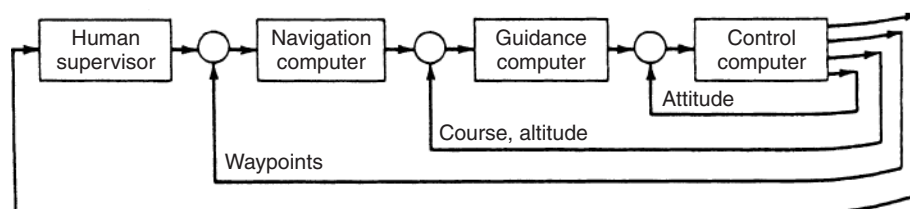


**Figure 13** Nested control loops of aerospace vehicle.

inner loops are automatic and the outer one is manual. In aerospace vehicles the innermost of four nested loops is typically called "control," the next "guidance," and the next "navigation," each having a set point determined by the next outer loop. Hess and McNally (1997) have shown how conventional manual control models can be extended to such multiloop situations. The outer loop in this generic aerospace vehicle includes the human operator, who, given mission goals, programs in the destination. In driving a car the functions of navigation, guidance, and control are all done by a person and can be seen to correspond roughly to knowledge-based, skill-based, and rule-based behavior.

Figure 14 is a qualitative functional model of supervisory control, showing the various cause–effect loops or relationships among elements of the system and emphasizing the symmetry of the system as viewed from top and bottom (human, task) of the hierarchy.

Figure 15 extends Rasmussen's model of skill-based, rule-based, and knowledge-based behavior to show various interactions with computer aids having comparable levels of intelligence.

One problem the supervisor faces is allocating attention between different tasks, where each time that he or she switches tasks there is a time penalty in transfer, typically different for different tasks and possibly



1. Task is observed directly by human operator's own senses.

2. Task is observed indirectly through artificial sensors, computers, and displays. This TIS feedback interacts with that from within HIS and is filtered or modified.

3. Task is controlled within TIS automatic mode.

4. Task is affected by the process of being sensed.

5. Task affects actuators and in turn is affected.

6. Human operator directly affects task by manipulation.

7. Human operator affects task indirectly through a controls interface, HIS/TIS computers, and actuators. This controls interaction with that from within TIS and is filtered or modified.

8. Human operator gets feedback from within HIS, in editing a program, running a planning model, etc.

9. Human operator orients him or herself relative to control or adjusts control parameters.

10. Human operator orients him or herself relative to display or adjusts display parameters.

**Figure 14**　Multiloop model of supervisory control. (From Sheridan, 1984.)
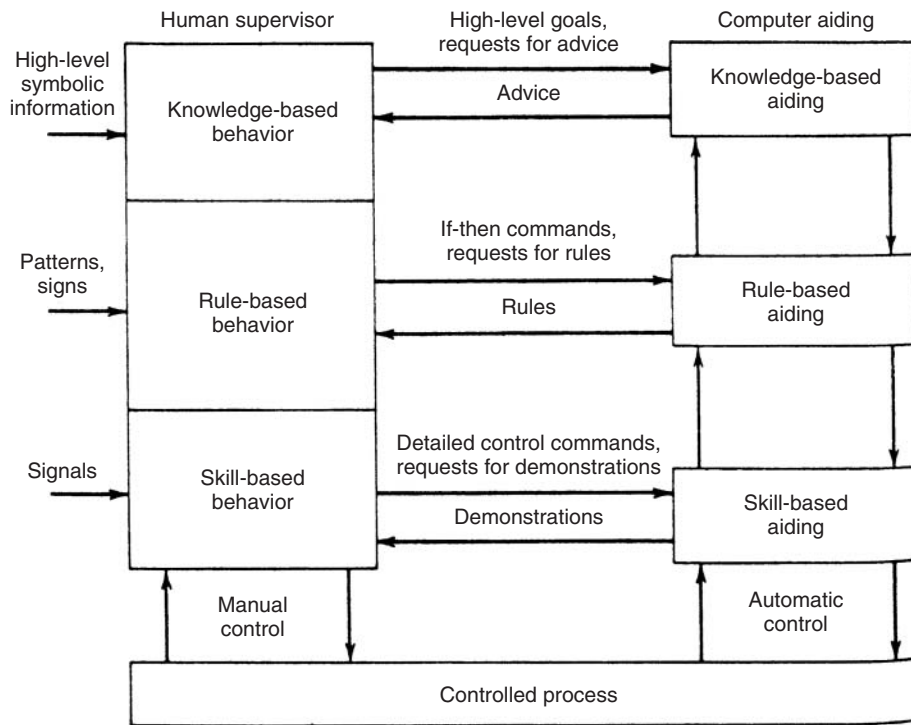
**Figure 15** Supervisor interactions with computer decision aids at knowledge, rule, and skill levels.

involving uses of different software procedures, different equipment, and even bodily transportation of himself or herself to different locations. Given relative worths for time spent attending to various tasks, it has been shown (Sheridan, 1970) that dynamic programming enables the optimal allocation strategy to be established. Moray et al. (1982) applied this model to deciding whether human or computer should control various variables at each succeeding moment. For simpler experimental conditions, the model fit the experimental data (subjects acted like utility maximizers), but as task conditions became complex, apparently it did not. Wood and Sheridan (1982) did a similar study where supervisors could select among alternative automatic machines (differing in both rental cost and productivity) to do assigned tasks or do the tasks themselves. Results showed the supervisors to be suboptimal, paying too much attention to costs and too little to productivity, and in some cases using the automation when they could have done the tasks more efficiently manually. Govindaraj and Rouse (1981) modeled the supervisor's decisions to divert attention from a continuous task to perform or monitor a discrete task.

Rouse (1977) utilized a queueing theory approach to model whether from moment to moment a task should be assigned to a computer or to the operator. The allocation criterion was to minimize service time under cost constraints. Results suggested that human–computer "misunderstanding" of one another degraded efficiency more than limited computer speed. In a related flight simulation study, Chu and Rouse (1979) had a computer

perform those tasks that had waited in the queue beyond a certain time. Chu et al. (1980) extended this idea to have the computer learn the pilot's priorities and later make suggestions when the pilot was under stress.

Tulga and Sheridan (1980) and later Pattipatti et al. (1983) utilized a model of allocation of attention among multiple task demands, a task displayed on the computer screen to the subject as is represented in Figure 16. Instead of being stationary, these demands appear at random times (not being known until they appear), exist for given periods of time, then disappear at the end of that time with no more opportunity to gain anything by attending to them, While available, they take differing amounts of time to complete and have differing rewards for completion, which information may be available after they appear and before they are "worked on." The human decision maker in this task need not allocate attention in the same temporal order in which the task demands become known, nor in the same order in which their deadline will occur. Instead, he or she may attend first to that task which has the highest payoff or takes the least time and/or may plan ahead a few moves so as to maximize gains. The Tulga–Sheridan experimental results suggest that subjects approach optimal behavior, which, when heavily loaded (i.e., there are more opportunities than he or she can possibly cope with), simply amounts to selecting the task with highest payoff regardless of time to deadline. These subjects also reported that their sense of subjective workload was greatest when by arduous planning they could barely keep up with all
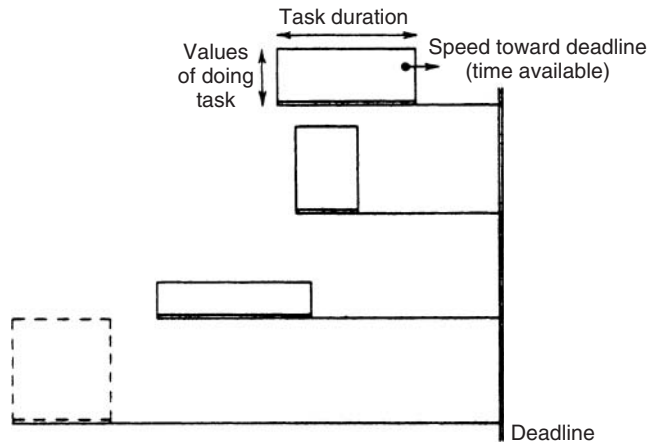
**Figure 16** Multitask computer display used in the Tulga–Sheridan experiment.

tasks presented. When still more tasks came at them and they had to select which they could do and which they had to off-load, subjective workload decreased.

Researchers and designers of supervisory control systems must cope with a number of questions. Among these are (1) how much autonomy is appropriate for the TIC, (2) how much the TIC and the HIC should tell the human supervisor, and (3) how responsibilities should be allocated among the TIC, HIC, and supervisor (Johannsen, 1981).

The famous Yogi Berra allegedly counseled: "Never make predictions, especially about the future!" Nevertheless, it is ethically mandatory that we predict as best we can. However, recent decades have seen a shift away from monolithic, computationally predictive models toward frameworks or categorizations of models, each of which may be quite simple—involving elementary control laws, a few heuristics, or pattern recognition rules. Thus, as knowledge and understanding of supervisory control have grown, along with its complexity, researchers have come to realize that they need not and cannot be held to comprehensive predictive models, desirable as they may be.

The most difficult, and it might even be said impossible, aspect of supervisory control to model is that of setting in goals, conditions, and values. Even though overall goals may be given to an actual system (or given in an experiment), how those are translated into subgoals and conditional statements remains elusive. The same is true for communicating values (criteria, coefficients of utility, etc.). Although this act of evaluation remains the sine qua non of why human participation in system control must remain, there is little prospect for mathematical modeling of this aspect in the near future.

## 11 POLICY ALTERNATIVES FOR HUMAN SUPERVISORY CONTROL

This section confronts the question of what policies might be adopted in dealing with the human–automation

interaction challenges that are unavoidable in the systems discussed here. Dilemmas will surely arise with regard to (a) when not to follow the recommendation of a decision support tool or when to bypass automation when either is believed to have failed or not be appropriate to the current situation and (b) how long to wait for automation to act before intervening manually.

The public will still demand both safety and efficiency and may continue to place stringent expectations on the human to provide both unless and until automation can prove itself sufficiently robust and reliable. New technology permits closer surveillance of human behavior. These facts could even exacerbate the pressures to maintain the "blame game" of punishing what may be seen as errant behavior, even though it is fully recognized that all human beings are inclined to err from time to time (Kohn et al., 2000). Institutional cultures change slowly, even given that large human–machine system developments such as NextGen have explicitly embodied efforts to work toward a "just culture" (Dekker, 2007) in dealing with human frailties. That puts a new emphasis on learning from mistakes rather than on meting out punishment.

It seems that five alternative policy approaches with regard to human operator roles and responsibilities can be distinguished (Sheridan, 2010). These are offered as contrasting approaches. Most likely some amalgam of these will be adopted by management in different system contexts.

1. *Maintain the typical status quo of full human operator responsibility.* Operators would undergo extensive training in decision support tools and automation so they could understand their use and limitations. Accordingly, they would be expected to use them wisely and continue to be responsible for safety and operations. The advantage of this approach

would be to minimize the need for policy and training changes in an evolving system. The disadvantage would be in leaving open dilemmas faced by the human operators as to what to do when automation seems to be inadequate to handle a situation or the controller is unsure of whether the automation will act before it is too late.

2. *Define explicit behavior thresholds and criteria to determine when controllers would be held responsible*. For example, in NextGen the automation will assume certain control functions previously performed by ground controllers communicating with pilots and vectoring aircraft, so controllers might be instructed not to bypass or override the automation unless and until certain explicit criteria (with respect to time, distance, etc.) are met. The ground controllers would be trained accordingly. The advantage of this approach would be that the controller would have very specific rules as to his or her responsibility. The disadvantage would be that defining such rules might be difficult to agree on and in any case seem to limit the controller's discretion. Further, the more detailed the rules are that the controller is asked to commit to memory, the more likely that some details will be forgotten or confused.

3. *Define and emphasize in both training and operation the ideal behavior and rationale to be used for each operation*. The advantage would be that with operators understanding and appreciating the basic operational concepts they could make best use of their professional expertise and experience. The disadvantage would be that there may not be uniformity in their response to events, particularly off-nominal events.

4. *Expect operators to "always do their best" in deciding when and how to employ automation or to bypass or override*. The (refined) record-keeping would determine whether they would be exonerated in any mishap. The advantage of this approach would be that the operator would be somewhat protected if the evidence showed he or she was really trying but the constraints of the situation were just too much to handle. The disadvantage would be that operators might be motivated toward laxity, hoping in any case to claim they were doing their best based on the evidence.

5. *Expect operators to "always do their best" but allow them to signal in real time when they feel they must intervene in an automated process*. Encourage them to announce when they have a decision dilemma or regard the situation as untenable. The advantage would be to add evidence to the record of what happened. The disadvantage would be the same as that under 4 above.

## 12 SOCIAL IMPLICATIONS AND THE FUTURE OF HUMAN SUPERVISORY CONTROL

One near certainty is that, as technology of computers, sensors, and displays improves, supervisory control will become more prevalent. This should occur in two ways: (1) a greater number of semiautomated tasks will be controlled by a single supervisor (a greater number of TICs will be connected to a single HIC) and (2) the sophistication of cognitive aids, including expert systems for planning, teaching, monitoring, failure detection, and learning, will increase and include more of what we now call knowledge-based behavior in the HIC.

The World Wide Web has enabled easy worldwide communication (for those properly equipped). One aspect of that communication that up to now has hardly become manifest is the ability to exercise remote control. A number of experimental demonstrations have been performed on controlling robots between continents, and in military operations UAVs are being controlled this way, but delayed feedback still poses a difficulty for continuous control, so supervisory control clearly has an advantage here. In the future we should see many more applications of moderate- and long-distance remote control.

Concurrently, understanding by the layperson (including those of both corporate and government bureaucracies) should come to understand the potential of supervisory control much better. At the present time the layperson tends to see automation as "all or none," where a system is controlled either manually or automatically, with nothing in between. In robotized factories the media tend to focus on the robots, with little mention of design, installation, programming, monitoring, fault detection and diagnosis, maintenance, and various learning functions that are performed by people. In the space program the same is true; options are seen to be either "automated," "astronaut in extravehicular activity (EVA)," or "astronaut or ground controlling telemanipulator" without much appreciation for the potential of supervisory control.

In considering the future of supervisory control relative to various degrees of automation and to the complexity or unpredictability of task situations to be dealt with, a representation such as Figure 17 comes to mind. The meaning of the four extremes of this rectangle are quite identifiable. Supervisory control may be considered to be a frontier (line) advancing gradually toward the upper right-hand corner.

For obvious reasons, the tendency has been to automate what is easiest and to leave the rest to the human. This has sometimes been called the technological imperative. From one perspective this dignifies the human contribution; from another it may lead to a hodge-podge of partial automation, making the remaining human tasks less coherent and more complex than need be, resulting in overall degradation of system performance (Bainbridge, 1983; Parsons, 1985)

"Human-centered automation" has become a popular phrase (Billings, 1991) and is often used in relation to human supervisory control. Therefore, to end this chapter, we might consider its alternative meanings. Below are 10 alternative meanings (stated in italics) that
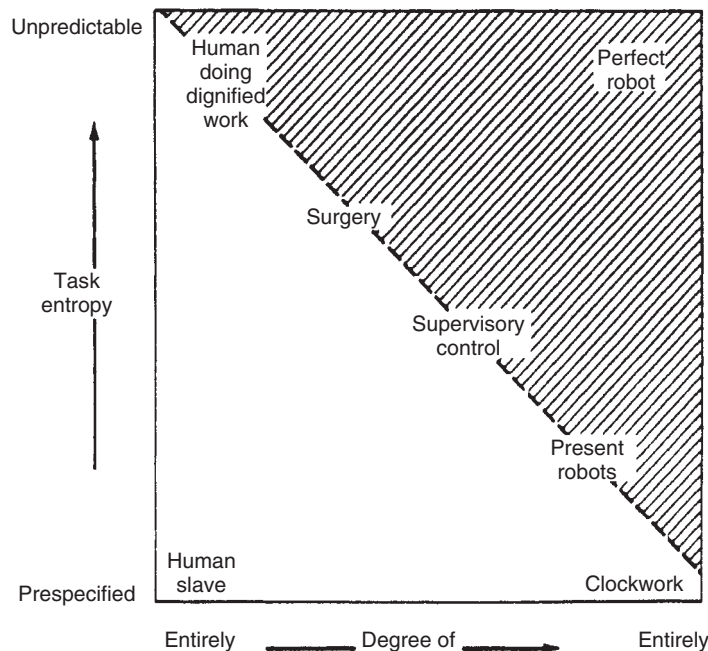
**Figure 17** Combinations of human and computer control to achieve tasks at various levels of difficulty.

the author has gleaned from current literature. In every case the meaning must be qualified, as is done by the one or two sentences following each particular meaning of the phrase.

1.  *Allocate to the human the tasks best suited to the human, allocate to the automation the tasks best suited to it.* Yes, but for some tasks it really is easier to do them manually than to initialize the automation to do them. And at the other end of the spectrum are tasks that require so much skill or art or creativity that it simply is not possible to program a computer to do them.

2.  *Keep the human operator in the decision and control loop.* That is a good idea provided that the control tasks are of appropriate bandwidth, attentional demand, and so on.

3.  *Maintain the human operator as the final authority over the automation.* Realistically, this is not always the safest solution. It depends on the task context. In nuclear plants, for example, there are safety functions that cannot be entrusted to the human operator and cannot be overridden by him or her. Examples have been given previously in the case of aircraft automation.

4.  *Make the human operator's job easier, more enjoyable, or more satisfying through friendly automation.* That is fine if operator ease and enjoyment are the primary considerations and if ease and enjoyment necessarily correlate with operator responsibility and system performance, but often these conditions are not the case.

5.  *Empower the human operator to the greatest extent possible through automation.* Again one must remember that operator empowerment is not the same as system performance. Maybe the designer knows best. Don't encourage megalomaniacal operators.

6.  *Support trust by the human operator.* Trust of the automation by the operator is often a good thing, but not always. Too much trust is just as bad as not enough trust.

7.  *Give the operator computer-based information about everything that he or she should want to know.* We now have many examples of where too much information can overwhelm the operator to the point where performance breaks down and even when the operator originally wanted "all" the information.

8.  *Engineer the automation to reduce human error and keep response variability to the minimum.* This, unfortunately, is a simplistic view of human error. Taken literally it reduces the operator to an automaton, a robot. Modest levels of error and response variability enhance learning (Darwin's requisite variety).

9.  *Make the operator a supervisor of subordinate automatic control system(s).* Although this is a chapter on supervisory control, it must be noted that for some tasks direct manual control may be best.

10. *Achieve the best combination of human and automatic control, where best defined by explicit*

*system objectives.* Again, in some ideal case, where objectives can reliably be reduced to mathematics, this would be just fine. Unfortunately, automatic judgment of what is good and bad in a particular situation is seldom possible, even for a machine programmed with the best available algorithms or heuristics. Fortunately, judgment of what is good and bad in a particular situation is almost the essence of what it is to be human.

The bottom line is that proper use of automation depends upon context, which in turn depends upon designer and operator judgment.

I have written elsewhere about the long-term social implications of supervisory control (Sheridan, 1980; Sheridan et al., 1983). My concerns are reviewed here very briefly:

1. *Unemployment.* This is the factor most often considered. More supervisory control means more efficiency, less direct control, and fewer jobs.

2. *Desocialization.* Although cockpits and control rooms now require two- to three-person teams, the trend is toward fewer people per team, and eventually one person will be adequate in most installations. Thus, cognitive interaction with computers will replace that with other people. As supervisory control systems are interconnected, the computer will mediate more and more interpersonal contact.

3. *Remoteness from the Product.* Supervisory control removes people from hands-on interaction with the workpiece or other product. They become not only separated in space but also desynchronized in time. Their functions or actions no longer correspond to how the product itself is being handled or processed mechanically.

4. *Deskilling.* Skilled workers "promoted" to supervisory controller may resent the transition because of fear that when and if called on to take over and do the job manually they may not be able to. They also feel loss of professional identity built up over an entire working life.

5. *Intimidation by Higher Stakes.* Supervisory control will encourage larger aggregations of equipment, higher speeds, greater complexity, higher costs of capital, and probably greater economic risk if something goes wrong and the supervisor does not take the appropriate action.

6. *Discomfort in the Assumption of Power.* The human supervisor will be forced to assume more and more ultimate responsibilities. Depending on one's personality, this could lead to insensitivity to detail, anxiety about being up to the job requirements, or arrogance.

7. *Technological Illiteracy.* Supervisory controllers may lack the technological understanding of how the computer does what it does. They may come to resent this and resent the elite class who do understand.

8. *Mystification.* Human supervisors of computer-based systems could become mystified about the power of the computer, even seeing it as a kind of magic or "big brother" authority figure.

9. *Sense of Not Being Productive.* Although the efficiency and mechanical productivity of a new supervisory control system may far exceed that of an earlier manually controlled system that a given person has experienced, that person may come to feel no longer productive as a human being.

10. *Eventual Abandonment of Responsibility.* As a result of the factors described previously, supervisors may eventually feel that they are no longer responsible for what happens; the computers are.

These 10 potential negatives may be summarized with a single word: alienation. In short, if human supervisors of the new breed of computer-based systems are not given sufficient familiarization with and feedback from the task, sufficient sense of retaining their old skills, or ways of finding identity in new ones, they may well come to feel alienated. They must be trained to feel comfortable with their new responsibility, must come to understand what the computer does and not be mystified, and must realize that they are ultimately in charge of setting the goals and criteria by which the system operates. If these principles of human factors are incorporated into the design, selection, training, and management, supervisory control has a positive future.

## 13 CONCLUSIONS

Computer technology, both hard and soft, is driving the human operator to become a supervisor (planner, teacher, monitor, and learner) of automation and an intervener within the automated control loop for abnormal situations. A number of definitions, models, and problems have been discussed. There is little or no present consensus that any one of these models characterizes in a satisfactory way all or even very much of supervisory control with sufficient predictive capability to entrust to the designer of such systems. It seems that for the immediate future we are destined to run breathless behind the lead of technology, trying our best to catch up.

## REFERENCES

Abbott, K., Slotte, S., Stimson, D., Amalberti, R. R., Bollin, G., Fabre, F., Hecht, S., Imrich, T., Lalley, R., Lydanne, G., Newman, T., and Thiel, G. (1996), *The Interfaces between Flightcrews and Modern Flight Deck Systems*, report of the FAA Human Factors Team, Federal Aviation Administration, Washington, DC.

Bainbridge, L. (1983), "Ironies of Automation," *Automatica*, Vol. 19, pp. 775–779.

Billings, C. S. (1991), *Human-Centered Aircraft Automation: A Concept and Guideline*, NASA TM-103885, NASA Ames Research Center, Moffett Field, CA.

Boehm-Davis, D., Curry, R., Wiener, E., and Harrison, R. (1983), "Human Factors of Flight Deck Automation: Report on a NASA-Industry Workshop," *Ergonomics*, Vol. 26, pp. 953–961.

Boer, E. R. (2000), "Behavioral Entropy as an Index of Workload,"in *Proceedings of the 44th Annual Meeting of the Human Factors and Ergonomics Society*, San Diego, CA, July 30-August 4.

Brooks, T. L. (1979), "SUPERMAN: A system for Supervisory Manipulation and the Study of Human-Computer Interactions,"S.M. Thesis, Massachusetts Institute of Technology, Cambridge, MA.

Chu, Y., and Rouse, W. B. (1979), "Adaptive Allocation of Decision Making Responsibility between Human and Computer in Multi-task Situations," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 9, pp. 769–778.

Chu, Y. Y., Steeb, R., and Freedy, A. (1980), *Analysis and Modeling of Information Handling Tasks in Supervisory Control of Advanced Aircraft*, PATR-1080-80-6, Perceptronics, Woodlands, CA.

Cummings, M. L. (2005), "Human Supervisory Control Challenges in Network Centric Operations,"in *Proceedings of the Unmanned Vehicle Systems Canada Conference*, Banff, CA.

Curry, R. E., and Ephrath, A. R. (1977), "Monitoring and Control of Unreliable Systems," in *Monitoring Behavior and Supervisory Control*, T. B. Sheridan and G. Johannsen, Eds., Plenum, New York, pp. 193–203.

Degani, A. (2004), *Taming HAL: Designing Interfaces Beyond 2001*, Palgrave, New York.

Dekker, S. (2007), *Just Culture: Balancing Safety and Accountability*, Ashgate, Surrey, United Kingdom.

Duncan, K. D. (1981), "Training for Fault Diagnosis in Industrial Process Plants," in *Human Detection and Diagnosis of System Failures*, J. Rasmussen and W. B. Rouse, Eds., Plenum, New York, pp. 553–524.

Edwards, E., and Lees, F. (1974), *The Human Operator in Process Control*, Taylor & Francis, London.

Ephrath, A. R., and Young, L. R. (1981), "Monitoring vs. Man-in-the-Loop Detection of Aircraft Control Failures," in *Human Detection and Diagnosis of System Failures*, J. Rasmussen and W. B. Rouse, Eds., Plenum, New York, pp. 143–154.

Falzon, P. (1982), "Display Structures: Computability with the Operator's Mental Representation and Reasoning Processes,"in *Proceedings of the 2nd Annual Conference on Human Decision Making and Manual Control*, pp. 297–305.

Ferrell, W. R. (1966, October), "Delayed Force Feedback," *Human Factors*, October, pp. 449–455.

Ferrell, W. R., and Sheridan, T. B. (1967), "Supervisory Control of Remote Manipulation," *IEEE Spectrum*, Vol. 4, No. 10, pp. 81–88.

Ferris, T., Sarter, N., and Wickens, C. D. (2010), "Cockput Automation: Still Struggling to Catch Up," in *Human Factors in Aviation*, E. Salas and D. Maurino, Eds., Academic, San Diego, CA, pp. 479–504.

Gai, E. G., and Curry R. E. (1978), "Preservation Effects in Detection Tasks with Correlated Decision Intervals," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 8, pp. 93–110.

Gaines, B. R. (1976), "On the Complexity of Casual Models," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 6, pp. 56–59.

Gentner, D., and Stevens, A. L., Eds. (1983), *Mental Models*, Lawrence Erlbaum Associates, Mahwah, NJ.

Govindaraj, T., and Rouse, W. B. (1981), "Modeling the Human Controller in Environments That Include Continuous and Discrete Tasks," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 11, pp. 411–417.

Hess, R. A., and McNally, B. D. (1997), "Automation Effect in a Multi-Loop Manual Control System," *IEEE Transactions on Systems, Man Cybernetics*, Vol. SMC-16, No. 1, pp. 111–121.

Johannsen, G. (1981), "Fault Management and Supervisory Control of Decentralized Systems," in *Human Detection and Diagnosis of System Failures*, J. Rasmussen and W. B. Rouse, Eds., Plenum, New York.

Kohn, L. T., Corrigan, J. M., and Donaldson, M. S., Eds. (2000), *To Err Is human: Building a Safer Health System*, National Academies Press, Washington, DC.

Lee, J. D., McGehee, D., Brown, T. L., and Reyes, M. (2002), "Collision Warning Timing, Driver Distraction, and Driver Response to Imminent Rear End Collision in a High Fidelity Driving Simulator," *Human Factors*, Vol. 44, No. 2, pp. 314–334.

Llaneras, R. E. (2000), "NHTSA Driver Distraction Internet Forum," available: www.nrd.nhtsa.dot.gov/departments/nrd-13/DriverDistraction.html.

Machida, K., Toda, Y., Iwata, T., Kawachi, M., and Nakamura, T. (1988), "Development of a Graphic Simulator Augmented Teleoperator System for Space Applications,"in *Proceedings of the 1988 AIAA Conference on Guidance, Navigation, and Control*, Part I, pp. 358–364.

Moray, N. (1982), "Subjective Mental Workload," *Human Factors*, Vol. 24, pp. 25–40.

Moray, N. (1986), "Monitoring Behavior and Supervisory Control," in *Handbook of Perception and Human Performance*, Vol. 2, K. Boff, L. Kaufman, and J. P. Thomas, Eds., Wiley, New York.

Moray, N. (1997), "Models of Models of . . . Mental Models," in *Perspectives on the Human Controller*, T. Sheridan and T. van Lunteren, Eds., Lawrence Erlbaum Associates, Mahwah, NJ.

Morris, N. M., and Rouse, W. B. (1985), "The Effects of Type of Knowledge upon Human Problem Solving in a Process Control Task," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 15, pp. 698–707.

Nandi, H., and Ruhe, W. (2002), "On Line Modeling and New Generation of Supervisory Control System for Sintering Furnaces," Comp US Controls, Indiana, PA.

Norman, D. A. (1981), "Categorization of Action Slips," *Psychological Review*, Vol. 88, pp. 1–15.

Norman, D. A. (1986), "Cognitive Engineering," in *User Centered System Design: New Perspectives in Human-Computer Interaction*, D. A. Norman and S. Draper, Eds., Lawrence Erlbaum Associates, Hillsdale, NJ.

Noyes, M. V., and Sheridan, T. B. (1984), "A Novel Predictor for Telemanipulation through a Time Delay," in *Proceedings of the Annual Conference on Manual Control*, NASA Ames Research Center, Moffett Field, CA, Human Factors and Ergonoomics Society, Santa Monica, CA.

Parasuraman, R. (1987), "Human-Computer Monitoring," *Human Factors*, Vol. 29, pp. 695–706.

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000), "A Model for Types and Levels of Human Interaction with Automation," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 30, No. 3, pp. 286–297.

Park, J. H. (1991, August), "Supervisory Control of Robot Manipulators for Gross Motions," Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Parsons, H. M. (1985), "Automation and the Individual: Comprehensive and Comparative Views," *Human Factors*, Vol. 27, No. 1, pp. 99–111.

Pattipatti, K. R., Kleinman, D. L., and Ephrath, A. R. (1983), "A Dynamic Decision Model of Human Task Selection Performance," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 13, pp. 145–166.

Rasmussen, J. (1982), "Human Errors: A Taxonomy for Describing Human Malfunction in Industrial Installations," *Journal of Occupational Accidents*, Vol. 4, pp. 311–335.

Rassmussen, J., and Rouse, W. B., Eds. (1981), *Human Detection and Diagnosis of System Failures*, Plenum, New York.

Reason, J. T., and Mycielska, K. (1982), *Absent Minded? The Psychology of Mental Lapses and Everyday Errors*, Prentice-Hall, Englewood Cliffs, NJ.

Rouse, W. B. (1977), "Human-Computer Interaction in Multi-Task Situations," IEEE *Transactions on System, Man and Cybernetics*, Vol. 7, No. 5, pp. 384–392.

Rouse, W. B. (1985), "Supervisory Control and Display Systems," in *Human Productivity Enhancement*, J. Zeidner, Ed., Praeger, New York.

Rouse, W. B., and Morris, N. M. (1984), *On Looking into the Black Box: Prospects and Limits in the Search for Mental Models*, Search Technology, Norcross, GA.

Rouse, W. B., and Rouse, S. H. (1983), "Analysis and Classification of Human Error," IEEE *Transactions on System, Man and Cybernetics*, Vol. 13, No. 4, pp. 539–599.

Sarter, N. B., and Amalberti, R., Eds. (2000), *Cognitive Engineering in the Aviation Domain*, Lawrence Erlbaum Associates, Mahwah, NJ.

Sarter, N. B., and Woods, D. D. (2000), "Learning from Automation: Surprises and 'Going Sour' Accidents," in *Cognitive Engineering in the Aviation Domain*, N. B. Sarter and R. Amalberti, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, p. 327.

Sarter, N. B., Mumaw, R. J., and Wickens, C. D. (2007), "Pilots' Monitoring Strategies and Performance on Automated Flight Decks: An Empirical Study Combining Behavioral and Eye-Tracking Data," *Human Factors*, Vol. 49, pp. 347–357.

Seiji, T., Ohga, Y., and Koyama, M. (2001), "Advanced Supervisory Control Systems for Nuclear Power Plants," *Hitachi Review*, Vol. 50, No. 3.

Sheridan, T. (2010), *Controller-Automation Issues in the Next Generation Air Transportation System (NextGen)*, DOT-VNTSC-FAA-10-11, Volpe Center, Cambridge, MA.

Sheridan, T., and Parasuraman, R. (2006), "Human-Automation Interaction," *Reviews of Human Factors and Ergonomics*, Vol. 1, pp. 89–129.

Sheridan, T. B. (1960), "Human Metacontrol," in *Proceedings of the Annual Conference on Manual Control*, Wright-Patterson Air Force Base, OH, Institute of Electrical and Electronics Engineers, New York.

Sheridan, T. B. (1970), "Optimum Allocation of Personal Presence," *IEEE Transactions on Human Factors in Electronics*, Vol. 10, pp. 242–249.

Sheridan, T. B. (1976), "Toward a General Model of Supervisory Control," in *Monitoring Behavior and Supervisory Control*, T. B. Sheridan and G. Johannsen, Eds., Plenum, New York.

Sheridan, T. B. (1980, October), "Computer Control and Human Alienation," *Technology Review*, Vol. 83, pp. 60–73.

Sheridan, T. B. (1983), "Measuring, Modeling and Augmenting Reliability of Man-Machine Systems," *Automatica*, Vol. 19, pp. 637–645.

Sheridan, T. B. (1984), *Interaction of Human Cognitive Models and Computer-Based Models in Supervisory Control, Man-Machine Systems Laboratory*, Massachusetts Institute of Technology, Cambridge, MA.

Sheridan, T. B. (1992), *Telerobotics, Automation, and Human Supervisory Control*, Massachusetts Institute of Technology, Cambridge, MA.

Sheridan, T. B. (1997, October), "Function Allocation: Algorithm, Alchemy or Apostasy?" paper presented at the Conference on Function Allocation, ALLFN 97, Galway, Ireland

Sheridan, T. B. (2002), *Humans and Automation*, Wiley, Hoboken, NJ.

Sheridan, T. B., and Ferrell, W. R. (1974), *Man-Machine Systems*, Massachusetts Institute of Technology, Cambridge, MA.

Sheridan, T. B., and Hennessy, R. T., Eds. (1984), *Research and Modeling of Supervisory Control Behavior*, National Research Council, Committee on Human Factors, National Academy Press, Washington, DC.

Sheridan, T. B., and Johannsen, G., Eds. (1976), *Monitoring Behavior and Supervisory Control*, Plenum, New York.

Sheridan, T. B., and Verplank, W. L. (1978), *Human and Computer Control of Undersea Teleoperators*, Man-Machine Systems Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Tulga, M. K., and Sheridan, T. B. (1980), "Dynamic Decisions and Workload in Multi-Task Supervisory Control," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 10, No. 5, pp. 217–231.

Wickens, C. D., and Kessell, C. (1979), "The Effects of Participatory Model and Task Workload on the Detection of Dynamic System Failures," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 9, pp. 24–34.

Wickens, C. D., and Kessell, C. (1981), "Failure Detection in Dynamic Systems," in *Human Detection and Diagnosis of System Failures*, J. Rasmussen and W. B. Rouse, Eds., Plenum, New York.

Wickens, C., Mavor, A., and McGee, J. (1997), *Flight to the Future*, National Academy Press, Washington, DC.

Wiener, E. L. (1988), "Cockpit Automation," in *Human Factors in Aviation*, E. L. Wiener and D. Nagel, Eds., Academic, San Antonio, TX, pp. 433–461.

Wiener, E. L., and Curry, R. E. (1980), "Flight Deck Automatic: Promises and Problems," *Ergonomics*, Vol. 23, pp. 995–1011.

Wood, W., and Sheridan, T. B. (1982), "The Use of Machine Aids in Dynamic Multi-Task Environments: A Comparison of an Optimal Model to Human Behavior," in *Proceedings of the IEEE International Conference on Cybernetics and Society*, Seattle, WA, pp. 668–672, Institute of Electrical and Electronics Engineers, New York.

Yoerger, D. (1982), "Supervisory Control of Underwater Telemanipulators: Design and Experiment," Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

# CHAPTER 35

# HUMAN DIGITAL MODELING IN DESIGN

**Vincent G. Duffy**
**Purdue University**
**West Lafayette, Indiana**

## 1   DEFINITIONS AND FOUNDATIONS OF HUMAN DIGITAL MODELING

Human digital modeling can be considered a digital representation of the human inserted into a simulation or virtual environment to facilitate prediction of safety and/or performance (Duffy, 2009a; Demirel and Duffy, 2007a). These include some visualization as well as the math and/or science in the background (Duffy, 2009a; Demirel and Duffy, 2007b). Applications in this field demonstrate how to reduce the need for prototyping and incorporate ergonomics and human factors earlier in the design process (Duffy, 2010b; Applied Human Factors and Ergonomics International, 2012). Recent new model development and applications include aviation models, manufacturing and service industries, virtual ergonomic assessment, anthropometrics, automotive design, human shape design, Bayesian modeling, human behavior modeling, risk assessment modeling, and validation (Duffy, 2010b). These will be outlined in this chapter. In the consumer-driven marketplace, the time it takes to bring a new or modified product to market can support or hinder new product launch. Human digital modeling can improve time to market and increase safety and ultimately the profitability of the organization (Duffy, 2010b).

As was stated in the preface in a recent related book, the growing body of literature in human digital modeling makes it difficult for newcomers to identify the key elements quickly (Duffy, 2009b). This chapter is intended to summarize what is available as new developments in the field while incorporating the foundations that these recent developments were built upon. Within this emerging area, there exists opportunity for human factors and ergonomics practitioners and researchers to develop a common language for better communication with practicing engineers and product designers. At present, very few in the human factors and ergonomics community have the opportunity for formal coursework that incorporates digital human modeling. In addition, Chaffin notes that "we are graduating very few engineers (probably less than 10%) who even have a first course in human factors and ergonomics" (Chaffin, 2005) and encourages participation from the human factors and ergonomics community in this emerging area (Chaffin, 2002).

For a human factors community, the term *human digital modeling* may really make more sense than *digital human modeling*, which is why human digital modeling has really become a human factor and the title of this chapter has used the reverse order wording "human digital modeling" to draw the attention of human factors and ergonomics practitioners who may consider "digital human modeling" to be somehow not very accessible. The remainder of the chapter is intended to demonstrate that the field has carefully incorporated the foundations of human factors and ergonomics and provides a great set of tools for bringing human factors and ergonomics earlier into the design process and the potential for new ones to still be developed or incorporated into commercially available tools. However, as the field has developed under the terminology digital human modeling (DHM), that term will be used in this chapter when that referenced literature also used that DHM terminology.

For the practicing engineer, human digital modeling represents the opportunity to reduce the need for physical prototyping as it typically makes the analyses available through commercial computer-aided engineering (CAE) and product life-cycle management (PLM) software packages. It also provides opportunities for engineers to facilitate faster product development efforts and reduce time to market for new products. For the field to continue in development, there needs to be the continuing dialogue between engineers and human factors and ergonomics specialists. Where the engineers also have human factors and ergonomics expertise or where the human factors and ergonomics specialists are engineers, there is the opportunity for faster adoption of

these methods. For the methods previously developed in human digital modeling or DHM to have greater adoption, it will take the readers of this chapter to take it upon themselves to do so.

Since a very small percentage of engineers have been trained in human factors and ergonomics, and only a small percentage of human factors and ergonomics specialists have the opportunity to learn about DHM as a part of the course curriculum, the developers of the analysis tools are currently tasked with that effort of facilitating adoption within their client organizations. They may at times be viewed somewhat suspiciously because they typically come from organizations with commercial objectives or they may be ignored from a variation of the NIH (not-invented-here) syndrome. Actually, commercial software developers such as Siemens-Jack, CATIA-Delmia, and Ramsis have been great facilitators so far for this emerging area that has such great potential for impacting product design and consumer applications in a positive way. However, just as in medicine Dr. Jim Bagian tells about doctors who may not appreciate medical techniques that were not taught formally during their medical training (Williams and Bagian, 2010), so engineers appear to be falling into this potential pitfall by ignoring emerging capabilities of PLM packages that could support their efforts to bring consumer-friendly designs to the marketplace.

Some early DHM models that have been available for up to 35 years have not been rapidly assimilated into organizations where they could improve the ergonomic design of most of the hardware and software systems used today (Chaffin, 2009). The benefits of using related technologies have been well documented over the last decade in various books, papers, and conference presentations. In outlining the past and present in North America and Europe from a scientific perspective, Bubb and Fritzsche (2009) focuses on five different lines of development, including anthropometric models, models for production design, biomechanical models, anatomical models, and cognitive models. Sheridan (2009) extends the discussion on the cognitive side and considers a historical perspective on human performance modeling. Though this tradition developed professionally through a technical group within the Human Factors and Ergonomics Society beginning in 2004, Sheridan traces the development over 50 years focusing on models intended to predict the results of future measurements that have been popular and widely applicable.

Researchers in the United States and Sweden highlighted some organizational and technical conditions that may be inhibiting faster adoption, including lack of trained DHM personnel (Chaffin, 2009; Hanson et al., 2009). Some use cases, and some of the connecting points to the human factors and ergonomics methodologies are specifically outlined in this chapter (Brazier et al., 2003; Bubb and Fritzsche, 2009; Chaffin, 2009; Duffy, 2007b; Peacock and Karwowski, 1993). Recent advances in human digital modeling demonstrate that this emerging area is developing as an international effort with contributions recently from Belgium, Canada, China, France, Greece, Ireland, Italy, Japan, Germany, Hong Kong, Korea, Malaysia, The Netherlands, Poland, Singapore, Spain, Sweden, Taiwan, the United Kingdom, and the United States (Duffy, 2009b, 2009c, 2010a).

## 2 MODELING FUNDAMENTALS

For the purposes of understanding the current state of development in the field of human modeling, modeling fundamentals will be outlined in relation to stages of development, that is, first, second, and third generation. Within this outline, one may distinguish those different generations of models as follows. First-generation models tended to be more empirically or data driven and focused on the physical aspects of work or human–system design. Second-generation models have tended to be more computationally driven. Those first built into the commercially available CAE and PLM software would be considered first-generation models. As the need for advanced product assessment often included multiple objectives and multiple constraints, the proposed solutions developed in research were more mathematically based.

First-generation models tended to have more validation and data to support predictions. Second-generation models may have broader applicability due to the multiple objectives and constraints that they address. Challenges arise as to the validity and limits of such applications. These have also tended to be more focused on the physical aspects, capitalizing on existing models that could be considered mathematically for further development. As the sponsor of some of the early models referred to by Sheridan were the military, with concerns in aviation at the time, some second-generation models (outlined in Abdel-Malek and Arora, 2009; Marler et al., 2009; Yang, 2009) and shifting paradigms also came from the military. The Virtual Soldier Project, which drove this second-generation modeling perspective, initially focused on the physical aspects as the cognitive models that existed were not part of the commercially available tools or original scope. As the nature of the work is changing, the requirements for modeling the cognitive aspects have increased. As these become more well known and understood across disciplines outside of psychology and cognitive science, there is greater potential for their integration into the commercially available tools. Representative samples of each will be outlined in Sections 2.1 and 2.2.

### 2.1 Physical Aspects of Work

Commercially available first-generation DHM simulation and analysis tools are outlined in Table 1. These have enabled the analysis of tasks that required lifting, carrying, and lowering.

A typical computer manikin application is shown in Figure 1 for an assembly task. Figure 2 shows the field of vision and reach envelope used in modern applications of DHM.

Advanced first-generation models enable one to move beyond some limiting assumptions of the first-generation models and analysis tools and have tended to focus on verification and validation. These tend to not

**Table 1 Selected First-Generation Analysis Tools Available in Commercially Available DHM Simulation Tools**

| Performance Model | Data Source | Input Parameters |
| --- | --- | --- |
| National Institute of Occupational Safety and Health (NIOSH) lifting equation | Waters et al., 1993 | Posture and lift begin and end, object weight, hand coupling |
| Low-back injury risk assessment | Chaffin et al., 1999 | Joint torques, postures |
| Strength assessment | Ciriello and Snook, 1991 | Task description, hand coupling, gender |
| Fatigue analysis | Rohmert, 1973 | Joint torques, strength equations |
| Metabolic energy expenditure | Garg et al., 1978 | Task descriptions, gender, load description |
| Rapid upper limb assessment | McAtamney and Corlett, 1993 | Posture assessment, muscle use, force description |
| Ovako working posture | Karhu et al., 1977 | Posture assessment |
| Comfort | Variety of sources, including Dreyfuss, 1993; Porter and Gyi, 1998; Rebiffe, 1966 | Posture assessment |

*Source*: Adapted from Raschke et al. (2001).
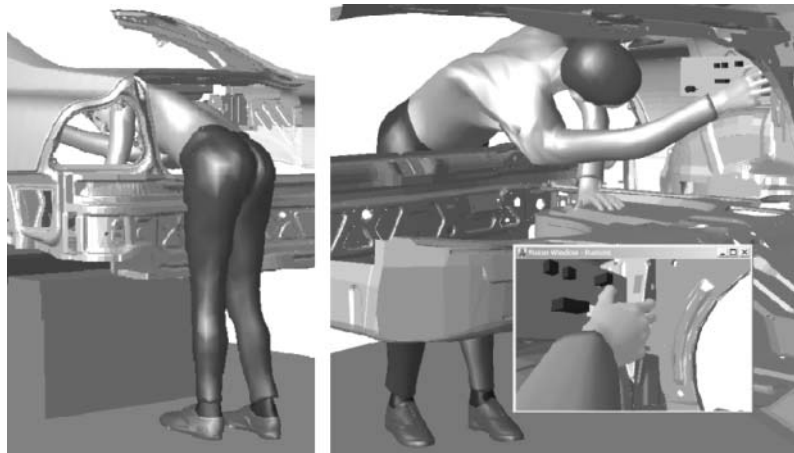


**Figure 1** Typical computer manikin application (eM-Human Advanced/RAMSIS): analysis of a preliminary assembly position at the Volvo S60 Road Traffic Information (RTI) unit. (Modified from Sundin and Ortengren, 2006. Courtesy of Volvo Car Corporation/Wiley.)

yet be available in commercial CAE or PLM packages. For instance, Marras (2006) highlights the need to consider the dynamic aspects of task, where the static posture assumption is a limitation of the revised NIOSH lift equation. It is estimated that the potential risk to the lower back from lifting may be underestimated by as much as 40% by assuming a static posture (Feyen et al., 2000). Marras outlines options by considering use of a lumbar motion monitor to estimate the potential risk due to the dynamic aspects of the task. The lumbar motion monitor (Figure 3; Marras, 2006) and motion capture systems (Cappelli and Duffy, 2007; Tian et al., 2007) allow for the capture of velocities and accelerations and angular velocities and angular accelerations that can be used to determine the likelihood of high risk for a task. Additional information on the capabilities and limitations of motion capture systems and other instrumentation in support of dynamic digital human modeling can be found in Wang (2009) and Morr

et al. (2009). Wang describes the use of motion capture for discomfort evaluation and notes the challenges in capturing a measure of discomfort, as it is not a measure that can be considered directly the opposite of comfort. In their chapter on instrumentation, Morr et al. (2009) consider digital human models and validation-related matters when a human is a passenger in a vehicle, but not as part of the product interaction. This is the case in vehicle crash simulations.

Efforts to extend the use of existing models such as comfort analysis were undertaken in various studies including use of comfort angles in vehicles when designing an ATM for people with limited mobility. Figure 4 shows that old and new designs can be compared to validate the use of previous measures in circumstances other than those for which they were intended. Some methodologies established for verification and validation in first-generation and advanced first-generation DHM model development are outlined by Oudenhuijzen and
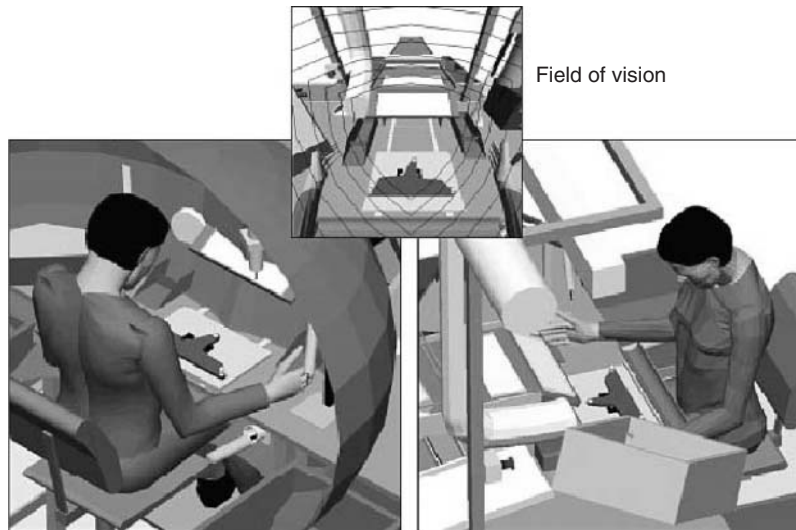
**Figure 2** Analysis functions. Top: field of vision, where lines represent the range of vision from close to peripheral. To the left: a reach envelope depicting points that can be reached by the right hand, where all equipment and details within the envelope can be reached comfortably without the need to "lean." (From Sundin and Ortengren, 2006. Courtesy of Delmia/Wiley.)



**Figure 3** Marras's lumbar motion monitor exoskeletal system enabling measures of the dynamic aspects of task. It captures information for calculation of the velocities and accelerations and angular velocities and angular accelerations that assist in predicting risk of lower back musculoskeletal injury (Marras, 2006).

colleagues (2009). Examples for the creation of three-dimensional (3D) libraries for aviation are also available (Oudenhuijzen et al., 2010). Their suggested methodologies are particularly relevant to the further development of advanced first-generation, data-driven models and second-generation, computationally driven models.

## 2.2 Assessment of Cognitive-Based Tasks

Cognitive modeling within the DHM community really represents the third generation of models. The cognitive aspects have been debated within the DHM community. The initial focus of DHM was on improving time to market by leveraging use of the CAE models considering predictions about risk to the lower back, for instance, to reduce the need for physical prototyping. As noted in Sundin and Ortengren (2006), some within the field believe that the cognitive models are underdeveloped. Yamaguchi and Proctor (2009) take exception and suggest that the cognitive models are simply not well known or well integrated into commercially available tools. They refer to some models such as Hick's law (Hick, 1952) that enable predictions of reaction time. This model is also noted by Sheridan in describing the history of human performance modeling. Referring back to Table 1, one can see that this and Shannon's model (Shannon, 1949) on communication in the presence of noise actually precede even the oldest of those initial analysis tools in commercially available DHM software packages. An additional representation of the cognitive aspects of task, independent of the quantitative aspects, is shown pictorially as Wickens's information-processing model in Figure 5 (Wickens et al., 2004; Wickens and Carswell, 2006).

The difference in what is traditionally considered human performance modeling and digital human modeling comes back to the original definition given earlier. The digital human model has a visualization, whereas the human performance modeling community has not included that as an expectation in their model development efforts. So the two communities serve different purposes and continue to coexist without a great deal of overlap. The impact is related to Proctor's comment that
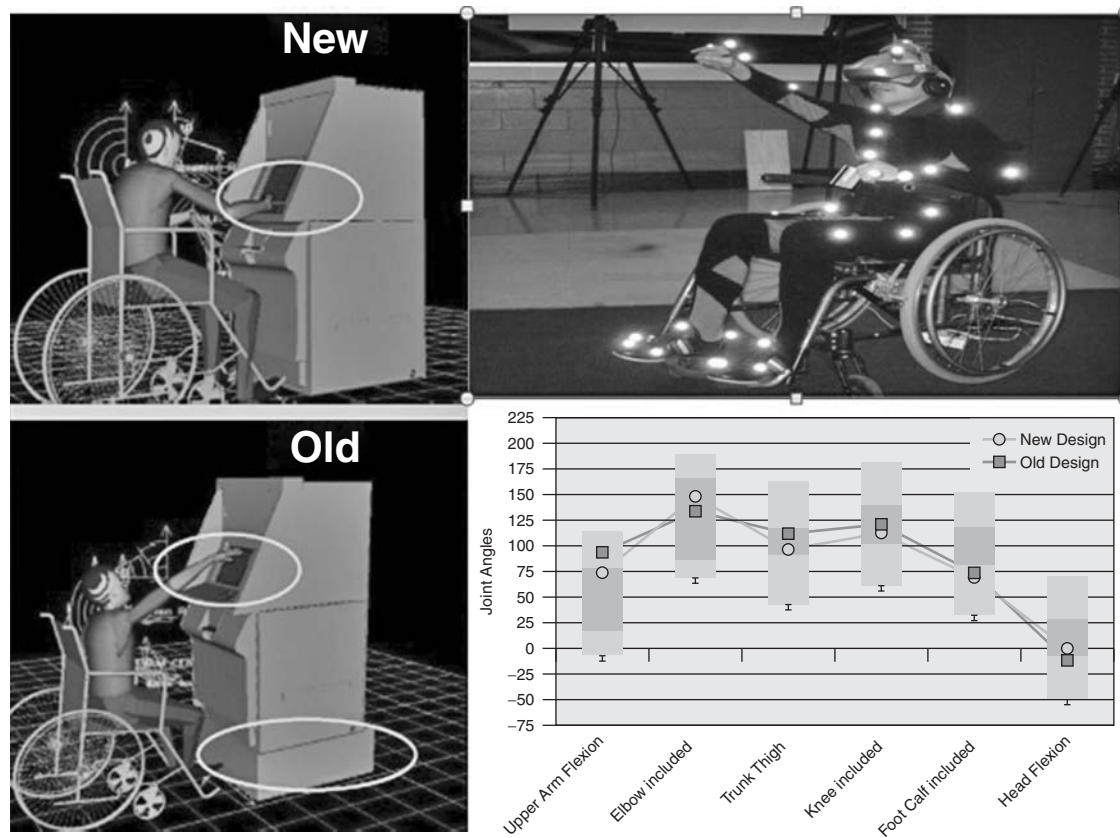
**Figure 4** A person with limited mobility with a head-mounted virtual reality display and reflective markers on a motion capture suit in the upper right. Motion capture, when integrated with computer-aided engineering designs to allow more realistic interaction with devices, can be used to help improve prediction of performance and safety for new designs. In this example, the old ATM design is 30 cm taller, as illustrated in the lower left. The new design, in the upper left, provides improved head flexion and upper arm flexion. Results given in the lower right are based on analysis tools originally developed for automobile design but showed similar outcomes with subjective measures of comfort. (Adapted from Li et al., 2006.)

the cognitive models are not as well known. The digital human modeling community focused on incorporating the visualizations, since the models incorporating the visualization with the math and/or science in the background have lent themselves well to communication that impacts decisions, as evidenced by the ergonomics practitioners at Ford Motor Company (Brazier et al., 2003; Stephens, 2006; Stephens and Jones, 2009).

Examples of cognitive models that incorporate the visualization aspect have been shown in considering various physiological measures such as facial skin temperature with visualizations that provide part of a multimodal measure of mental workload (Or and Duffy, 2007; Thomas et al., 2009). Others that have the cognitive models integrated with the visualization are highlighted by Gore, including MIDAS, the Man-Machine Integration Design and Analysis System (Gore, 2009). In the context of decisions about man–machine integration in the military Lockett and Archer (2009) and Meunier et al. (2009) outline clear examples where the

integration of cognitive and physical considerations in the modeling and visualization domain impact decisions about systems development. These, in most cases, were the predecessors to the Virtual Solider project intended to serve a broader set of product and systems design objectives than the previous examples reviewed by Lockett and Archer. The rationale for the computationally driven approach (previously described as second-generation models) is given by Abdel-Malek and Arora (2009) with consideration for limitations in empirically or data-driven model development (Abdel-Malek and Arora, 2009; Marler et al., 2009) and the synergies that are possible if the two approaches are considered in parallel.

Where models like ACT-R (Adaptive Control of Thought–Rational; Anderson and Libiere, 1998) have been built upon for some integrated and hybrid models (Lockett and Archer, 2009), some have noted limitations of these models and architectures in their current form in terms of integration into the Virtual Soldier Project
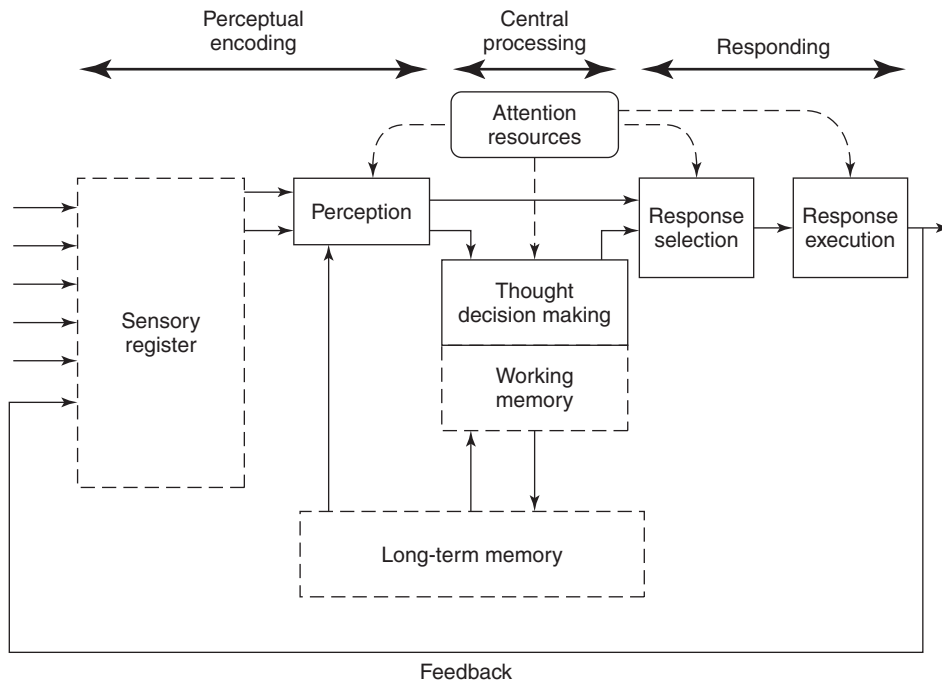
**Figure 5** Wickens's information-processing model. (From Wickens et al., 2004; see also Wickens and Carswell, 2006.)

(Marler et al., 2009). Some driving-related models that consider both vehicle and driver control are based on the ACT-R concept (Salvucci, 2007). It is fortunate that, while noting the limitations for developing and applying driving-related models based on ACT-R, some researchers in Europe (Lenk and Mobus, 2010) are proposing alternatives such as a Bayesian approach (Lenk and Mobus, 2010; Mobus and Eilers, 2010).

## 3 METHODS OF EVALUATION AND ANALYSIS

Depending on the level of complexity of the human–system interaction requirements and the past expertise

developed in simulating the human performance and safety-related outcomes for that type of task, there may be various aspects of the physical world needed in testing that may represent elements of a physical prototype. These may be considered on a continuum (Duffy, 2007a); see Figure 6. This suggests that currently there are limitations in human digital models and in new devices that require hand controls, and in assembly in some cases there may be a need for physical prototyping either fully or partially.

With regard to interactive virtual design there has been discussion about when the actual boundaries of the virtual environment need to be represented physically (Szczerba et al., 2007; Stephens, 2006; Li et al.,
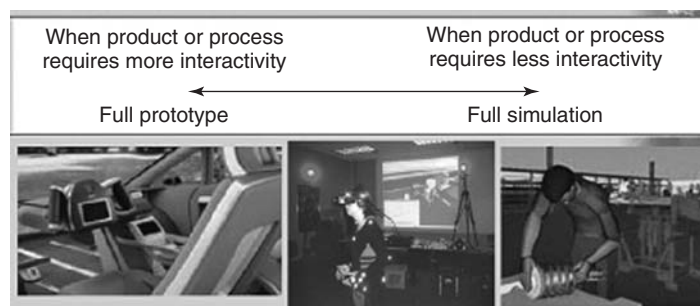


**Figure 6** Virtual prototyping capabilities and limitations: DHM in virtual design shown on a continuum. (Adapted Duffy, 2007a.)

2006). Demirel and Duffy (2009) demonstrate the potential benefits of incorporating force feedback in virtual interactive design assessments. LaFiandra (2009) outlines available methods, models, and technology related to lifting biomechanics. Types of models include the NIOSH lift equation, Snook tables, link segment models, electromyography (EMG) models, and recommendations regarding team lifting. Predicting maximum foot and hand force is highly desired not only for specifying the force limit of industrial workers but also for evaluating hand or foot control which requires a high demand of force (Wang et al., 2010).

Related software referred to by LaFiandra (2009) not previously discussed in Section 2.1 includes University of Michigan's 3D Static Strength Prediction Program (3DSSPP), AnyBody Technology (Denmark: Aalborg University), and Santos (Virtual Soldier Research, University of Iowa). 3DSSPP software predicts static strength requirements for tasks such as lifts, presses, pushes, and pulls considering postural data, force parameters, and anthropometry as input. Even for the controls requiring low force demand, maximum static strength can be used as an objective indicator for defining discomfort evaluation criteria (Wang et al., 2010). AnyBody provides the opportunity to model a range from some subset of the musculoskeletal system to the entire body. Santos is the only human digital modeling software that incorporates an optimization-based approach for analyzing the human in the loop.

Other human digital models are not readily available in commercial tools but serve to support footwear development (Luximon and Goonetilleke, 2009; Luximon et al., 2010). Other clothing development and physically based grasp models are described (Armstrong et al., 2009; Endo et al., 2007, 2010) and are currently available as specialized tools directly from the researchers.

## 4 ORGANIZATIONAL ASPECTS

The visualization aspect of digital human models has helped the decision maker in an organization to understand potential outcomes based on variations in design. In the past, the level of sophistication of the manikin led some to criticize models that had correct math and science, but were not well visualized. As the DHM field is defined by models that include visualization, the quality of that visualization became a measure of interest within the field. As illustrated in Figure 7, as recently as 2006, the visualizations in commercially available CAE and PLM software packages were still not well refined.

Cheng showed that recent research has contributed to more realism and an improvement in predicted shapes for certain manikin poses (Figure 8) (Cheng et al., 2010). A summary of available human digital modeling packages and characteristics is given in Li (2009) with criteria to help in the assessment.

Ways in which various human digital modeling tools could be utilized in an industrial workstation assessment of occupational ergonomics are described in the literature (Du and Duffy, 2007; Lamkull et al. 2009a, 2009b; Raschke et al., 2001). These are mainly in the context of manufacturing ergonomics. Commercially available tools now are distinguishing themselves by noting the improved manikin and visualizations. See Figure 9. Examples of emerging areas and newly developed tools are described in Section 6.

### 4.1 DHM Applications

In this section applications are grouped according to the three areas of development outlined in Bubb and Fritzsche (2009): anthropometric models, biomechanical models, and models for production design. When including cognitive models as one of the categories,
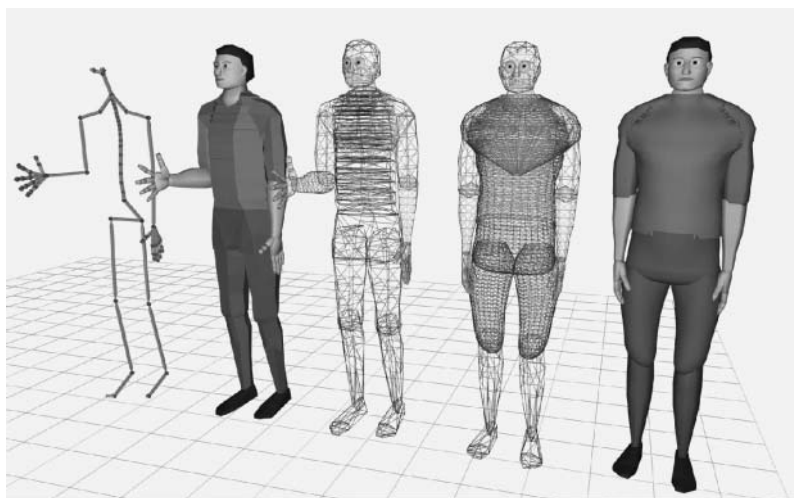


**Figure 7** Representations of a computer manikin (Jack); the linked segments are visible on the extreme left, in the wire frame third from left, and in the shaded frame second from right. (From Sundin and Ortengren, 2006. Courtesy of H. Sjoberg, Chalmers University of Technology/Wiley.)
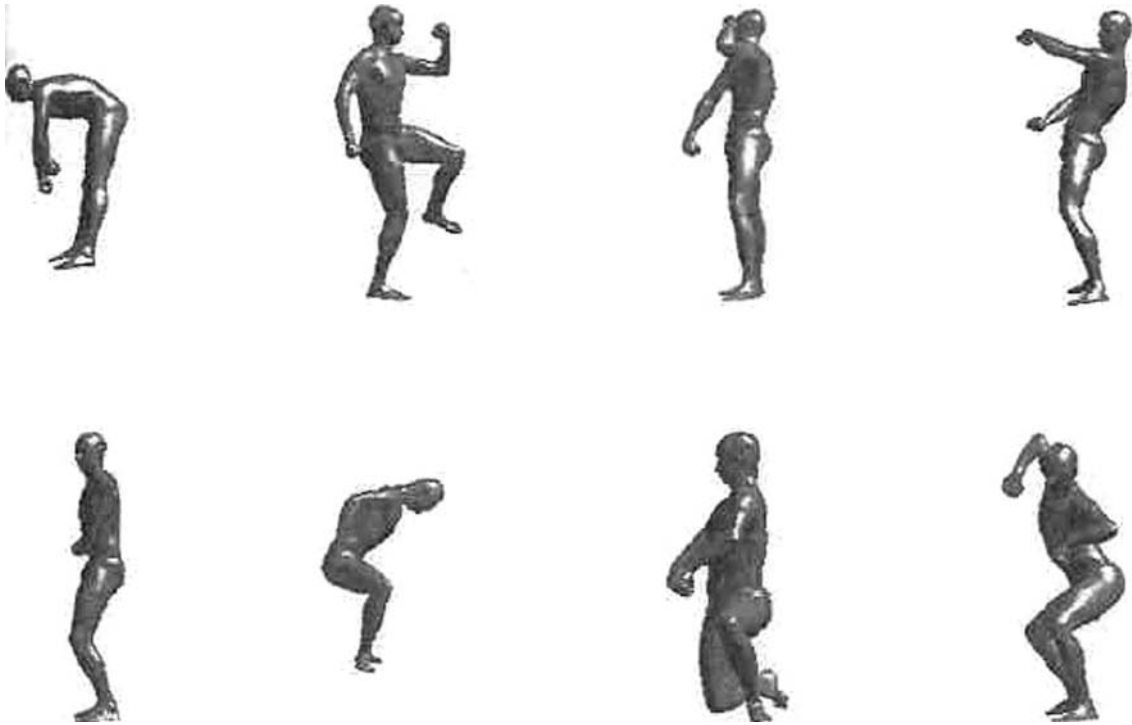
**Figure 8** Predicted shapes in poses demonstrated in recent related research have more realism (Cheng et al., 2010).



**Figure 9** Delmia's new manikin within V.6 of CATIA is shown (left) and in various product promotions including one recently on You Tube. Initially the new manikin is moving side by side with the less sophisticated prior manikin and then suggests the other move to ''retirement'' (Dassault Systemes, 2010).

they suggest that all of the human models would fit within one of the categories, though some features may have crossover between categories. Cognitive models previously were reviewed in Section 2.2. These models have been typically developed for application in workplace design, product design, safety evaluation, and documentation of planning and/or production.

### 4.1.1 Anthropometric

Anthropometric models consider especially shape and size. Anthropometric and anatomical models will be combined and a distinction will be briefly made for anatomical models related to shape. Efforts to measure male and female hand, arm, and leg lengths are

intended to better design products and workstations with a focus on minimizing the number of people excluded (Konz and Johnson, 2008). Frey Law et al. (2009) summarize the body of literature related to modeling human physical capability, including joint strength and range of motion. Comfort reach, maximum reach, and unreachable areas are summarized for standing, kneeling, and prone positions in task and military equipment design for Chinese soldiers (Dong et al., 2010). Researchers in Sweden have demonstrated keen interest in the further development of human digital models for production (Lamkull et al., 2009b). Bertilsson et al. (2010) recently interviewed personnel from car companies about anthropometric diversity in the early stages of development. They generalize the results and suggest that only one or a few human models are actually considered. Personnel claim it is a time-consuming process to create and correctly position the model in the computer-aided design (CAD) environment. A matrix describing important model capabilities and the ability of certain commercial packages to meet those is provided. The challenges described highlight the need for current research, such as that by Green and Hudson (2010) demonstrating specialized methods for positioning such models in airplane passenger seats.

### 4.1.2 Biomechanical

Biomechanical models can consider the mass, inertia, spring, and damping elements that represent the body parts connected by joints (Bubb and Fritzsche, 2009). Section 3 included a brief review of available methods, models, and technology related to lifting biomechanics (LaFiandra, 2009). Park's (2009) overview of data-based human motion simulation gives insight into empirically or data-driven modeling techniques for reach, motion prediction, and obstruction avoidance. Suggestions for future work highlight current limitations, including the cost of determining optimum posture, difficulties for simulations to link long sequences of motion behavior, challenges in integrating multiple motion databases and datasets, and constraints due to individual differences in strength, range of motion, or obesity (Park, 2009). Additional insight is provided by optimization-based posture prediction in recent research using joint angles to serve as design variables and vision-based constraints such as eye movement and visual obstacle avoidance (Knake et al., 2010). This research builds on experience from the Virtual Soldier Project.

Other biomechanical models can simulate physical dynamic behavior. Happee and Wismans (2009) summarize issues related to simulations of human body impact to give insight into the prediction of injury mechanisms and injury criteria. The finite-element modeling methods described represent another example of computationally driven modeling that is shaping the field. The MADYMO (Mathematical Dynamic Models) software can be used for accident reconstruction and can substitute in some cases for data that may otherwise be obtained from the use of crash test dummies that are limited in what they can provide (Happee and Wismans, 2009). Practical applications include occupant response to mine blasts. Recent research provides insight into

comparisons of the use of dummies and commercial finite-element software code (Irde et al., 2010). It is suggested that risk assessment is one of the most fundamental methods for designing safe products and work tasks (Marras, 2006; Koizumi et al., 2010). Injury prevention is one of the most important and urgent issues in children's health since the primary cause of death of children is unintentional injury (Koizumi et al., 2010). Efforts in this recent research consider how to incorporate knowledge of child behavior predictions with biomechanical models, biomechanical simulations, and an injury model.

### 4.1.3 Production

Production models can simulate the work process through discrete-event simulation and can predict the necessary working time while incorporating humans in the loop. Those that would be categorized with human digital models would allow some opportunity to visualize and optimize processes. They can help to identify and eliminate bottlenecks and optimize throughput in manufacturing and services; help in the visualization of product movement and material handling; and improve terminal operations in container terminals and streamline patient flow in health care (Flexsim, 2010). In reviewing virtual environments (Alexander and Ellis, 2009), one is reminded that virtual environments are expected to be interactive, and if digital human models could be more than "puppets," there could be additional value in seeing user or model response. For instance, if intelligent agents or avatars or intelligent conversational agents are included, the emotional, psychological, and sociological aspects are referred to as "missing" from current DHM (Boucsein and Backs, 2009).

## 4.2 Educational Aspects

Undergraduate students planning to work in engineering or product design should have access to commercially available tools before graduation. Their participation does not need to be product specific but should center around a set of capabilities that a well-trained human factors and ergonomics specialist would appreciate as a potential time savings in analysis and communication tool within the organization. These students should become aware of ways to create and interpret the visualizations that communicate information about the user aspects of product and process design to key decision makers within the design process. Zhang et al. (2007) presented methods for the design and implementation of an ergonomics evaluation system of a 3D airplane cockpit.

As the work is changing from more physical to more cognitive and opportunities are growing in the services industries, it will be important for students to identify, encourage, and incorporate these models as they become available as specialized tools or in the human modeling packages that incorporate a variety of analysis tools. Earlier in the development of human modeling tools that were part of PLM software suites, such as Dassault Systemes (2010) CATIA/DELMIA and UGS-Siemens/Jack (Siemens, 2010), new tools were incorporated based on a "demand-driven" basis. When clients

had a need for a particular model that was not yet a part of the PLM package, they would request and sponsor their integration. There has recently been some shift in that expectation, in that PLM developers seem to be more willing to seek and incorporate available and tested models and create capacity for human modeling, letting the user ultimately determine which model to use and when. For instance, discomfort measurements incorporated with motion capture systems that are designed to assess consumer products such as dishwashers (see Preatoni et al., 2010) may be of interest in consumer product development. Their model utilizes gesture and gait and is compared to postural loading on the upper body based on joint discomfort and maximum holding time, a method developed by Kee and Karwowski (2001).

Even though the software may be capable of such analyses, there will be an increasing need in the workplace to consider the cross-disciplinary and system aspects of designs. Though some would like to see such tasks relegated to a checklist-type review, ultimately, in the foreseeable future, the analysis and development considering the human aspects of new products and processes will remain an engineering function, particularly in emerging areas and where there are new applications for human digital models. Andreoni et al. (2010) rely on direct observations through motion capture as well as subjective questionnaire data to develop their consumer product dishwasher assessment tool. As interest in safety of products and customization continues to increase, one important market continues to be the product market for children. To design safe products for children, one must measure and understand foreseeable use and children's behavior quantitatively (Nishida et al., 2010). And as the services sector of the economy continues to grow, so too will the need to construct human models to effectively predict user behavior and consumer preferences (Motomura, 2010).

## 5 CURRENT CHALLENGES

As outlined by Badler and Allbeck (2009), current challenges include issues related to sensing and reacting. Currently the manikins are not modeled to account for the impact of talking on a cell phone on performance (Badler and Allbeck, 2009) or the emotional aspects (Backs and Boucsein, 2009; Boucsein and Backs, 2009). Certainly health care is an application area that provides both opportunities and challenges with regard to current and emerging human digital models. At times, modeling the cognitive or procedural aspects of task is straightforward when compared to modeling the parallel physical aspects. Efforts to model patient safety and reduction of medication administration error are currently considered in the literature (Boone-Seals and Duffy, 2005). Anecdotes and the number of documented incidents have shown to be powerful motivators for improving patient safety. It is apparent, however, that human factors–related health care delivery models lack the visualizations that may facilitate decision making to improve patient safety (Duffy, 2010a). Methods such as data mining have separate bodies of literature from

DHM (Liu, 2009). With the emergence of inexpensive sensor technologies, these data-mining techniques could be blended with human models and visualizations to affect decisions in health care and ultimately safety.

Williams et al. (2009) demonstrated a methodology for assessing a high-fidelity DHM with force or haptic feedback during training. Rapala and Novak (2009) have outlined the field of health care delivery and the need for simulation in training. They consider nurses' perspectives in the context of the health care delivery team. Training and experience can affect perception of hazard and risk (Duffy, 2003). The level of fidelity of a simulation can influence medical training. Professional fragmentation and a tradition of individualism provide some barriers to teamwork (Leape and Berwick, 2005). Behaviors and interactions in teams described by Caldwell (2009) highlight successes centered on resource coordination and information flow efficiency that were demonstrated in the aviation industry and that have some applicability to human digital models. DeLaurentis (2009) suggests methods for modeling the role of human behavior as part of a system of systems. His examples again demonstrate insight from air transportation systems that appear to also have applicability to health care, particularly in agent-based simulations.

## 6 EMERGING OPPORTUNITIES

Issues related to improved driver information systems are outlined by Mobus and colleagues (Mobus and Eilers, 2009, 2010; Mobus et al. 2009). Methodologies for text mining (Noorinaeini and Lehto, 2009) may be helpful for predicting human performance in various cognitive-based tasks. As noted by Badler and Allbeck (2009), once one begins to look at human performance at the task level, new issues arise that transcend biomechanical models and simulations of the past. For instance, as shown in Figure 10, new automation can control a vehicle more effectively than the most accident-prone drivers (Eby and Kantowitz, 2006). However, it is not clear when certain vulnerabilities in task performance warrant acceptance of support or an augmented reality (Schmorrow et al., 2009). A whole body of literature on augmented reality and augmented cognition could yield further insights.

Established industries such as mining need advanced methods and technology to measure safety. Methods established in cooperation with NIOSH Pittsburgh are contributing to that effort (Ambrose, 2009). Virtual reality training also provides opportunities for improving human performance (Sadasivan et al., 2009). Workload assessment capabilities, in the context of discomfort due ultimately to early design decisions, can provide insight into system incompatibilities (Grobelny et al., 2009). People's preferences are a crucial part of the decision-making process—both the potential user and the designer. Modeling usability can help to justify a greater emphasis on the user in design (Grobelny and Michalski, 2010). Utilizing optimization in design can give important insight into design trade-offs that may exist (Parkinson, 2009). Emerging needs in human modeling include the need to identify individuals based
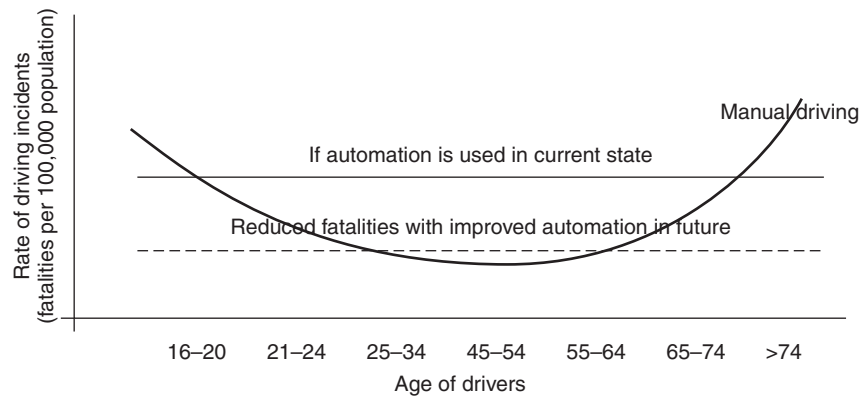
**Figure 10** U.S. fatal crash rates per 100,000 population in 2002. As technology improves, drivers may be willing to relinquish control of the vehicle. Predictive models are needed for circumstances when behavior changes and allows the automation to take over. As well, predictive models are needed for how many lives may be saved as the automation improves and the horizontal line in the figure shifts vertically downward with improvements of in-vehicle automation systems that sense the vehicle environment and inform the driver. [Data from the National Highway Traffic Safety Administration (NHTSA), 2002. Adapted from Eby and Kantowitz (2006), which showed the fatality data without the automation curves.]

on biometric measurement technology. An overview of the body of literature in biometrics, including measures of fingerprint, face, iris, voice, and mult-modal efforts, is provided in Du (2009).

Recent research characterizes hand strength while an extravehicular activity (EVA) glove was worn in an extravehicular mobility unit (EMU) suit. Data were collected in various hand postures under bare-handed, gloved with no thermal micrometeoroid garment (TMG) and gloved with a TMG (Mesloh et al., 2010). It was found that the TMG reduced grip strength to 55% of bare-hand strength in the unpressurized condition and 46% in the pressurized condition. Lateral pinch strength increased (>100%, original), as it seems the glove shape contributed to additional support (Mesloh et al., 2010). Space exploration is one emerging area for clothing design and human performance modeling. Human digital models with a data-based grasp synthesis approach support ergonomic assessment for hand-held product design (Kawaguchi et al., 2009). These can enable some additional predictive capability in computer-aided engineering design environments. As previously mentioned, clothing for soldiers is of current interest within the military (Marler et al., 2009) and footwear design (Luximon and Goonetilleke, 2009) is very popular among a more diverse engineering student body. In order to address these emerging application areas, it is important to incorporate lessons from past DHM related to shape and size analysis (Godil and Ressler, 2009) as well as some scanner-based anthropometry methods (Lu and Wang, 2009).

## REFERENCES

Abdel-Malek, K., and Arora, J. (2009), "Physics-Based Digital Human Modeling: Predictive Dynamics," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 5, pp. 1–33.

Alexander, T. and Ellis, S. R. (2009), "Virtual Environments and Digital Human Models," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 7, pp. 1–17.

Ambrose, D. H. (2009), "Advanced Measurement Methods in Mining," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 43, pp. 1–21.

Anderson, J. R., and Lebiere, C. (1998), *The Atomic Components of Thought*, Lawrence Erlbaum, Mahwah, NJ.

Andreoni, G., Anselmi, L., Costa, F., Mazzola, M., Preatoni, E., Romero, M., and Simionato, B. (2010), "Human Behavior Modeling: A Mixed Approach," in *Advances in Human Factors and Ergonomics in Healthcare*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 77–83.

Applied Human Factors and Ergonomics International (AHFE) (2012). Available at http://www.ahfe2012.org/, accessed December 2011.

Armstrong, T. J., Choi, J., and Ahuja, V. (2009), "Development of Hand Models for Ergonomic Applications," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 12, pp. 1–20.

Backs, R. W., and Boucsein, W. (2009), "Psychophysiology in Digital Human Modeling," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 16, pp. 1–14.

Badler, N. I., and Allbeck, J. M. (2009), "Future Needs and Developments in Support of Computer-Based Human Models," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 56, pp. 1–13.

Bertilsson, E., Hogberg, D., and Hanson, L. (2010), "Digital Human Model Module and Work Process for Considering Anthropometric Diversity," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 568–575.

Boone-Seals, A., and Duffy, V. G. (2005). "Toward a Model for Reducing Medication Administration Errors," *Ergonomics*, Vol. 48, No. 9, pp. 1151–1168.

Boucsein, W., and Backs, R. W. (2009), "The Psychophysiology of Emotion, Arousal, and Personality: Methods and Models," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 35, pp. 1–18.

Brazier, J. B., Tomko, B., and Brow, T. (2003), "The Car That Jill Built," paper presented at The 6th Annual Applied Ergonomics Conference, Dallas, TX, March 11–15, 2003.

Bubb, H., and Fritzsche, F. (2009), "A Scientific Perspective of Digital Human Models: Past, Present, and Future," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 3, pp. 1–30.

Caldwell, B. S. (2009), "Digital Modeling of Behaviors and Interaction in Teams," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 47, pp. 1–10.

Cappelli, T., and Duffy, V. G. (2007), "Motion Capture for Job Risk Classifications Incorporating Dynamic Aspects of Work," *SAE 2006 Transactions Journal of Passenger Cars—Electronic and Electrical Systems*, 2006-01-2317, pp. 1069–1072.

Chadwick, L., and Fallon, E. F. (2010), "HFMEA of a Radiotherapy Information System—Challenges & Recommendations for Future Studies," in *Advances in Human Factors and Ergonomics in Healthcare*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 631–641.

Chaffin, D. B. (2002), "On Simulating Human Reach Motions for Ergonomic Analyses," *HFEM*, Vol. 12, No. 3, pp. 1–13.

Chaffin, D. B. (2005), "Engineers with HFE Education—Survey Results," *ETG Newsletter of the Human Factors and Ergonomics Society*, Vol 3, pp. 2–3.

Chaffin, D. B. (2009), "Some Requirements and Fundamental Issues in Digital Human Modeling," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 2, pp. 1–10.

Chaffin, D. B., Andersson, G. B. J., and Martin, B. J. (1999), *Occupational Biomechanics*, 3rd ed., Wiley, New York.

Cheng, Z., Smith, J., Parakkar, J., Cheng, H., and Robinette, K. (2010), "Human Shape Modeling in Various Poses," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 396–406.

Ciriello, V. M., and Snook, S. H. (1991), "The Design of Manual Handling Tasks: Revised Tables of Maximum Acceptable Weights and Forces," *Ergonomics*, Vol. 34, pp. 1197–1213.

Dassault Systemes (2010), "V6 PLM Video Gallery," available: http://www.3ds.com/products/v6/resource-center/gallery/, accessed November 1, 2010.

DeLaurentis, D. (2009), "Modeling the Role of Human Behaviors in a System-of-Systems," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 22, pp. 1–16.

Demirel, H. O., and Duffy, V. G. (2007a), "Applications of Digital Human Modeling in Industry," in *Digital Human Modeling*, V. G. Duffy, Ed., Lecture Notes in Computer Science, Vol. 4561, Springer-Verlag, Berlin, pp. 824–832.

Demirel, H. O., and Duffy, V. G. (2007b), "Digital Human Modeling for Product Lifecycle Management," in *Digital Human Modeling*, V. G. Duffy, Ed., Lecture Notes in Computer Science, Vol. 4561, Springer-Verlag: Berlin. pp. 372–381.

Demirel, H. O., and Duffy, V. G. (2009), "Impact of Force Feedback on Computer Aided Ergonomic Analyses," in *Digital Human Modeling*, V. G. Duffy, Ed., Lecture Notes in Computer Science, Vol. 5620, Springer-Verlag, Berlin, pp. 608–613.

Dong, D., Wang, L., and Yuan, X. (2010), "Experimental Research on the Confirmation of Human Body Surface Reachable Grade," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 407–414.

Dreyfuss, H. (1993), *The Measure of Man and Woman: Human Factors in Design*, Whitney Library of Design, New York.

Du, J. C., and Duffy, V. G. (2007), "Development of a Methodology for Assessing Industrial Workstations Using Optical Motion Capture Integrated with Digital Human Models," *Occupational Ergonomics*, Vol. 7, No. 1, pp. 11–25.

Du, Y. (2009), "Biometrics," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 36, pp. 1–15.

Duffy, V. G. (2003), "Effects of Training and Experience on Perception of Hazard and Risk," *Ergonomics*, Vol. 46, No. 1, pp. 114–125.

Duffy, V. G. (2007a), "Using the Virtual Build Methodology for Computer-Aided Ergonomics and Safety," *Human Factors and Ergonomics in Manufacturing*, Vol. 17, No. 5, pp. 413–422.

Duffy, V. G., Ed. (2007b), *Digital Human Modeling*, Lecture Notes in Computer Science, Vol. 4561, Springer-Verlag, Berlin.

Duffy, V. G. (2009a) "Introduction," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 1, pp. 1–5.

Duffy, V. G., Ed. (2009b), *Handbook of Digital Human Modeling*, CRC Press, Taylor & Francis Group, Boca Raton, FL.

Duffy, V. G., Ed. (2009c), *Digital Human Modeling*, Lecture Notes in Computer Science, Vol. 5620, Springer-Verlag, Berlin.

Duffy, V. G., Ed. (2010a), *Advances in Human Factors and Ergonomics in Healthcare*, CRC Press, Taylor & Francis Group, Boca Raton, FL.

Duffy, V. G., Ed. (2010b), *Advances in Applied Digital Human Modeling*, CRC Press, Taylor & Francis Group, Boca Raton, FL.

Eby, D. W., and Kantowitz, B. H. (2006), "Human Factors and Ergonomics in Motor Vehicle Transportation," in *Handbook of Human Factors and Ergonomics*, 3rd ed., G. Salvendy, Ed., Wiley, Hoboken, NJ, pp. 1538–1569.

Endo, Y., Kanai, S., Kishinami, T., Miyata, N., Kouchi, M., and Mochimaru, M. (2007), "A Computer-Aided Ergonomic Assessment and Product Design System Using Digital Hands," in *Digital Human Modeling*, V. G. Duffy, Ed., Lecture Notes in Computer Science, Vol. 4561, Springer-Verlag, Berlin, pp. 833–842.

Endo, Y., Miyata, N., Kouchi, M., Mochimaru, M., and Kanai, S. (2010), "Physically-Based Grasp Posture Generation for Virtual Ergonomic Assessment Using Digital Hand," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 167–175.

Feyen, R., Liu, Y., Chaffin, D., Jimmerson, G., and Joseph, B. (2000), "Computer-Aided Ergonomics: A Case Study of Incorporating Ergonomics Analyses into Workplace Design," *Applied Ergonomics*, Vol. 31, No. 3, pp. 291–300.

Flexsim (2010), "General Purpose Simulation Software," available: http://www.flexsim.com, accessed November 5, 2010.

Frey Law, L., Xia, T., and Laake, A. (2009), "Modeling Human Physical Capability: Joint Strength and Range of Motion," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 50, pp. 1–12.

Garg, A., Chaffin, D. B., and Herrin, G. D. (1978), "Prediction of Metabolic Rates for Manual Materials Handling Jobs," *American Industrial Hygiene Association Journal*, Vol. 39, No. 8, pp. 661–674.

Godil, A., and Ressler, S. (2009), "Shape and Size Analysis and Standards," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 14, pp. 1–15.

Gore, B. F. (2009), "Human Performance: Evaluating the Cognitive Aspects," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 32, pp. 1–18.

Green, R. F., and Hudson, J. (2010), "A Method for Positioning Digital Human Models in Airplane Passenger Seats," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 1–8.

Grobelny, J., and Michalski, R. (2010), "Human Preferences Modeling in Usability of Graphical Interfaces," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 510–522.

Grobelny, J., Michalski, R., and Karwowski, W. (2009), "Workload Assessment Predictability for Digital Human Models," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 28, pp. 1–13.

Hanson, L., Hogberg, D., and Nabo, A. (2009), "Digital Human Modeling in Automotive Product Applications," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 40, pp. 1–16.

Happee, R., and Wismans, J. (2009), "Impact Simulation and Biomechanical Human Body Models," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 11, pp. 1–13.

Hick, W. E. (1952), "On the Rate of Gain of Information," *Quarterly Journal of Experimental Psychology*, Vol. 4, pp. 11–26.

Irde, K., Lou, K.-A., Wilhelm, T., and Perciballi, W. (2010), "Occupant Response in Mine Blast Using Different Dummies from Two Commercial FE Codes," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 536–543.

Karhu, O., Kansi, P., and Kuorina, I. (1977), "Correcting Working Postures in Industry: A Practical Method for Analysis," *Applied Ergonomics*, Vol. 8, pp. 199–201.

Kawaguchi, K., Endo, Y., and Kanai, S. (2009), "Database-Driven Grasp Synthesis and Ergonomic Assessment for Handheld Product Design," in *Digital Human Modeling*, V. G. Duffy, Ed., Lecture Notes in Computer Science, Vol. 5620, Springer-Verlag: Berlin. pp. 642–652.

Kee, D., and Karwowski, W. (2001), "LUBA: An Assessment Technique for Postural Loading on the Upper Body Based on Joint Motion Discomfort and Maximum Holding Time," *Applied Ergonomics*, Vol. 32, No. 4, pp. 357–366.

Knake, L., Mathai, A., Marler, T., Farrell, K., Johnson, R., and Abdel-Malek, K. (2010), "New Capabilities for Vision-Based Posture," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 628–638.

Koizumi, Y., Nishida, Y., Motomura, Y., Miyazaki, Y., and Mizoguchi, H. (2010), "A System for Presenting Potential High-Risk Situation by Integrating Biomechanical, Injury, and Child Behavior Model," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 597–606.

Konz, S., and Johnson, S. (2008), *Work Design: Occupational Ergonomics*, 7th ed., Holcomb-Hathaway, Scottsdale, AZ.

LaFiandra, M. (2009), "Methods, Models, and Technology for Lifting Biomechanics," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 8, pp. 1–27.

Lamkull, D., Berlin, C., and Ortengren, R. (2009a), "Digital Human Modeling: Evaluation Tools," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 24, pp. 1–22.

Lamkull, D., Ortengren, R., and Malmskold, L. (2009b), "Digital Human Modeling Automotive Manufacturing Applications," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 42, pp. 1–42.

Leape, L., and Berwick, D. (2005), "Five Years After To Err Is Human: What Have We Learned? *Journal of the American Medical Association*, Vol. 292, No. 19, pp. 2384–2390.

Lenk, J., and Mobus, C. (2010), "Modeling Lateral and Longitudinal Control of Human Drivers with Multiple Regression Models," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 446–456.

Li, K., Duffy, V. G., and Zhang, L. (2006), "Universal Accessibility Assessments through Virtual Interactive Design," *International Journal of Human Factors Modelling and Simulation*, Vol. 1, No. 1, pp. 52–68.

Li, Z. (2009), "Digital Human Modeling Packages," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 54, pp. 1–21.

Liu, Y. (2009), "Data Mining and its Applications in Digital Human Modeling," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 37, pp. 1–15.

Lockett, J. F., and Archer, S. (2009), "Impact of Digital Human Modeling on Military Human-Systems Integration and Impact of the Military on Digital Human Modeling," in

*Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 51, pp. 1–30.

Lu, J.-M., and Wang, M.-J. (2009), "Digital Human Modeling and Scanner-Based Anthropometry," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 53, pp. 1–11.

Luximon, A., and Goonetilleke, R. (2009), "Foot Modeling and Footwear Development," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 13, pp. 1–21.

Luximon, A., Luximon, Y., Wong, D.W-C., and Zhang, M. (2010), "The Geriatric 3D Foot Shape," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 176–185.

Marler, T., Arora, J., Beck, S., Lu, J., Mathai, A., Patrick, A., and Swan, C. (2009), "Computational Approaches in Digital Human Modeling," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 10, pp. 1–33.

Marras, W. S. (2006), "Basic Biomechanics and Workstation Design," in *Handbook of Human Factors and Ergonomics*, 3rd ed., G. Salvendy, Ed., Wiley, Hoboken, NJ, pp. 340–370.

McAtamney, L., and Corlett, E. N. (1993), "RULA: A Survey Method for the Investigation of Work-Related Upper Limb Disorders," *Applied Ergonomics*, Vol. 24, pp. 91–99.

Mesloh, M., England, S., Benson, E., Thompson, S., and Rajulu, S. (2010), "The Effects of Extravehicular Activity (EVA) Glove Pressure on Hand Strength," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 156–165.

Meunier, P., Cain, B., and Morrissey, M. (2009), "Advanced Human Modeling in Support of Military Systems," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 52, pp. 1–10.

Mobus, C., and Eilers, M. (2009), "Further Steps towards Driver Modeling According to the Bayesian Programming Approach," in *Digital Human Modeling*, V. G. Duffy, Ed., Lecture Notes in Computer Science, Vol. 5620, Springer-Verlag, Berlin, pp. 413–422.

Mobus, C., and Eilers, M. (2010), "Mixture-of-Behaviors and Levels-of-Expertise in a Bayesian Autonomous Driver Model," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 425–435.

Mobus, C., Eilers, M., Garbe, H., and Zilinski, M. (2009), "Probabilistic and Empirical Grounded Modeling of Agents in (Partial) Cooperative Traffic Scenarios," in *Digital Human Modeling*, V. G. Duffy, Ed., Lecture Notes in Computer Science, Vol. 5620, Springer-Verlag: Berlin. pp. 423–432.

Morr, D. R., Wiechel, J. F., and Metzler, S. A. (2009), "Instrumentation in Support of Dynamic Digital Human Modeling," in *Handbook of Digital Human Modeling*, Vol. 33, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 1–34.

Motomura, Y. (2010), "Large-Scale Data Based Modeling in Everyday Life for Service Engineering," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 576–586.

National Highway Traffic Safety Administration (NHTSA) (2003), "Traffic Safety Facts, 2002," U.S. Department of Transportation, Washington, D.C.

Nishida, Y., Motomura, Y., Kitamura, K., and Yamanaka, T. (2010), "A Social System That Circulates Safety Knowledge for Injury Prevention," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 587–596.

Noorinaeini, A., and Lehto, M. R. (2009), "Mathematical Models of Human Text Classification," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 17, pp. 1–15.

Or, C. K. L., and Duffy, V. G. (2007), "Development of a Facial Skin Temperature-Based Methodology for Non-Intrusive Mental Workload Measurement," *Occupational Ergonomics*, Vol. 7, No. 2, pp. 83–94.

Oudenhuijzen, A., Zehner, G. F., and Hudson, J. A. (2009), "Verification and Validation of Human Modeling Systems," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 23, pp. 1–11.

Oudenhuijzen, A., Zehner, G., Hudson, J., and Choi, H. (2010), "On the Creation of 3D Libraries for F-16 Pilots in Their Crew Station: Method Development, Library Creation and Validation," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 9–18.

Park, W. (2009), "Data-Based Human Motion Simulation," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 9, pp. 1–17.

Parkinson, M. B. (2009), "Optimization in Design: A DHM Perspective," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 26, pp. 1–12.

Peacock, B., and Karwowski, W., Eds. (1993), *Automotive Ergonomics*, Taylor & Francis, Washington, DC.

Porter, J. M., and Gyi, D. E. (1998), "Exploring the Optimum Posture for Driver Comfort," *International Journal of Vehicle Design*, Vol. 19, No. 3, pp. 255–266.

Porter, J. M., Marshall, R., Case, K., Gyi, D. E., Sims, R. E., and Summerskill, S. (2009), "Inclusive Design for Mobility Impaired," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 41, pp. 1–13.

Preatoni, E., Mazzola, M., and Andreoni, G. (2010), "Assessment of Dynamic Reaching Tasks through Novel Ergonomic Index," in *Advances in Human Factors and Ergonomics in Healthcare*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 102–108.

Rapala, K., and Novak, J. C. (2009), "Health Care Delivery and Simulation," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 49, pp. 1–8.

Raschke, U., Schutte, L. M., and Chaffin, D. B. (2001), "Ergonomics in Digital Environments," in *Handbook of Industrial Engineering*, 3rd ed., G. Salvendy, Ed., Wiley, New York, pp. 1111–1130.

Rebiffe, R. (1966), "An Ergonomic Study of the Arrangement of Driving Position in Motor Cars," *Journal of the Institute of Mechanical Engineers Symposium* (*London*).

Rohmert, W. (1973), "Problems in Determining Rest Allowances, Part 1: Use of Modern Methods to Evaluate Stress and Strain in Static Muscular Work," *Applied Ergonomics*, Vol. 4, No. 2, pp. 91–95.

Sadasivan, S., Stringfellow, P. F., and Gramopadhye, A. (2009), "Virtual Reality Training to Improve Human Performance," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, chapter 44, pp. 1–21.

Salvucci, D. D. (2007), "Integrated Models of Driver Behavior," in *Integrated Models of Cognitive Systems*, W. D. Gray, Ed., Oxford University Press, New York, pp. 356–367.

Schmorrow, D. D., Muller, P. B., Kobus, D. A., Van Orden, K., Reeves, L., and Rossi, K. A. (2009), "Modeling and Augmented Cognition: Supporting Optimal Individual Warfighter Human Performance," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 55, pp. 1–13.

Shannon, C. E. (1949), "Communication in the Presence of Noise," *Proceedings of the IRE*, Vol. 37, pp. 10–22.

Sheridan, T. (2009), "Historical Perspectives on Human Performance Modeling," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 4, pp. 1–10.

Siemens (2010), "Jack and Process Simulate Human," available: http://www.plm.automation.siemens.com/en_us/products/tecnomatix/assembly_planning/jack, accessed November 5, 2010.

Stephens, A. M. (2006), "The Truck That Jack Built: Digital Human Models and Their Role in the Design of Work Cells and Product Design," paper presented at the SAE International Conference, Digital Human Modeling for Design and Engineering, Lyon, France, July 2–6.

Stephens, A., and Jones, M. L. H. (2009), "Workplace Methods and Use of Digital Human Models," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 6, pp. 1–11.

Sundin, A., and Ortengren, R. (2006), "Digital Human Modeling for CAE Applications," in *Handbook of Human Factors and Ergonomics*, 3rd ed., G. Salvendy, Ed., Wiley, Hoboken, NJ, pp. 1053–1078.

Szczerba, J., Duffy, V. G., Geisler, S., Rowland, Z., and Kang, J. (2007), "A Study in Driver Performance: Alternative Human-Vehicle Interface for Brake Actuation," *Society of Automotive Engineers, SAE 2007 Transactions—Journal of Engines*, 2006-01-1060, Vol. 115, Part 3, pp. 605–610.

Thomas, N. L., Du, Y., Artavatkun, T., and She, J.-H. (2009), "Non-Intrusive Personalized Mental Workload Evaluation for Exercise Intensity Measure," in *Digital Human Modeling*, V. G. Duffy, Ed., Lecture Notes in Computer Science, Vol. 5620, Springer-Verlag: Berlin. pp. 315–322.

Tian, R., Duffy, V. G., and McGinley, J. (2007), "Effecting Validity of Ergonomics Analysis during Virtual Interactive Design," in *Digital Human Modeling*, V. G. Duffy, Ed., Lecture Notes in Computer Science, Vol. 4561, Springer-Verlag, Berlin, pp. 988–997.

Wang, X. (2009), "Discomfort Evaluation and Motion Measurement," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 25, pp. 1–8.

Wang, X., Berelle, C., Pannetier, R., Numa, J., and Chapuis, T. (2010), "A Data-Based Approach for Predicting Variation Range of Hand and Foot Maximum Force on a Control: Application to Hand Brake," in *Advances in Applied Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 369–376.

Waters, T. R., Putz-Anderson, V., Garg, A., and Fine, L. J. (1993), "Revised NIOSH Equation for the Design and Evaluation of Manual Lifting Tasks," *Ergonomics*, Vol. 36, No. 7, pp. 749–776.

Wickens, C. D., and Carswell, C. M. (2006), "Information Processing," in *Handbook of Industrial Engineering*, 3rd ed., G. Salvendy, Ed., Wiley, New York, pp. 111–149.

Wickens, C. D., Lee, J., Liu, Y., and Gordon-Becker, S. (2004), *An Introduction to Human Factors Engineering*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ.

Williams, L., and Bagian, J. (2010), "Teaching Healthcare Clinicians to Demand Safe Healthcare Delivery Systems," in *Advances in Human Factors and Ergonomics in Healthcare*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 19–25.

Williams, R. L., Howell, J. N., and Conaster, R. R. (2009), "Digital Human Modeling for Palpatory Medical Training with Haptic Feedback," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 48, pp. 1–11.

Yamaguchi, M., and Proctor, R. W. (2009), "Modeling Response Time and Accuracy for Digital Humans," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 15, pp. 1–11.

Yang, J. (2009), "Human Modeling and Simulation," in *Handbook of Digital Human Modeling*, V. G. Duffy, Ed., CRC Press, Taylor & Francis Group, Boca Raton, FL, Chapter 29, pp. 1–23.

Zhang, L., Yuan, X., Wang, L., and Dong, D. (2007), "Design and Implementation of Ergonomics Evaluation System of 3D Airplane Cockpit," in *Digital Human Modeling*, V. G. Duffy, Ed., Lecture Notes in Computer Science, Vol. 4561, Springer-Verlag, Berlin, pp. 1047–1055.

# CHAPTER **36**

# VIRTUAL ENVIRONMENTS*

**Kay M. Stanney**
**Design Interactive, Inc.**
**Oviedo, Florida**


**Joseph V. Cohn**
**Office of Naval Research**
**Arlington, Virginia**

## 1 INTRODUCTION

Virtual environments (VEs) aim to immerse users in realistic settings, allowing them to engage in an intuitive and intimate manner with their digital universe. Even a decade ago, VE technologies and their applications were immature and few in number. This has changed dramatically, with a recent analysis of the virtual simulation training market revealing that today commercial and customized virtual simulation and training products abound (King, 2009). Virtual environment technologies have many advantages, including the ability to provide adaptable, modest cost, deployable, and safe training solutions, offer rehabilitation and medical applications that reach far beyond the conventional, and create learning and game-based virtual experiences that would otherwise be impossible to explore. Yet limitations do exist, in that virtual experiences cannot fully replace the benefits of real experiences, there is not yet a full understanding on how best to use this technology,

and the technology does not always meet the expectations of its users due to issues such as cybersickness and lack of presence. The future looks bright, however, as the gaming industry is pushing the realm of the possible and making it ever more feasible to "learn by doing," "train like we fight," and "involve me and I understand." This chapter reviews the current state-of-the-art in VE technology, provides design and implementation strategies, discusses health and safety concerns and potential countermeasures, and presents the latest in VE usability engineering approaches. Current efforts in a number of application domains are reviewed. The chapter should enable readers to better specify design and implementation requirements for VE applications and prepare them to use this advancing technology in a manner that minimizes health and safety concerns.

## 2 SYSTEM REQUIREMENTS

A VE is a computer-generated immersive environment that can simulate both real and imaginary worlds, oftentimes in three dimensions. Current VE applications are primarily intriguing visual and auditory experiences,

---

* This chapter represents a modified version of K. M. Stanney and J. V. Cohn, "Virtual Environments," in *Handbook of Human−Computer Interaction* (2012, 3rd Edition), J. Jacko, Ed., Taylor & Francis.
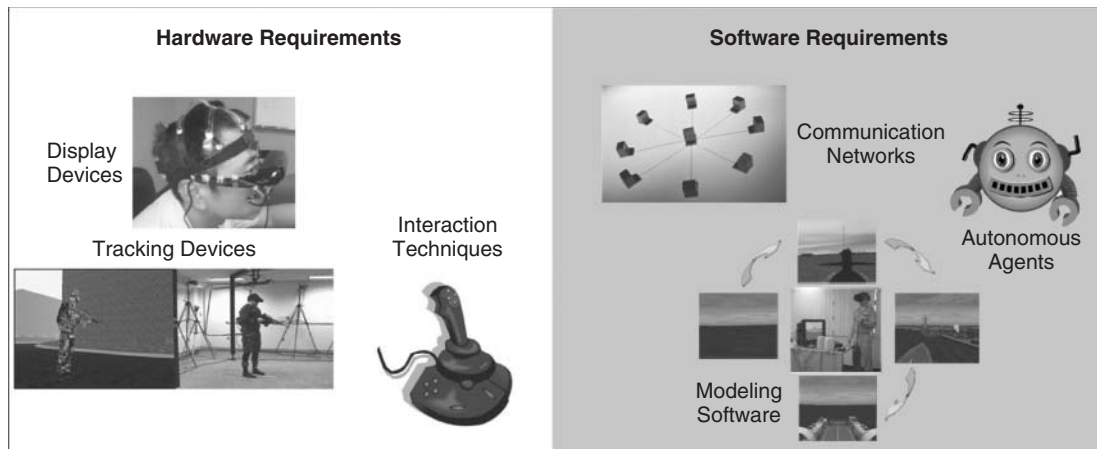
**Figure 1**  Hardware and software requirements for virtual environment generation.

with a smaller number incorporating additional sensory modalities, such as haptics and smell. These worlds are driven by hardware, which provides the hosting platform and multimodal presentation, which allows for physical interaction and tracks the whereabouts of users as they traverse the virtual world, as well as software to model and generate the virtual world and their autonomous agents and support communication networks that link multiple users (see Figure 1).

More specifically, hardware interfaces consist primarily of:

- Interface devices used to present multimodal information and sense the VE
- Tracking devices used to identify head and limb position and orientation
- Interaction techniques that allow users to navigate through and interact with the virtual world

Software interfaces include:

- Modeling software used to generate VEs
- Autonomous agents that inhabit VEs
- Communication networks used to support multiuser virtual environments

## 2.1  Hardware Requirements

Virtual environments require very large physical memories, high-speed processors, high-bandwidth mass storage capacity, and high-speed interface ports for interaction devices (Durlach and Mavor, 1995). These requirements are easily met by today's high-speed, high-bandwidth computing systems, many of which have surpassed the gigahertz barrier. The future looks even brighter, with promises of massive parallelism in multicore and many-core processor architectures (Holmes et al., 2010), which will allow tomorrow's computing systems to be exponentially faster than their ancestors. With the rapidly advancing ability to generate complex

and large-scale virtual worlds, hardware advances in multimodal input/output (I/O) devices, tracking systems, and interaction techniques are needed to support generation of increasingly engaging virtual worlds. In addition, the coupling of augmented cognition and VE technologies can lead to substantial gains in the ability to evaluate their effectiveness.

### 2.1.1  Multimodal I/Os

To present a multimodal VE (see Chapter 14), multiple devices are used to present information to VE users. In terms of VE projection systems, the one that has received the greatest attention, both in hype and disdain, is almost certainly the head-mounted display (HMD). One benefit of HMDs is their compact size, as an HMD when coupled with a head tracker can be used to provide a similar visual experience as a multitude of bulky displays associated with spatially immersive displays (SIDs) and desktop solutions. In addition, HMDs are suggested to enhance situation awareness, enable correct decision making, and reduce workload by allowing users to turn their head and eyes to fully perceive the environment, decreasing multimodal clutter, providing an intuitive means of presenting spatialized multimodal warnings and alerts, and redundantly coding critical cues (e.g., external threats, navigational waypoints), for example, by using audio cues to direct visual attention (Melzer and Rash, 2009).

There are three main types of HMDs: monocular (e.g., one image source is viewed by a single eye), biocular (e.g., one image source viewed by both eyes), and binocular (e.g., stereoscopic viewing via two image generators, with each eye viewing an independent image source) (Melzer et al., 2009). A monocular HMD design is best when projecting moving maps or text information that must be read on the move (e.g., dismounted Warfighter) or to allow viewing of imagery with the simplest, lightest (in terms of head-supported weight/mass), and least costly (both monetarily and in terms of power consumption) solution. The downside of monocular displays is that they have a small field of view (FOV),

convey no stereoscopic depth information, have the potential for a laterally asymmetric center of mass (CM), and may have issues associated with focus, eye dominance, binocular rivalry, and ocular-motor instability. For a wide FOV, more effective target recognition, and a more comfortable viewing experience, a biocular or binocular solution is needed. Biocular solutions present no interocular rivalry and are lighter, easier to adjust, and less expensive than binocular solutions. The disadvantages of biocular displays are that they are heavier, more complex to align, focus, and adjust, and have reduced luminance as compared to monocular displays. Binocular displays have a symmetrical CM and can present stereo viewing (via field-sequential single-screen displays with shutter glasses, single-screen polarized displays, or dual-screen HMDs), which provides for better depth information than monocular and biocular solutions. On the downside, binocular solutions are heavy, require more complex alignment, focus, and adjustments than monocular, and are expensive. Biocular and binocular solutions are particularly well suited when creating fully immersive VEs for gaming or training systems, as their large FOV provides a more compelling sense of immersion.

When coupled with tracking devices, HMDs can be used to present three-dimensional (3D) visual scenes that are updated as a user moves his or her head about a virtual world. Although this often provides an engaging experience, due to poor optics, sensorial mismatches, and slow update rates, these devices are also often associated with adverse effects such as eyestrain and nausea (Stanney and Kennedy, 2008). In addition, while HMDs have come down substantially in weight, rendering them more suitable for extended wear, they are still hindered by cumbersome designs, obstructive tethers, suboptimal resolution, and insufficient FOVs. These shortcomings may be the reason behind why, in a review of HMD devices, approximately a third had been discontinued by their manufacturers (Bungert, 2007). Nevertheless, of the HMDs available, there are several low- to mid-cost models, which are relatively lightweight and provide a horizontal FOV and resolution far exceeding predecessor systems.

Low-technology stereo viewing VE display options include anaglyph methods, where a viewer wears glasses with distinct color-polarized filters, usually with the left-image data placed in the red channel of an electronic display and the right-image data in the blue channel; parallel or cross-eyed methods, in which right and left images are displayed adjacently (parallel or crossed), requiring the viewer to actively fuse the separate images into one stereo image; parallax barrier displays, in which an image is made by interleaving columns of two images from a left- and right-eye perspective image of a 3D scene; polarization methods, in which the images for the left and right eyes are projected on a plane through two orthogonal linearly polarizing filters (e.g., the right image is polarized horizontally; the left is polarized vertically) and glasses with polarization filters are donned to see the 3D effect; Pulfrich methods, in which an image of a scene moves sideways across the viewer's FOV and one eye is covered by a dark filter so

that the darkened image reaches the brain later, causing stereo disparity; and shutter glass methods in which images for the right and left eyes are displayed in quick alternating sequence and special shutter glasses are worn that "close" the right or left eye at the correct time (Konrad and Halle, 2007; Vince, 2004). All of these low-technology solutions are limited in terms of their resolution, the maximum number of views that they can display, and clunky implementation; they can also be associated with pseudoscopic images (e.g., the depth of an object can appear to flip inside out).

Other options in visual displays include SIDs (e.g., displays that surround viewers physically with panoramic large FOV imagery generally projected via fixed front or rear projection display units; Konrad and Halle, 2007; Majumder, 2003), desktop stereo displays, and volumetric displays that fill a volume of space with a "floating" image (Konrad and Halle, 2007). Examples of SIDs include the Cave Automated Virtual Environment (CAVE) (Cruz-Neira et al., 1993), Blue-c, ImmersaDesk, PowerWall, Infinity Wall, and VisionDome (Majumder, 1999). Issues with SIDs include a stereo view that is correct for only one or a few viewers, noticeable overlaps between adjacent projections, and image warp on curved screens. Blue-c addresses some of these concerns by combining simultaneous acquisition of multiple 3D video streams with advanced 3D projection technology (Gross et al., 2003). Desktop display systems have advantages over SIDs because they are smaller, easier to configure in terms of mounting cameras and microphones, easier to integrate with gesture and haptic devices, and more readily provide access to conventional interaction devices, such as mice, joysticks, and keyboards. Issues with such displays include stereo that is only accurate for one viewer and a limited-display volume. Volumetric displays provide visual accommodation depth cues and vertical parallax, which are particularly useful for scenes that require viewing from a multitude of viewing angles, generally without the need for goggles; however, they do not maintain accurate occlusion cues (often considered the strongest depth cues) for all viewers (Konrad and Halle, 2007). Perspecta is an example of a swept-volume display that uses a flat, double-sided screen with a rotating projected image to sweep out a hemispherical image volume (Favalora, 2005). DepthCube is an example of a static-volume display that uses electronically addressable elements [i.e., a digital micromirror device (DMD) imaging system] to scan out the image volume (Sullivan, 2004). Issues with volumetric displays include low resolution and the tendency for transparent images to lose interposition cues. Also, view-independent shading of objects is not possible with volumetric displays, and current solutions do not exhibit arbitrary occlusion by interposition of objects (Konrad and Halle, 2007). The way of the future seems to be direct virtual retinal displays, where images are projected directly onto the human retina with a low-energy laser or liquid crystal displays (LCDs) (McQuaide et al., 2003), as well as displays that represent the physical world around us, such as autostereoscopic omnidirectional light field displays,

which present interactive 3D graphics to multiple simultaneous viewers 360° around the display (Jones et al., 2007). If designed effectively, these next-generation devices should eliminate the tethers and awkwardness of current designs while enlarging the FOV and enhancing resolution.

When virtual environments provide audio (see Chapter 9), the interactive experience is generally greatly enhanced (Shilling and Shinn-Cunningham, 2002). Audio can be presented via spatialized or nonspatialized displays. Just as stereo visual displays are a defining factor for VE systems, so are "interactive" spatialized audio displays (e.g., those with "on-the-fly" positioning of sounds). VRSonic's SoundScape3D (http://www.vrsonic.com/), Firelight's FMod (http://www.fmod.org/), and AuSIM3D (http://ausim3d.com/) are examples of positional 3D audio technology. There have been promising developments in new sound modeling paradigms (e.g., VRSonic's ViBe technology) and sound design principles that will hopefully lead to a new generation of tools for designing effective spatial–audio environments (Fouad, 2004; Fouad and Ballas, 2000; Jones et al., 2005).

Developers must decide if sounds should be presented via headphones or speakers. For nonspatialized audio, most audio characteristics (e.g., timbre, relative volume) are generally considered to be equivalent whether projected via headphones or speakers. This is not so for spatialized audio, in which the presentation technique impacts how audio is rendered for the display and presents the developer with important design choices.

While in the past headphone spatialization required expensive, specialized hardware to achieve real-time rates, modern multicore processors as well as the availability of powerful graphics processing units (GPUs) have made it possible to render complex audio environments over headphones using general-purpose computers. With binaural rendering, a sound can be placed in any location, right or left, up or down, near or far, via the use of a head-related transfer function (HRTF) to represent the manner in which sound sources change as a listener moves his or her head (Begault, 1994; Butler, 1987; Cohen, 1992). For optimal results, however, the HRTFs used for rendering must be personalized for each individual user. One method of doing this is to actually measure each user's HRTF for use in rendering. This approach generally involves a fairly lengthy measurement procedure using specialized hardware. Recently, there have been efforts to develop fast and low-cost approaches to HRTF measurements (Zotkin et al., 2006) that may, in the future, make personalized HRTF rendering practical for general use. An alternative approach to measured HRTFs is to use a best-fit HRTF selection process in which one finds the nearest matching HRTF in a database of candidate HRTFs by either comparing the physiological characteristics of stored HRTFs to those of a target user (Algazi et al., 2001) or using a subjective selection process to find the best-fit HRTF (Seeber, 2003). Other considerations that should be taken into account when choosing headphone rendering are that, for immersive displays, head trackers must be used to

achieve proper relative positioning of sound sources. Also, rendering spatial audio for groups of users over headphones may not be practical for more than a few users.

An alternative approach to headphone spatialization is the use of loudspeaker arrays (Ballas et al., 2001). Loudspeaker arrays can range in size from relatively small surround-sound configurations with 2, 4, 5, 7, or 10 loudspeakers up to hundreds of loudspeakers. The differentiating factors among loudspeaker arrays are the speaker layouts, number of loudspeakers comprising the array, and algorithms used to render spatial audio. Generally speaking, increasing the number of loudspeakers in the array results in more accurate spatialization. The manner in which loudspeakers are laid out in the listening area is closely related to the size of the array. Planar loudspeaker configurations require a smaller number of loudspeakers but are only capable of creating a 2D sound field. Volumetric configurations, on the other hand, can create a 3D sound field but require a larger number of loudspeakers and a more elaborate setup. Recently, VRSonic introduced a spherical loudspeaker array system called the AcoustiCurve. It provides a volumetric array in a spherical configuration around the listening space.

The rendering algorithm used for spatialization is also closely tied to the loudspeaker array size and configuration. Pairwise panning algorithms are the simplest form of spatialization and create a positional sound source by manipulating the amplitude of the signal arriving at two adjacent loudspeakers in the array (Mouba, 2009). An extension to this idea is vector base amplitude panning (VBAP), where the source is panned among three loudspeakers forming a triangle in a volumetric array (Pulkki, 1997). Another spatialization algorithm that is gaining popularity is wave field synthesis (WFS), a technique based on Huygens's principle (Spors and Ahrens, 2010). The WFS technique creates a positional source within the listening space by re-creating the incident wave front of a virtual source using a loudspeaker array. The advantage of WFS is that it does not suffer from the "sweet spot" problem so listeners can get an accurate impression of the synthesized sound field at any location within the listening space; this is not the case with pairwise panning (Shilling and Shinn-Cunningham, 2002). The primary drawback of WFS is that it requires a large number of loudspeakers and considerable processing power to re-create the incident wave front.

Whether using headphones or loudspeaker arrays, spatialization is only one component of simulating a sound field and developers should carefully consider the level of fidelity required by the application when choosing an audio rendering system. Properly synthesizing a virtual soundscape requires modeling the full propagation path of sound, including source model, spreading loss, air absorption, material absorption, and material reflection. Accurately modeling the full propagation path in real time is beyond the capabilities of current computers. There is, however, promising research in the use of GPU processors to achieve real-time rates using ray casting methods (Jedrzejewski, 2004).

While not as commonly incorporated into VEs as visual and auditory interfaces, haptic devices (see Chapter 10) can be used to enhance aspects of touch and movement of the hand or body segments while interacting with a virtual environment. Haptic devices have been classified as passive (unidirectional, e.g., keyboard, mouse, trackball) versus active (bidirectional, thereby supporting two-way communications between human and interactive system; Hale and Stanney, 2004; e.g., force reflecting robotic arm), grounded (e.g., joystick) versus ungrounded (e.g., exoskeleton-type haptic devices), net-force (e.g., PHANTOM device or textured surfaces) versus tactile devices (e.g., tactile pin arrays), and impedance control (i.e., user's input motion is measured and an output force is returned) versus admittance control (e.g., user's input forces are measured and motion is fed back to the user) (Basdogan and Loftin, 2008). In general, haptic displays are effective at alerting people to critical tasks (e.g., warning), providing a spatial frame of reference within one's personal space, and supporting hand–eye coordination tasks. Texture cues, such as those conveyed via vibrations or varying pressures, are effective as simple alerts and may speed reaction time and aid performance in degraded visual conditions (Akamatsu, 1994; Biggs and Srinivasan, 2002; Massimino and Sheridan, 1993; Mulgund et al., 2002). Kinesthetic devices are advantageous when tasks involve hand–eye coordination (e.g., object manipulation), where haptic sensing and feedback are key to performance. Currently available haptic interaction devices include static displays (e.g., convey deformability or Braille); vibrotactile, electrotactile, and pneumatic displays (e.g., convey tactile sensations such as surface texture and geometry, surface slip, surface temperature); force feedback systems (e.g., convey object position and movement distances); and exoskeleton systems (e.g., enhance object interaction and weight discrimination) (Hale and Stanney, 2004). Minamizawa et al. (2008) suggest that to provide natural haptic feedback, such interfaces should be bimanual and wearable and aim to enhance the existence and operability of virtual objects while not disturbing the motion and behavior of users. Currently, there are several wearable haptic displays that can be used in virtual environments, such as Cyber-Glove Systems' CyberGlove, CyberTouch, CyberGrasp, and CyberForce (http://www.cyberglovesystems.com/) and Immerz's KOR-fx (Kinetic Omnidirectional Resonance effect) acousto-haptic technology, the latter of which translates the audio signals from an interactive environment into vibrations that can be felt throughout the body and experienced as the sensation of rain, wind, weight shift, and G-forces (www.Immerz.com). Beyond supporting hand–eye coordination tasks and conveying simple alerts, haptics can be used to communicate grammar structured strings of tactile symbols (Fuchs et al., 2008). Such a tactile language has been used at a concept level to support urban military operations, specifically in support of unit coordination and room clearing tasks (Johnston, Hale, & Axelsson, 2010). Beyond communicating a command-based vocabulary, haptics can also be used to provide exteroceptive feedback, for example, by presenting tactile cues to enhance situation awareness

or optimize human performance. It has been suggested that such a solution could more closely couple operators with unmanned aerial systems (Johnston et al., 2010). The future may bring volumetric haptic displays, which project a touch-based representation of a surface onto a 3D volumetric space and allow users to feel the projected surface with their hands (Acosta and Liu, 2007) through haptic rending techniques (Basdogan et al., 2008), tearables that allow users to experience the real sense of tearing paper (Maekawa et al., 2009), and other such interactive tactile solutions.

The "vestibular system can be exploited to create, prevent, or modify acceleration perceptions" in virtual environments (Lawson et al., 2002, p. 137). For example, by simulating acceleration cues, a person can be psychologically transported from his or her veridical location, such as sitting in a chair in front of a computer, to a simulated location, such as the cockpit of a moving airplane. While vestibular cues can be stimulated via many different techniques in VEs, three of the most promising methods are physical motion of the user (e.g., motion platforms), wide FOV visual displays that induce vection (e.g., an illusion of self-motion), and locomotion devices that induce illusions of self-motion without physical displacement of the user through space (e.g., walking in place, treadmills, pedaling, foot platforms) (Hettinger, 2002; Hollerbach, 2002; Lawson et al., 2002). Of these options, motion platforms are probably the most advanced. For example, Sterling et al. (2000) integrated a small motion-based platform with a VE designed for helicopter landing training and found it to be comparable to a high-cost, large-scale helicopter simulator in terms of training effectiveness. Motion platforms are generally characterized via their range of motion/degrees of freedom (DOF) and actuator type (Isdale, 2000). In terms of range of motion, motion platforms can move a person in many combinations of translational (e.g., surge-longitudinal motion, sway-lateral motion, heave-vertical motion), and rotational (e.g., roll, pitch, yaw) DOF. A single-DOF translational motion system might provide a vibration sensation via a "seat shaker." A common 6 DOF configuration is a hexapod, which consists of a frame with six or more extendable struts (actuators) connecting a fixed base to a movable platform. In terms of actuators, electrical actuators are quiet and relatively maintenance free; however, they are not very responsive and they cannot hold the same load as can hydraulic or pneumatic systems. Hydraulic and pneumatic systems are smoother, stronger, and more accurate; however, they require compressors, which may be noisy. Servos are expensive and difficult to program.

Olfaction could be added to VE systems to stimulate emotion or enhance recall (Basdogan and Loftin, 2008). There have been several efforts made to support advances in olfactory interaction (Gutierrez-Osuna, 2004; Jones et al., 2004; Washburn and Jones, 2004; Washburn et al., 2003). One example of an olfactory system is the Scent Pallet (http://www.enviroscent.com/), which is a computer peripheral, universal serial bus (USB) device that uses up to eight scent cartridges, fans, and an air compressor to deliver different types

of scents. This system has been incorporated into the Full Spectrum Virtual Iraq/Afghanistan PTSD Therapy Application to provide the smell of rubber, cordite, garbage, body odor, smoke, diesel fuel, gunpowder, and other scents of the battlefield (S. Rizzo et al., 2006). These scents can be used as direct stimuli (e.g., scent of burning rubber) or as general cues to increase immersion (e.g., ethnic food cooking). The Scent Pallet was used to present vanilla, pizza, coffee, whiskey, beer, brandy, tequila, gin, scotch, red wine, white wine, cigarette smoke, and pine tree scents in an alcohol cue reactivity assessment system, which was found to be highly effective in stimulating subjective alcohol cravings (Bordnick et al., 2008). While several have mentioned the incorporation of gustatory stimulation, there are currently no functioning systems (Basdogan and Loftin, 2008).

### 2.1.2 Tracking Systems

Tracking systems allow determination of users' head or limb position and orientation or the location of handheld devices in order to allow interaction with virtual objects and traversal through 3D computer-generated worlds (Foxlin, 2002). Tracking is what allows the visual scene in a VE to coincide with a user's point of view, thereby providing an egocentric real-time perspective. Tracking systems must be carefully coupled with the visual scene, however, to avoid unacceptable lags (Kalawsky, 1993). Advances in tracking technology have been realized in terms of drift-corrected gyroscopic orientation trackers, outside-in optical tracking for motion capture, and laser scanners (Foxlin, 2002). The future of tracking technology is likely hybrid tracking systems (http://www.intersense.com/hybrid_technology .aspx), such as optical-inertial, GPS-inertial, magnetic-inertial, digital acoustic-inertial, and optical-magnetic hybrid solutions.

Tracking technology also allows for gesture recognition, in which human position and movement are tracked and interpreted to recognize semantically meaningful gestures (Turk, 2002). Gestures can be used to specify and control objects of interest, direct navigation, manipulate the environment, and issue meaningful commands. Gesture tracking devices that are worn (e.g., gloves, bodysuits) are currently more advanced than passive techniques (e.g., computer vision), yet the latter hold much promise for the future, as they can provide more natural, noncontact, and less obtrusive solutions than those that must be worn; limitations need to be overcome in terms of accuracy, processing speed, and generality (Erol et al., 2007).

### 2.1.3 Interaction Techniques

While one may think of joysticks and gloves when considering VE interaction devices, there are many techniques that can be used to support interaction with and traversal through a virtual environment. Interaction devices support traversal, pointing and selection of virtual objects, tool usage (e.g., through force and torque feedback), tactile interaction (e.g., through haptic devices), and environmental stimuli (e.g., temperature, humidity) (Bullinger et al., 2001).

Supporting traversal throughout a VE, via motion interfaces, is of primary importance (Hollerbach, 2002). Motion interfaces are categorized as either active (e.g., locomotion) or passive (e.g., transportation). Active-motion interfaces require self-propulsion to move about a virtual environment (e.g., treadmill, pedaling device, foot platforms). Passive-motion interfaces transport users within a VE without significant user exertion (e.g., inertial motion, as in a flight simulator, or noninertial motion, such as in the use of a joystick or gloves). The utility, functionality, cost, and safety of locomotion interfaces beyond traditional options (e.g., joysticks) have yet to be proven. In addition, beyond physical training, concrete applications for active-motion interfaces have yet to be clearly delineated. There are, however, some example applications, such as Arch-Explore, which is a real walking user interface that adapts redirected walking to allow exploration of large-scale virtual models of architectural scenes in a room-sized virtual environment (Bruder et al., 2009).

Another interaction option is speech control (see Chapters 8 and 35). Continuous speech recognition systems are currently under development, such as Parakeet (Vertanen and Kristensson, 2009), PocketSphinx (Huggins-Daines et al., 2006), and PocketSUMMIT (Hetherington, 2007). For these systems to provide effective interaction, however, additional advances are needed in acoustic and language-modeling algorithms to improve the accuracy, usability, and efficiency of spoken language understanding; such systems are still a ways away from offering conversational speech.

To support natural and intuitive interaction, a variety of interaction techniques can be coupled. For example, combining speech interaction with nonverbal gestures and motion interfaces can provide a means of interaction that closely captures real-world communications.

### 2.1.4 Augmented Cognition Techniques

Augmented cognition is an emerging computing paradigm in which users and computers are tightly coupled via physiological gauges that measure the cognitive state of users and adapt interaction to optimize human performance (Stanney et al., 2009). If incorporated into VE applications, augmented cognition could provide a means of evaluating their validity and compelling nature. For example, neuroscience studies have established that differential aspects of the brain are engaged when learning different types of materials and the areas in the brain that are activated change with increasing competence (Carroll et al., 2010a; Kennedy et al., 2005). Thus, if VE users were immersed in an educational experience, augmented cognition technology could be used to gauge if targeted areas of the brain were being activated and dynamically modify the content of a VE learning curriculum if desired activation patterns were not being generated. Physiological measures could also be used to detect the onset of cybersickness (see Section 4.1) and to assess the engagement, awareness, and anxiety of VE users, thereby potentially providing much more robust measures of immersion and presence (see Section 5.2). Such techniques could prove invaluable to entertainment VE applications (cf. Badiqué et al., 2002)

that seek to provide the ultimate experience, military training VE applications (cf. Knerr et al., 2002) that seek to emulate the "violence of action" found during combat, medical training applications (Wiecha et al., 2010) that seek to enhance traditional lab-based and classroom training practices, and therapeutic VE applications (cf. North et al., 2002; Strickland et al., 1997) that seek to overcome disorders such as fear of heights or flying.

## 2.2 Software Requirements

Software development of VE systems has progressed tremendously, from proprietary and arcane systems to development kits that run on multiple platforms (e.g., general-purpose operating systems to workstations). Virtual environment system components have become modular and distributed, thereby allowing VE databases (e.g., editors used to design, build, and maintain virtual worlds) to run independently of visualizers and other multimodal interfaces via network links. Standard APIs (application program interfaces) (e.g., OpenGL, Open Inventor, Direct3D, Mesa3D) allow multimodal components to be hardware independent. Virtual environment programming languages are maturing, with APIs, libraries (OpenGL Performer), and scripting languages (e.g., JavaScript, Lua, Linden, Mono, Perl, Python, Ruby) allowing nonprogrammers to develop virtual worlds (Stanney and Zyda, 2002). Advances are also being made in modeling of autonomous agents and communication networks used to support multiuser virtual environments.

### 2.2.1 Modeling

A VE consists of a set of geometry, the spatial relationships between the geometry and the user, and the change in geometry invoked by user actions or the passage of time (Kessler, 2002). Generally, modeling starts with building the geometry components (e.g., graphical objects, sensors, viewpoints, animation sequences) (Kalawsky, 1993). These are often converted from computer-aided design (CAD) data. These components then get imported into the VE modeling environment and rendered when appropriate sensors are triggered. Color, surface textures, and behaviors are applied during rendering. Programmers control the events in a VE by writing task functions, which become associated with the imported components.

A number of 3D modeling languages and toolkits are available that provide intuitive interfaces and run on multiple platforms and renderers (e.g., 3D Studio Max, AC3D, ZBrush, modo 401, Nexus, AccuRender, 3d ACIS Modeler, Ashlar-Vellum's Argon/Xenon/Cobalt, Carrara, CINEMA 4D, DX Studio, EON Studio, solidThinking) (Ultimate 3D Links, 2010). In addition, there are scene management engines (e.g., OpenSceneGraph, NVIDIA's SceniX) and game engines (e.g., Real Virtuality) that allow programmers to work at a higher level, defining characteristics and behaviors for more holistic concepts (Karim et al., 2003; Menzies, 2002). There have also been advances in photorealistic rendering tools (e.g., EI Technology's Amorphium), which are

evolving toward full-featured physics-based global illumination rendering systems (e.g., RenderPark). Taken together, these advances in software modeling allow for the generation of complex and realistic VEs that can run on a variety of platforms, permitting access to VE applications by both small- and large-scale application development budgets.

### 2.2.2 Autonomous Agents

Autonomous agents are synthetic or virtual human entities that possess some degree of autonomy, social ability, reactivity, and proactiveness (Allbeck and Badler, 2002; also see Chapter 15). There are several types of agents (Serenko and Detlor, 2004), including user agents (i.e., assist users by interacting with them, knowing their preferences and interests, and acting on their behalf), service agents (i.e., seamlessly collaborate with different parts of a system and perform more general tasks in the background, unbeknownst to users), embedded agents (i.e., interact with user and system to hide task complexity and make the overall user experience more exciting and enjoyable), and stand-alone agents (i.e., employ leading edge technologies and lay down the foundation for new architectures, standards, and innovative formats of agent-based computing). Autonomous agents can have many forms (e.g., human, animal), which are rendered at various levels of detail and style, from cartoonish to physiologically accurate models, and the form of the agent has been found to influence behavior both during and post VE exposure (i.e., the Proteus effect, where people infer their expected behaviors and attitudes from observing the appearance of their avatar; Yee et al., 2009). Such agents are a key component of many VE applications involving interaction with other entities, such as adversaries, instructors, or partners (Stanney and Zyda, 2002). Considerable work is being done to enhance the believability of such agents. For example, Heylen et al. (2008) found that when humanlike eye gaze behavior was incorporated into agents, that users communicated with such agents more effectively, and of utmost importance, human performance was also found to be enhanced with the more lifelike agents. As our understanding of how best to design autonomous agents evolves, such principles will be important to incorporate into their design to enhance the overall engagement and effectiveness of virtual worlds.

There has been significant research and development in modeling embodied autonomous agents. As with object geometry, agents are generally modeled off-line and then rendered during real-time interaction. While the required level of detail varies, modeling of hair and skin adds realism to an agent's appearance (Allbeck and Badler, 2002). There are a few toolkits available to support agent development, with one of the most notable offered by Boston Dynamics, Inc. (BDI) (http://www.bostondynamics.com/bd_diguy.html/), a spin-off from the MIT Artificial Intelligence Laboratory. BDI's DI-Guy allows VE developers to quickly integrate humans into their VEs, providing artificial intelligence to the characters, thereby enabling agents to autonomously navigate and react to their changing environment. Another option is ArchVision's 3D Rich

Photorealistic Content (RPC) People (http://www.archvision.com/RPCPeople.cfm).

### 2.2.3 Networks

Distributed networks allow multiple users at diverse locations to interact within the same virtual environment. Improvements in communication networks are required to allow realization of such shared experiences in which users, objects, processes, and autonomous agents from diverse locations interactively collaborate (Durlach and Mavor, 1995). Yet the foundation for such collaboration has been built within Internet2 (http://www.internet2.edu/), a next-generation Internet Protocol (IP) that delivers production network services for research and education institutions. This optical network could meet the high-performance demands of VEs, as it allows user-based allocation of high-capacity data circuits over a fiber-optic network. In addition, the Large Scale Networking (LSN) Coordinating Group (http://www.nitrd.gov/subcommittee/lsn.aspx) aims to develop leading-edge networking technologies and services, including programs in network security, new network architectures, heterogeneous networking (optical, mobile wireless, sensornet, etc.), federation across networking domains, grid and collaboration networking tools and services, with a goal of assuring that the next generation of the Internet will be scalable, trustworthy, and flexible. There are additional novel network technologies, including IP multicasting (i.e. a routing technique to prioritize one-to-many communication over an IP infrastructure in a network), quality of service (i.e., resource reservation control mechanisms), and IPv6 [i.e., also called IPng (or IP Next Generation), is a next generation IP addressing system] that could support distributed VE applications, which can leverage the special capabilities (e.g., high bandwidth, low latency, low jitter) of these advancing network technologies to provide shared virtual worlds.

## 3 DESIGN AND IMPLEMENTATION STRATEGIES

While many conventional HCI techniques can be used to design and implement VE systems, there are unique cognitive, content, product liability, and usage protocol considerations that must be addressed (see Figure 2).

### 3.1 Cognitive Aspects

The fundamental objective of VE systems is to provide multimodal interaction or, when sensory modalities are missing, perceptual illusions that support human
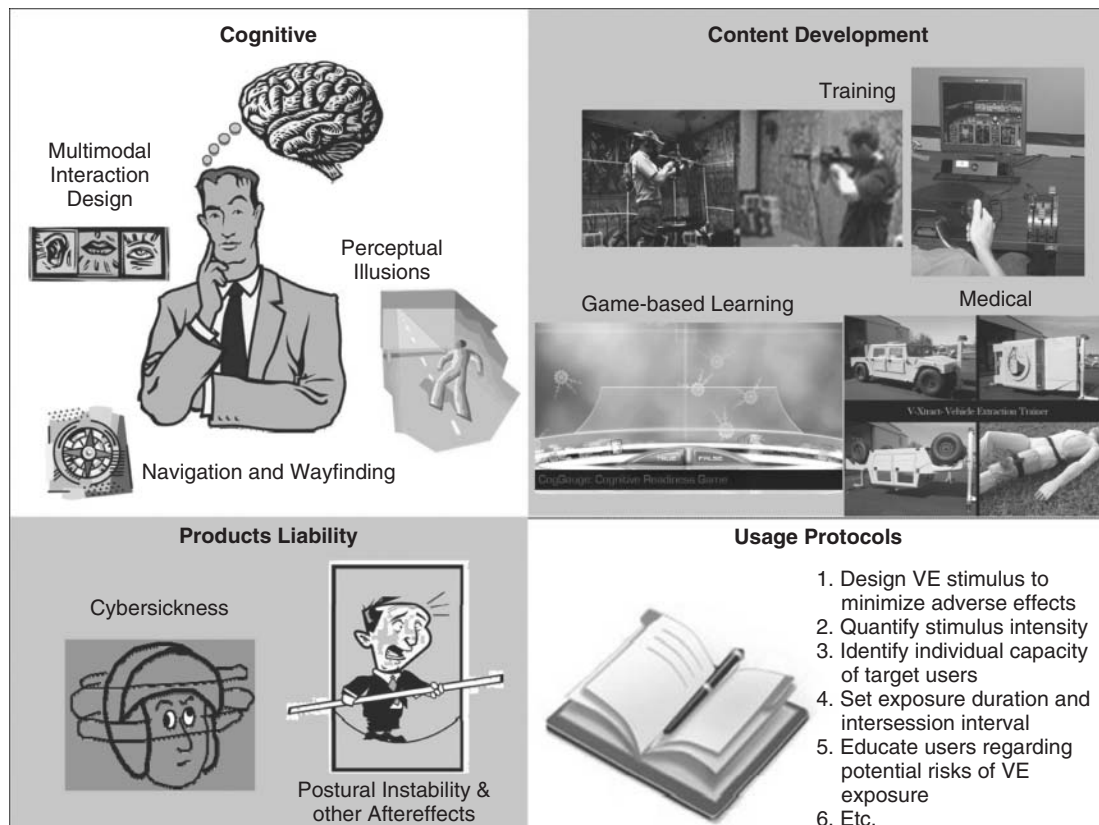


**Figure 2** VE design and implementation strategies.

information processing in pursuit of a VE application's goals, which could range from training to entertainment. Ancillary yet fundamental to this goal is to minimize cognitive obstacles, such as navigational difficulties, that could render a VE application's goals inaccessible.

### 3.1.1 Multimodal Interaction Design

Virtual environments are designed to provide users with immersive experiences that allow for direct manipulative and intuitive interaction with multisensory stimulation (Bullinger et al., 2001). The goals of providing this multimodal interaction within a VE are to achieve human–human communication/human–system interaction that is as natural as possible and to increase the robustness of this interaction by using redundant or complementary cues (Reeves et al., 2004). If designed effectively, engagement in such immersive multimodal VE experiences can lead to high levels of situation awareness and in turn high levels of human performance; however, the multimodal interaction within the VE must be appropriately designed to lead to this enhanced awareness. Specifically, the number of sensory modalities stimulated and the quality of this multisensory interaction are critical to the immersiveness and potential effectiveness of VE systems (Popescu et al., 2002). There are some emerging guidelines in the design of such multimodal interaction. For example, Stanney et al. (2004) provided a set of preliminary cross-modal integration rules. These rules consider aspects of multimodal interaction, including (a) temporal and spatial coincidence, (b) working memory capacity, (c) intersensory facilitation effects, (d) congruency, and (e) inverse effectiveness. When multimodal sensory information is provided to users, it is essential to consider such rules governing the integration of multiple sources of sensory feedback. VE users have adapted their perception–action systems to "expect" a particular type of information flow in the real world; VEs run the risk of breaking these perception–action couplings if the full range of sensory is not supported or if it is supported in a manner that is not contiguous with real-world expectations. Such pitfalls can be avoided through consideration of the coordination between sensing and user command and the transposition of senses in the feedback loop. Specifically, command coordination considers user input as primarily monomodal and feedback to the user as multimodal. Designers need to consider which input modalities are most appropriate to support execution of a given task within the VE, if there is any need for redundant user input, and whether or not users can effectively handle such parallel input (Stanney et al., 1998a, 2004). Additional multimodal design guidelines have been provided by Hale et al. (2009), who have outlined how a number of sensory cues may effectively be used to enhance specific situation awareness (SA) components (i.e., object recognition, spatial, temporal) within a VE, with the goal of optimizing SA development.

A limiting factor in supporting multimodal sensory stimulation in VEs is the current state of interface technologies. With the exception of the visual modality, current levels of technology simply cannot even begin to reproduce virtually those sensations, such as haptics and olfaction, which users expect in the real world. One solution to current technological shortcomings, sensorial transposition, occurs when a user receives feedback through senses other than those expected, which may occur because a command coordination scheme has substituted available sensory feedback for those that cannot be generated within a virtual environment. Sensorial substitution schemes may be one for one (e.g., visual for force) or more complex (e.g., visual for force and auditory; visual and auditory for force). If designed effectively, command coordination and sensory substitution schemes should provide multimodal interaction that allows for better user control of the virtual environment. On the other hand, if designed poorly, these solutions may in fact exacerbate interaction problems.

### 3.1.2 Perceptual Illusions

When sensorial transpositions are used, there is an opportunity for perceptual illusions to occur. With perceptual illusions, certain perceptual qualities perceived by one sensory system are influenced by another sensory system (e.g., "feel" a squeeze when you see your hand "grabbing" a virtual object). Such illusions could simplify and reduce the cost of VE development efforts (Storms, 2002). For example, when attending to a visual image coupled with a low-quality auditory display, auditory–visual cross-modal perception allows for an increase in the perceived quality of the visual image. Thus, in this case if the visual image is the focus of the task, there may be no need to use a high-quality auditory display.

There are several types of perceptual illusions that can be used in the design of virtual environments (Steinicke and Willemsen, 2010). Visual illusions can be used to substitute for missing proprioceptive and vestibular senses, as vision usually dominates these senses. For example, vection (i.e., a compelling illusion of self-motion throughout a virtual world) is known to be enhanced via a number of visual display factors, including a wide field of view and high spatial frequency content (Hettinger, 2002), as well as visual jitter (Kitazaki et al., 2010). In addition, change blindness (i.e., failing to notice alterations in a visual scene) can be used to apply subtle manipulations to the geometry of a VE and direct movement behavior, such as redirecting a user's walking path throughout a virtual environment (Suma et al., 2010). Other such illusions exist and could likewise be leveraged if perceptual and cognitive design principles are identified that can be used to trigger and capitalize on these illusory phenomena. For example, acoustic illusions (e.g., a fountain sound; Riecke et al., 2009) could also be used to create a sense of vection in a VE, even when no such visual motion is provided. In addition, haptic illusions (Hayward, 2008) could be used to provide users with the impression of actually feeling virtual objects when they are in fact touching real-world props or traveling along a trajectory path that may even vary in size, shape, weight, or surface from their virtual counterparts without users perceiving these discrepancies (e.g., feel an illusory bump when actually touching a flat surface [Robles-De-La-Torre and Hayward, 2001]; feel an illusory sharp edge when

hand actually travels along a smooth trajectory [Portillo-Rodríguez et al., 2006]).

### 3.1.3  Navigation and Wayfinding

Effective multimodal interaction design and use of perceptual illusions can be impeded if navigational complexities arise. Navigation is the aggregate of wayfinding (e.g., cognitive planning of one's route) and the physical movement that allows travel throughout a virtual environment (Darken and Peterson, 2002). A number of tools and techniques have been developed to aid wayfinding in virtual worlds, including maps, landmarks, trails, and direction finding. These tools can be used to display current position, current orientation (e.g., compass), log movements (e.g., "breadcrumb" trails), demonstrate or access the surround (e.g., maps, binoculars), or provide guided movement (e.g., signs, landmarks) (Chen and Stanney, 1999). For example, Burigat and Chittaro (2007) found 3D arrows to be particularly effective in guiding navigation throughout an abstract virtual environment. Darken and Peterson (2002) provided a number of principles concerning how best to use these tools. If effectively applied to VEs, these principles should lead to reduced disorientation and enhanced wayfinding in large-scale virtual environments.

### 3.2  Content Development

Content development is concerned with the design and construction of the virtual objects and synthetic environment that support a VE experience (Isdale et al., 2002). While this medium can leverage existing HCI design principles, it has unique design challenges that arise due to the demands of real-time, multimodal, collaborative interaction. In fact, content designers are just starting to appreciate and determine what it means to create a full sensory experience with user control of both point of view and narrative development. Aesthetics is thought to be a product of agency (e.g., pleasure of being), narrative potential, presence and co-presence (e.g., existing in and sharing the virtual experience), as well as transformation (e.g., assuming another persona) (Murray, 1997). Content development should be about stimulating perceptions (e.g., sureties, surprises) as well as contemplation over the nature of being (Isdale et al., 2002).

Existing design techniques, for example, from entertainment, video games, and theme parks, can be used to support VE content development (see Chapters 43 and 46). Game development techniques that can be leveraged in VE content development include but are not limited to providing a clear sense of purpose, emotional objectives, perceptual realism, intuitive interfaces, multiple solution paths, challenges, a balance of anxiety and reward, as well as an almost unconscious flow of interaction (Isdale et al., 2002). From theme park design, content development suggestions include (a) having a story that provides the all-encompassing theme of the VE and thus the "rules" that guide design, (b) providing location and purpose, (c) using cause and effect to lead users to their own conclusions, and (d) anchoring users in the familiar (Carson, 2000a, 2000b). While these suggestions provide guidelines for VE content development,

considerable creativity is still an essential component of the process.

While the content incorporated into the virtual worlds of today is mostly quite separate from the real world, in recent years life and technology have been more tightly coupled, the result being that computers are starting to have an awareness of themselves as well as the people who interact with them in 3D virtual spaces that are evolving into a "second life." Virtual worlds are in fact penetrating our native space and content development for future generations will likely aim to allow us to seamlessly use our own native language, with its wide range of verbal and physical gestures and emotions, thereby more fully entwining our first and second (virtual) lives (Rolston, 2010).

### 3.3  Product Liability

Those who implement VE systems must be cognizant of potential product liability concerns. Exposure to a VE system often produces unwanted side effects that could render users incapable of functioning effectively upon return to the real world. These adverse effects may include nausea and vomiting, postural instability, visual disturbances, and profound drowsiness (Stanney et al., 1998b). As users subsequently take on their normal routines, unaware of these lingering effects, their safety and well-being may be compromised. If a VE product occasions such problems, liability of VE developers or system administrators could range from simple accountability (e.g., reporting what happened) to full legal liability (e.g., paying compensation for damages) (Kennedy and Stanney, 1996; Kennedy et al., 2002). In order to minimize their liability, manufacturers and corporate users should design systems and provide usage protocols to minimize risks, warn users about potential aftereffects, monitor users during exposure, assess users' risk, and debrief users after exposure.

### 3.4  Usage Protocols

To minimize product liability concerns, VE usage protocols should be carefully designed. A comprehensive VE usage protocol will involve the following activities (see Stanney et al., 2005):

1. Designing VE stimulus to minimize adverse effects by minimizing lags and latencies, optimizing frame rates, and providing an adjustable interpupillary distance on visual display.

2. Quantifying stimulus intensity of a VE system using the Simulator Sickness Questionnaire (Kennedy et al., 1993) or other means and comparing the outcome to other systems (see Stanney et al., 2005). If a given VE system is of high intensity (say the 50th or higher percentile) and is not redesigned to lessen its impact, significant dropouts can be expected.

3. Identifying individual capacity of target user population to resist adverse effects of VE exposure via the Motion History Questionnaire (Kennedy and Graybiel, 1965) or other means.

4. Setting exposure duration and intersession interval to minimize adverse effects by limiting the duration of initial exposures, setting intersession exposure intervals two to five days apart, and moderating the stimulus intensity of virtual experiences (see Stanney et al., 2005).

5. Educating users regarding potential risks of VE exposure (e.g., inform users they may experience nausea, malaise, disorientation, headache, dizziness, vertigo, eyestrain, drowsiness, fatigue, pallor, sweating, increased salivation, and vomiting).

6. Educating users regarding potential adverse aftereffects of VE exposure (e.g., inform users they may experience disturbed visual functioning, visual flashbacks, and unstable locomotor and postural control for prolonged periods post-exposure).

7. Instructing users to terminate VE interaction if they start to feel ill.

8. Providing adequate air flow and comfortable thermal conditions.

9. Adjusting equipment to minimize fatigue.

10. For strong VE stimuli, warning users to avoid extraordinary maneuvers (e.g., flying backward or experiencing high rates of linear or rotational acceleration) during initial interaction.

11. Providing an attendant to monitor users' behavior and ensure their well-being.

12. Specifying amount of time post-exposure that users must remain on premises before driving or participating in other such high-risk activities. Do not allow individuals who fail post-exposure tests or experience adverse aftereffects to conduct high-risk activities until they have recovered (e.g., have someone else drive them home).

13. Calling users the next day or having them call to report any prolonged adverse effects.

Regardless of the strength of the stimulus or the susceptibility of the user, following a systematic usage protocol can minimize the adverse effects associated with VE exposure.

## 4 HEALTH AND SAFETY ISSUES

The health and safety risks associated with VE exposure complicate usage protocols and lead to product liability concerns. It is thus essential to understand these issues when utilizing VE technology. There are both physiological and psychological risks associated with VE exposure, the former being related primarily to sickness and aftereffects and the latter primarily being concerned with the social impact.

### 4.1 Cybersickness, Adaptation, and Aftereffects

Motion-sickness-like symptoms and other aftereffects (e.g., balance disturbances, visual stress, altered hand–eye coordination) are unwanted byproducts of VE exposure (Stanney and Kennedy, 2008). The sickness related to VE systems is commonly referred to as "cybersickness" (McCauley and Sharkey, 1992). Some of the most common symptoms exhibited include dizziness, drowsiness, headache, nausea, fatigue, and general malaise (Kennedy et al., 1993). More than 80% of users will experience some level of disturbance, with approximately 12% ceasing exposure prematurely due to this adversity (Stanney et al., 2003). Of those who drop out, approximately 10% can be expected to have an emetic response (e.g., vomit), however, only 1–2% of all users will have such a response. These adverse effects are known to increase in incidence and intensity with prolonged exposure duration (Kennedy et al., 2000). While most users will experience some level of adverse effects, symptoms vary substantially from one individual to another as well as from one system to another (Kennedy and Fowlkes, 1992). These effects can be assessed via the Simulator Sickness Questionnaire (Kennedy et al., 1993), with values above 20 requiring due caution (e.g., warn and observe users) (Stanney et al., 2005).

To overcome such adverse effects, individuals generally undergo physiological adaptation during VE exposure. This adaptation is the natural and automatic response to an intersensorily imperfect VE and is elicited due to the plasticity of the human nervous system (Welch, 1978). Due to technological flaws (e.g., slow update rate, sluggish trackers), users of VE systems may be confronted with one or more intersensory discordances (e.g., visual lag, a disparity between seen and felt limb position). In order to perform effectively in the VE, they must compensate for these discordances by adapting their psychomotor behavior or visual functioning. Once interaction with a VE is discontinued, these compensations persist for some time after exposure, leading to aftereffects.

Once VE exposure ceases and users return to their natural environment, they are likely unaware that interaction with the VE has potentially changed their ability to effectively interact with their normal physical environment (Stanney and Kennedy, 1998). Several different kinds of aftereffects may persist for prolonged periods following VE exposure (Welch, 1997). For example, hand–eye coordination can be degraded via perceptual–motor disturbances (Kennedy et al., 1997; Rolland et al., 1995), postural sway can arise (Kennedy and Stanney, 1996), as can changes in the vestibulo-ocular reflex (VOR) or one's ability to stabilize an image on the retina (Draper et al., 1997). The implications of these aftereffects are:

1. VE exposure duration may need to be minimized.

2. Highly susceptible individuals or those from clinical populations (e.g., those prone to seizures) may need to avoid or be banned from exposure.

3. Users should be closely monitored during VE exposure.

4. Users' activities should be closely monitored for a considerable period of time post-exposure to avoid personal injury or harm.

## 4.2 Social Impact

Virtual environment technology, like its ancestors (e.g., television, computers), has the potential for negative social implications through misuse and abuse (Kallman, 1993; also see Chapter 61). Yet violence in VE is nearly inevitable, as evidenced by the violent content of popular video games. Such animated violence is a known favorite over the portrayal of more benign emotions such as cooperation, friendship, or love (Sheridan, 1993; also see Chapter 4). The concern is that users who engage in what seems like harmless violence in the virtual world may become desensitized to violence and mimic this behavior in the look-alike real world.

Currently, it is not clear whether or not such violent behavior will result from VE exposure; early research, however, is not reassuring. Calvert and Tan (1994) found VE exposure to significantly increase the physiological arousal and aggressive thoughts of young adults. Perhaps more disconcerting was that neither aggressive thoughts nor hostile feelings were found to decrease due to VE exposure, thus providing no support for catharsis. Such increased negative stimulation may then subsequently be channeled into real-world activities. The ultimate concern is that VE immersion may potentially be a more powerful perceptual experience than past, less interactive technologies, thereby increasing the negative social impact of this technology (Calvert, 2002). A proactive approach is needed which weighs the risks and potential consequences associated with VE exposure against the benefits. Waiting for the onset of harmful social consequences should not be tolerated. Koltko-Rivera (2005) suggests that a proactive approach would involve determining (1) types and degree of VE content (e.g., aggressive, sexual), (2) types of individuals or groups exposed to this content (e.g., their mental aptitude, mental conditioning, personality, worldview), (3) circumstances of exposure (e.g., private experience, family, religion, spiritual), and (4) effects of exposure on psychological, interpersonal, or social function.

## 5 VIRTUAL ENVIRONMENT USABILITY ENGINEERING

Most VE user interfaces are fundamentally different from traditional graphical user interfaces, with unique I/O devices, perspectives, and physiological interactions. Thus, when developers and usability practitioners attempt to apply traditional usability engineering methods to the evaluation of VE systems, they find few if any that are particularly well suited to these environments (for notable exceptions see Gabbard et al., 1999; Hix and Gabbard, 2002; Stanney et al., 2000). There is a need to modify and optimize available techniques to meet the needs of VE usability evaluation as well as to better characterize factors unique to VE usability, including sense of presence.

## 5.1 Usability Techniques

Assessment of usability for VE systems must go beyond traditional approaches, which are concerned with the determination of effectiveness, efficiency, and user satisfaction (Bowman et al., 2002; see Chapter 55). Evaluators must consider whether multimodal input and output are optimally presented and integrated, navigation is supported to allow the VE to be readily traversed, object manipulation is intuitive and simple, content is immersive and engaging, and the system design optimizes comfort while minimizing sickness and aftereffects. The affective elements of interaction also become important when evaluating VE systems (see Chapter 58). It is an impressive task to ensure that all of these criteria are met.

Gabbard et al. (1999) have developed a taxonomy of VE usability characteristics that can serve as a foundation for identifying and evaluating usability criteria particularly relevant to VE systems. Stanney et al. (2000) used this taxonomy as the foundation on which to develop an automated system, MAUVE (Multicriteria Assessment of Usability for Virtual Environments), which assesses VE usability in terms of how effectively each of the following are designed: (a) navigation, (b) user movement, (c) object selection and manipulation, (d) visual output, (e) auditory output, (f) haptic output, (g) presence, (h) immersion, (i) comfort, (j) sickness, and (k) aftereffects. MAUVE can be used to support expert evaluations of VE systems, similar to the manner in which traditional heuristic evaluations are conducted. Due to such issues as cybersickness and aftereffects, it is essential to use these or other techniques (cf. Modified Concept Book Usability Evaluation Methodology; Swartz, 2003) to ensure the usability of VE systems, not only to avoid rendering them ineffective but also to ensure that they are not hazardous to users. Recently, guidelines have been evolving for enhancing the design of social VEs (e.g., Second Life by Linden Labs, Whyville by Numedeon, Inc.), such as those promoted by the Center for Disease Control and Prevention (CDC, 2010) for reaching individuals with timely health information that may relate to campaigns and upcoming events.

## 5.2 Sense Of Presence

A usability criterion unique to VE systems is sense of presence. Virtual environments have the unique advantage of leveraging the imaginative ability of individuals to psychologically "transport" themselves to another place, one that may not exist in reality (Sadowski and Stanney, 2002). To support such transportation, VEs provide physical separation from the real world by immersing users in the virtual world via, for example, an HMD, then imparting sensorial sensations via multimodal feedback that would naturally be present in the alternate environment. Focus on generating such presence is one of the primary characteristics distinguishing VEs from other means of displaying information.

Presence has been defined as the subjective perception of being immersed in and surrounded by a virtual world rather than the physical world one is currently

situated in (Stanney et al., 1998b). Virtual environments that engender a high degree of presence are thought to be more enjoyable, effective, and well received by users (Sadowski and Stanney, 2002). High-presence VEs are also suggested to be effective learning environments (Mantovani and Castelnuovo, 2003), as well as to enhance behavioral modeling outcomes and lead to greater imitation in the physical world (Fox et al., 2009b). Presence can be "broken" (i.e., lost) by external interference (e.g., people talking in the real-world during VE exposure), internal interference (e.g., daydreaming), inconsistent mediation (e.g., lag, distortions), contradictory mediation (e.g., when the virtual does not behave like the real), or unrefined mediation (e.g., information overload; Chertoff, Schatz, McDaniel, and Bowers, 2008). To enhance and maintain presence, designers of VE systems should spread detail around a scene, let user interaction determine when to reveal important aspects, maintain a natural and realistic, yet simple appearance, and utilize textures, colors, shapes, sounds, and other features to enhance realism (Kaur, 1999). To generate the feeling of immersion within the environment, designers should isolate users from the physical environment (use of an HMD may be sufficient), provide content that involves users in an enticing situation supported by an encompassing stimulus stream, provide natural modes of interaction and movement control, and utilize design features that enhance vection (Stanney et al., 2000). To enhance presence in learning environments, the design of perceptual features (i.e., perceptual realism, interactivity and control), individual factors (i.e., imagination and suspension of disbelief, identification, motivation and goals, emotional state), content characteristics (i.e., plot, story, narration, and dramaturgy), and interpersonal, social, and cultural context should be carefully considered (Mantovani and Castelnuovo, 2003). Presence can be assessed via Witmer and Singer's (1998) Presence Questionnaire or techniques used by Slater and Steed (2000) as well as a number of other means (Sadowski and Stanney, 2002).

## 6 APPLICATION DOMAINS

Virtual environments have been adopted by an ever-growing number of domains. Originally primarily used as a training platform, recent times have seen VEs in as diverse areas as operating rooms and courtrooms. These applications can provide adaptable, modest cost, deployable, and safe selection and training solutions, create game-based and learning virtual experiences that would otherwise be impossible to explore, and offer rehabilitation and medical applications that reach far beyond the conventional.

### 6.1 VE as a Selection and Training Tool

If one looks at training as a continuum across which a trainee matures in their declarative, procedural, and strategic knowledge, as well psychomotor skills and attitudes, then VE training is thought to be most suitable once the trainee has foundational declarative knowledge instantiated and some rudimentary procedural knowledge (Cohn et al., 2007). In general, van Merriënboer

and Kester (2005) recommend following the "fidelity principle," where learning is supported via a gradual increase in the fidelity of the training environment. Similarly, many training strategies reduce fidelity early in the training lifecycle to minimize complexity and avoid overloading the trainee (Regian et al., 1992). These suggestions are paralleled by stress-exposure training paradigms that suggest moving trainees through three stages, with an early focus on information provision and knowledge acquisition, followed by skills acquisition, and then culminating with practice of acquired skills under conditions that gradually approximate the stress environment (Driskell and Johnston, 1998). Taken together, these theories suggest the following (Cohn et al., 2007):

- Classroom lectures and low-fidelity training solutions (e.g., schematics, mock-ups) are most suitable for initial acquisition of declarative knowledge (i.e., general facts, principles, rules, and concepts) (Kelly et al., 1985; Rouse, 1982; 1991). VE training simulators would generally be less effective for such initial training, as they can be overly complex and confusing (Andrews, 1988; Boreham, 1985; Jones, 1990).

- Medium-fidelity VE training solutions, such as desktop VEs, are suggested to be suitable for training basic procedural knowledge and problem solving skills, and practice of such skills to mastery (Patrick, 1992; Pappo, 1998).

- High-fidelity VE training solutions, such as fully immersive VEs, can be used for consolidation of learned declarative knowledge and basic skills and procedures, practice of acquired knowledge and skills (e.g., mission rehearsal), as well as development of more advanced strategic knowledge and tactical skills (Forrest et al., 2002; Maran and Glavin, 2003; Vozenilek et al., 2004).

- High-fidelity VE training solutions, which are fully immersive and multisensory, may also be suitable for behavioral conditioning with stressors. Once basic knowledge and skills are mastered, attitudes and stress-induced behaviors are likely most appropriately trained in immersive and engaging solutions, which have the authenticity to generate realistic responses from trainees; yet there is limited research on this topic (Driskell and Johnston, 1998).

Beyond the ability of various forms of VEs to support several stages along the trainee continuum, they also offer the ability to immerse trainees in multiple contexts. This is important because learning is context specific (Anderson et al., 2000). By providing training in multiple context and from multiple points of view, VEs can be used in an effort to avoid the "reductive tendency" of learners to over-simplify new concepts, especially those gleaned in dynamic, highly-interactive environments, as well as the development of "knowledge-shields" erected to confirm simplified beliefs and understandings

(Feltovich et al., 2004). Consequently, while high fidelity VE training simulations oftentimes may be considered as the ultimate training solution, they are not suitable for all types of training and it is essential to determine the optimal level of fidelity that is required for a given training solution.

In terms of applying VE to enhance human performance, training is thus actually the second stage of a two-stage process. Ideally, one would like to select those individuals that have a certain degree of "performance capability" and are, in turn, ready for immersive VE training. Traditional approaches to selection focus on social and psychophysical assessments. For example, aptitude tests, ranging from traditional pen-and-paper-type to psychomotor tests to computer-based (but not VE!) assessments (Carretta and Ree, 1993; 1995), have all been used with varying levels of success. The single most important criticism of each of these approaches is that they are designed to be predictive of future performance and as such are more often than not abstractions of aspects of the larger task(s) for which the individual is being selected. An alternate approach would be to provide selectees with a method that provides a direct indication of their performance abilities. This distinction, essentially between a test being predictive of performance ability versus indicative of performance ability, has a great impact on selection. A meta-analysis performed within the aviation domain, where much of selection research has focused, found that typical predictive validities (most often reported as either the correlation coefficient $r$ or the multiple correlation coefficient $R$ and representing the degree to which given predictor/set of predictors and performance metrics are related) for such assessments range from a low of 0.14 to a "high" of about 0.40 (Martinussen, 1996). Yet, when a virtual simulation component is added to this mix, these values have been shown to improve considerably, pushing correlations towards the 0.60 level (Gress and Willkomm, 1996). This suggests that VE systems should be used as part of a comprehensive performance enhancement program that focuses on selecting those users with the correct set of knowledge, skills, and abilities (KSAs) and then providing, when needed, training to fine tune those KSAs. One approach to accomplishing this goal would be to assess trainee readiness by immersing them in virtual game-based cognitive assessment tools. For example, CogGauge immerses trainees in a mock spaceship cockpit in which the trainees must perform cognitive tasks at various celestial bodies and in so doing gain rewards that can culminate in the creation of a space station (Carpenter et al., 2010; Johnston, Carpenter, and Hale, 2011). The engaging nature of CogGauge serves to motivate trainees, making the KSA assessment process more interesting and engaging. Such immersive assessment tools can be used to determine where in the progression from novice to expert a given trainee is and then provide the most suitable form of training.

The suitability of virtual training solutions has been explored in a wide variety of areas, such as perceptual and cognitive performance (Carroll et al., 2010b), decision making under stress (Carroll et al., 2010a; Hill et al., 2003), operational readiness (Barba et al.,

2006), and cross-cultural communication (Deaton et al., 2005; Stanney et al., 2010). Similarly, VE applications are being used as interactive tools for teaching medical students, nurses, and doctors knowledge and skills as varied as the basics of human anatomy, complicated surgical procedures, communication skills, decision-making skills, and location of medical equipment within critical care vehicles (Grantcharov et al., 2004; Johnsen et al., 2005; Jones et al., 2010; Segal and Fernandez, 2009; Fried et al., 2010; Hassinger et al., 2010). As VE technology has matured, the breadth of VE training applications has likewise grown (King, 2009).

Flight skills are often trained in simulators and virtual environments (e.g., Aerosim's Virtual Flight Deck™, Microsoft Flight Sim, RealFlight®). Such applications provide a good example for demonstrating a key advantage of immersive training, which is the breadth of performance data that can be collected to evaluate training effectiveness. For example, behavioral and neurophysiological measures can be assessed during VE training and used to assess a learner's perceptual and cognitive processes (Carroll et al., 2010b; 2010c). These data can include such things as measurement and synching of eye tracking and electroencephalography (EEG) data with behavioral metrics (e.g., the actions taken in the VE) to capture unobservable performance characteristics, such as learner cognitive state (i.e., workload, engagement, distraction) and perceptual performance (i.e., scan data) during VE training. These data can, in turn, be used to identify the root cause of performance breakdowns and present instructors/ learners with performance summaries, such as scan data heat maps and cognitive replays illustrating how perceptual and cognitive processes contributed to performance breakdowns. The analysis can go even further, spatially correlating eye tracking data with VE scenario specific objects (e.g., specific flight gauges) and then diagnosing such things as the appropriateness of attention allocation (e.g., is the pilot scanning relevant instruments at the correct time?), root cause of performance errors (e.g., is the error due to inadequate scanning, lack of detection of critical events, or inappropriate actions?), and issues with cognitive state (e.g., is the pilot disengaged, overloaded, or distracted?). Thus, interactive VE training solutions allow trainees to not only consolidate their knowledge and practice their skills but also to provide adaptive training based on individual learner performance. The granular level of performance data available to VE training applications can also support "precision" training of various aspects of performance, such as perceptual skills used in search and detection task (e.g., security baggage screening, imagery analysis, threat detection, medical diagnostics; Carroll et al. 2010b, 2010c; Hale et al., 2007). In such applications, trainees can be shown not only the 'right' way to search a scene within a VE, but also how their specific search techniques differed from an expert and demonstrate the types of strategies that would be most helpful in improving an individual trainee's search strategies.

One might ask if the benefits of VE selection and training are worthwhile given the level of effort necessary to develop such immersive applications. In

the U.S. Air Force the cost of a single individual student pilot failing to complete basic flight school can run to $100,000 (Siem et al., 1988). Student failure can be attributed to both inadequate selection techniques and deficient training techniques. Clearly, both selection and training play critical roles in producing effective pilots. The challenge is to develop a training program that ensures a smooth union between the two, a solution that identifies the best candidates and then provides the optimum training. VE training solutions can be used to address both of these aspects of training by using immersive cognitive assessment tools to ensure the trainee is indeed ready for the complexities presented by immersive training and then providing detailed performance diagnostics that can realize "precision" training solutions that are uniquely tailored to a given trainee's performance deficiencies.

While VE applications may prove effective in selecting training candidates and providing them with training that is tailored to their given capabilities, the bottom-line for assessing the value of VE selection and training likely lies in how transferrable the skills from such training are to their target domain. A constant thread in training research is the notion that, in order for training to be effective, the basic skills being taught must show some degree of transfer to real-world performance. Over 100 years ago, Thorndike and Woodworth (1901) laid down the most basic training transfer principle when they proposed that transfer was determined by the degree of similarity between any two tasks. Applying this heuristic to VE design, one might conclude that the most basic way to ensure perfect transfer is to ensure that the real-world performance elements that are meant to be trained should be replicated perfectly in a virtual environment. This notion of "identical elements" could easily create a serious challenge for system designers even by today's technology standards, as VEs are still not able to perfectly duplicate the wide range of sensorial stimuli encountered during daily interactions with our world (Stoffregen et al., 2003). Countering this somewhat simplistic design approach is Osgood's (1949) principle that greater similarity between any two tasks along certain dimensions will not guarantee wholesale, perfect transfer. The challenge, as noted by Roscoe (1982), is to find the right balance between technical fidelity and training effectiveness. These issues are explored in detail by Stanney and Cohn (2012).

## 6.2 VE as an Entertainment and Education Tool

Virtual environments have reached beyond their original applications, primarily as military training tools, and have extended into a wide variety of entertainment applications. From interactive arcades to cybercafes, the entertainment industry has leveraged the unique characteristics of the VE medium, providing dynamic and exciting experiences in a multitude of forms. Virtual environment entertainment applications have found their way into games, sports, movies, art, online communities, location-based entertainment, theme parks, and other venues (Badiqué et al., 2002; Nakatsu et al., 2005). By

exploiting the unique interactive characteristics of VEs compared to more traditional entertainment media (e.g., film, play), VE technology provides a more immersive medium for entertainment through the use of simple artificial virtual characters (i.e., avatars), engaging narrative, and dynamic control to create an immersive interactive experience.

There are many forms of virtual entertainment, including:

- **Video games**. Immersive video games have become omnipresent (Chatfield, 2010). Generally, these games require their users to formulate hypotheses, learn game rules via trial-and-error, multitask, interactively develop strategies, and dynamically solve problems. These are skills that have life-relevance and thus the use of video games for edutainment has been widely considered. In fact, serious games, which are video games aimed at learning and other productive endeavors, are being taken much more *seriously* in recent years (Gunter, Kenny, and Vick, 2008). So, while some lambaste the amount of time young people are engaged in video games and suggest they are living in a media-saturated world (Rideout, Foehr, and Roberts, 2010), others are focused on leveraging this intense interest in productive ways, with one of the primary focuses being the use of interactive games in education (Aldrich, 2009; Squire, 2005). Some even suggest that such games can be used to get over the prepubescent literacy slump that leads to educational failures (Glee, 2008).

- **Computer role-playing games**. Computer role-playing video games (CRPGs; e.g., *Dungeons & Dragons*, *Ultima Underworld*, *Might and Magic*, *The Elder Scrolls*, *Diablo*), involve players in controlling one or more characters (i.e., a "party") as they seek to fulfill a series of quests (Barton, 2007). They involve fantasy, story-telling, and narrative progression, as well as evolving player character development (e.g., health, dexterity, strength). In 3D CRPGs, players typically navigate the game world from a first or third-person perspective. As with video games, CRPGs have been adapted to education purposes such as teaching literacy skills (Adams, 2009) and helping students craft interactive short stories (Carbonaro et al., 2008).

- **Massively multiplayer online role-playing games (MMORPGs)**. MMORPGs (e.g., *EverQuest*, *Meridian 59*, *Ultima Online*, *Final Fantasy*, *World of Warcraft*) are a genre of CRPGs that involve a very large number of players interacting with one another within a virtual game world (Bartle, 2003). They are similar to GRPGs in their makeup but are differentiated by the volume of players involved and the persistence of the virtual world, which evolves continuously, even when players are offline. The psychology of these games has become a topic of interest for academic researcher, with players being

classified into various psychological groups (i.e., achievers, explorers, socializers, killers; Bartle, 2003) and categories of motivation (i.e., immersion, cooperation, achievement, competition; Radoff, 2011). MMORPGs are starting to be used for educational applications, such as teaching science and English (Eustace et al., 2004), as well as supporting cooperative learning activities and exploring research questions (Childress and Braswell, 2006).

- **Massively multiplayer virtual worlds (MMVWs)**. Massively multiplayer virtual worlds (e.g., *Second Life*, *Active World*, *Twinity*, *Smeet*) have been developed that lack the inbuilt narrative, goals/objectives, and rule-based structure of games and instead provide the opportunity to explore and engage with other residents and avatars of the world through socializing, participating in activities, exploring other lands, and creating and trading virtual property and services with one another (Guest, 2008). Such virtual worlds have been used as a platform for educational purposes, scientific research, and the arts, as well as for launching personal relationships that can even lead to marriage (Dickey, 2005; Hayes, 2006).
- **Alternate reality games (ARGs)**. ARGs (e.g., *Dreadnot*, *The Art of the Heist*, *I Love Bees*, *The Beast*) are virtual games that involve intense player involvement with a real-time story that evolves according to the types of actions participants take in the virtual world (Kim, Allen, and Lee, 2008). Players, which can form into Guilds – associations of players, interact directly with characters, which are controlled by game designers (as opposed to AI), to solve plot-based challenges and puzzles. The ARG community can reach beyond the virtual through such means as websites, email messages, faxes, and voicemail messages, with players working together to analyze the story and coordinate real-life, as well as online activities. ARGs have extended into interactive television (e.g., *The Fallen Alternate Reality*, *ReGenesis Extended Reality*). An intriguing extension is to serious ARGs that focus on real-world problem solving, such as *World Without Oil* that focused on solving the issue of a global oil shortage (Egner, 2009) and *Foldit*, an ARG that reframes nettlesome scientific challenges as a competitive multiplayer computer game, the latter of which amazingly led to a breakthrough in HIV research (Khatib et al., 2011).

The crossover from purely entertainment to edutainment and real-world problem solving suggests that virtual interactive games have the potential to harness the ingenuity of game players into a formidable force that can be directed toward educational purposes, as well as solving a wide range of scientific problems. The future of interactive games is thus most intriguing . . . one cannot help but wonder where this creative energy will be directed in the future.

## 6.3 VE as a Medical Tool

What makes virtual reality application development in the assessment, therapy, and rehabilitation sciences so distinctively important is that it represents more than a simple linear extension of existing computer technology for human use. Virtual reality offers the potential to create systematic human testing, training, and treatment environments that allow for the precise control of complex, immersive, dynamic 3D stimulus presentations, within which sophisticated interaction, behavioral tracking, and performance recording are possible (A. Rizzo et al., 2006, p. 36).

Much has been written about applications for VE within the medical arena (cf. Moline, 1995; Satava and Jones, 2002). While some of these applications represent unique approaches to harnessing the power of VE, many other applications, such as simulating actual medical procedures, reflect training applications and therefore will not be discussed anew here. One area of medical application for which VE is truly coming into its own is medical rehabilitation. In particular, two areas of rehabilitation, behavioral/cognitive and motor, show strong promise.

### 6.3.1 Behavioral/Cognitive Rehabilitation Applications

In terms of behavioral rehabilitation applications, VE applications have been gaining prominence in behavioral science research over the past several years. For example, VE cue reactivity programs have been successfully tested for feasibility in nicotine (Bordnick et al., 2005), opiates (Kuntze et al., 2001), and alcohol (Bordnick et al., 2008) dependent individuals, as well as those with eating disorders (Gutiérrez-Maldonado et al., 2006). VE applications have also shown promise in modifying exercise behavior (Fox and Bailenson, 2009) and retirement savings behavior (Ersner-Hershfield et al., 2008) and managing pain (Dahlquist et al., 2007; Gold et al., 2007; Hoffman et al., 2008). Perhaps the fastest growing application for VEs in behavioral rehabilitation is in the area of exposure therapy (Fox et al., 2009a; Gregg and Tarrier, 2007; Parsons and Rizzo, 2008; Powers and Emmelkamp, 2008). For example, VE applications have been used to treat acrophobia (the fear of heights; Coelho et al., 2006), agoraphobia (fear of open spaces; Botella et al., 2007), arachnophobia (fear of spiders; Cote and Bouchard, 2005), aviophobia (fear of flying; Rothbaum et al., 2000), combat-related posttraumatic stress disorder (Reger and Gahm, 2008), panic disorder (Botella et al., 2007), public speaking anxiety (Harris et al., 2002), and social phobia (Roy et al., 2003). The reason for this broad use of VE technology for exposure therapy is likely due to the ideal matching between VE's strengths (presenting evolving information with which users can interact in various ways) and such therapy's basic requirements (incremental exposure to the offending environment). Importantly, compared to previous treatment regimens, which often-times simply required patients to mentally revisit their fears, VEs offer a significantly more immersive experience. In fact, it is quite likely that many of VE's

shortcomings, such as poor visual resolution, inadequate physics modeling underlying environmental cues, and failure to fully capture the wide range of sensorial cues present in the real world, will be ignored by the patient, whose primary focus is on overcoming anxiety engendered by her or his specific phobias. On a practical level, VEs enable patients to virtually visit their therapist's office, where they can be provided an individually tailored multimodal treatment experience (Rothbaum et al., 1996; Emmelkamp et al., 2001; Anderson et al., 2003).

Beyond behavioral rehabilitation, VE applications are being developed for the study, assessment, and rehabilitation of various types of cognitive processes, such as perception, attention, and memory. For example, VE applications are being used as perceptual skills trainers, such as for elderly drivers who have degraded visual scanning behavior (Romoser and Fisher, 2009) and for rehabilitating stroke victims who suffer from unilateral spatial neglect, where an individual fails to perceive stimuli presented to the contralesional hemivisual field even though they are not "blind" to this area (Katz et al., 2005). In terms of attention, attention-deficit hyperactivity disorder (ADHD) is an example of a cognitive dysfunction that has been addressed via VE rehabilitation applications (Parsons et al., 2007). Brooks and Rose (2003) suggest that VE rehabilitation applications can be used both in terms of assessment of memory impairments and memory remediation (e.g., use of reorganization techniques), where it has been found to promote procedural learning of those with memory impairments; importantly, this learning has been found to transfer to improved real-world performance. Examples of memory remediation in VEs include its use to enhance the ability of stroke victims to remember to perform actions in the future (Brooks et al., 2004), as well as its use in enhancing the performance of an individual with age-related impairment in memory-related cognitive processes (Optale et al., 2001). VEs have also been shown to uncover subtle cognitive impairments that might otherwise go undetected (Tippett et al., 2009). In general, VE applications can provide precisely controlled means of assessing cognitive impairments that are not available using more traditional evaluation methods. Specifically, VEs can deliver an assessment environment, where controlled stimuli can be presented at varying degrees of perception/attention/memory challenge and level of deficit can be assessed. This level of experimental control allows for the development of both cognitive impairment assessment and rehabilitation applications that have a high level of specificity and ecological validity.

### 6.3.2 Motor Rehabilitation Applications

Many of VE's qualities that make it an ideal tool for providing medical training—such as tactile feedback and detailed visual information (Satava and Jones, 2002)—also make it an ideal candidate for supplementing motor rehabilitation treatment regimens for such conditions as stroke (Deutsch and Mirelman, 2007; Yeh et al., 2007), cerebral palsy (Bryanton et al., 2006), and

amblyopia (i.e., lazy-eye; Eastgate et al., 2006). Specifically, Fox et al. (2009a) suggest that VEs have three features that make them uniquely suited to facilitating motor rehabilitation: the ability to review one's physical behavior and interactively examine one's progress, see one's own avatar from a third-person perspective in real time, and safely re-create real environments that cannot otherwise be experienced (e.g., crossing a busy intersection). In determining how best to apply VE in physical rehabilitation treatment regimens, Holden (2005) suggested considering three practical areas in which VE is strongest: repetition, feedback, and motivation. All three elements are critical to both effective learning and regaining motor function. The application of VE, in each case, provides a powerful method for rehabilitation specialists to maximize the effect of a treatment regimen for a given session and, because they may reduce the time investment required by therapists (one can simply immerse the patient, initiate the treatment, and then allow the program to execute), to also expand the access of such treatments to a wider population.

Since VE is essentially computer based, patients can effectively have their attention drawn to a specific set of movement patterns they may need to make to regain function; conducting this in a "loop" provides unlimited ability to repeat a pattern while using additional visualization aids, such as a rendered cursor or "follow-me" types of cues, to force the patient into moving a particular way (cf. Chua et al., 2003). As well, it is a relatively simple matter to digitize movement information, store it, and then, based on comparisons to previously stored, desired, movement patterns, to provide additional feedback to assist the patient. In terms of motivation, treatment scenarios can be tailored to capture specific events that individual patients find most motivating: a baseball fan can practice her movement in a baseball-like scenario; a car enthusiast can do so in a driving environment.

There are certain caveats that must be considered when exploiting VE for rehabilitation purposes, most significantly the potentially rapid loss of motor adaptations following VE exposure. Lackner and DiZio (1994) demonstrated that certain basic patterns of sensorimotor recalibrations learned in a given physical environment can diminish within an hour, postexposure, although subsequent findings (DiZio and Lackner, 1995) suggest that there are certain transfer benefits that are longer lasting. Brashers-Krugg et al. (1996) provided additional evidence that sensorimotor recalibrations of the type likely to be required for rehabilitation have postexposure periods in excess of 4 h during which their effects can be extinguished. Most importantly, Cohn et al. (2000) demonstrated that such recalibrations, when learned in VE, have essentially no transfer to real-world conditions postexposure. Clearly, more research is needed to understand the conditions under which such transfer effects can be made most effective within the clinical setting.

This is just a small glimpse of the potential applications for VE technology, the limit of which is bound only by our imagination. Recently, they have even been suggested as having potential value as a simulation of experience to enhance courtroom practice (Leonetti and

Bailenson, 2010). It will be interesting to see the vast variety of future VE applications that evolve.

## 7 CONCLUSIONS

Virtual environments have made substantial advances over the past decade, both in terms of the hardware and software used to generate them, as well as the breadth of their application. Yet, despite significant revolutions in component technology, many of the challenges addressed by incipient systems, such as multimodal sensori-interaction, visual representation, and scenario generation, have yet to be fully resolved. At the same time, our understanding of the potential such tools have to offer has advanced considerably. No longer simple amusements, these powerful machines can provide educational value, assist in treating physical and cognitive maladies, and even help design better VE systems. As the uses for which VEs are ideally suited continue to be defined and refined, one can anticipate that current development challenges will be resolved, allowing for a greater reach and more beneficial impact from applications of VE technology.

## ACKNOWLEDGMENTS

## REFERENCES

Acosta, E., and Liu, A. (2007), "Real-Time Volumetric Haptic and Visual Burrhole Simulation," in *Proceedings of IEEE Virtual Reality Conference*, VR 2007, Charlotte, NC, March 10–14, IEEE Press, Los Alamitos, CA, pp. 247–250.

Adams, M. G. (2009), "Engaging 21st-Century Adolescents: Video Games in the Reading Classroom," *English Journal*, Vol. 98, No. 6, pp. 56–59.

Akamatsu, M. (1994, July), "Touch with a Mouse: A Mouse Type Interface Device with Tactile and Force Display," in *Proceedings of 3rd IEEE International Workshop on Robot and Human Communication*, Nagoya, Japan, July 18–20, pp. 140–144.

Aldrich, C. (2009), "Virtual Worlds, Simulations, and Games for Education: A Unifying View," *Innovate: Journal of Online Education*, Vol. 5, No. 5, available: http://www.innovateonline.info/index.php?view=article&id=727, accessed October 28, 2011.

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (2001, October), "The CIPIC HRTF Database," in *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electro-Acoustics*, New Paltz, NY, pp. 99–102.

Allbeck, J. M., and Badler, N. I. (2002), "Embodied Autonomous Agents," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 313–332.

Anderson, J. R., Greeno, J. G., Reder, L. M., & Simon, H. A. (2000), "Perspectives on Learning, Thinking, and Activity," *Educational Researcher*, Vol. 29, No. 4, pp. 11–13.

Anderson, P., Rothbaum, B. O., and Hodges, L. F. (2003), "Virtual Reality Exposure in the Treatment of Social Anxiety," *Cognitive and Behavioral Practice*, Vol. 10, No. 3, pp. 240–247.

Andrews, D. H., Carroll, L. A., and Bell, H. H. (1995), "The Future of Selective Fidelity in Training Devices", Educational Technology, Vol. 35, No. 6, pp. 32–36.

Axelsson, P., and Johnston, M. R. (2010), "Results from the Empirical CRE Context Relevant Testing of the System for Tactile Reception of Advanced Patterns (STRAP)," in *Proceedings of 54th Human Factors and Ergonomics Society Annual Meeting*, San Francisco CA, September 27–October 1, 2010.

Badiqué, E., Cavazza, M., Klinker, G., Mair, G., Sweeney, T., Thalmann, D., and Thalmann, N. M. (2002), "Entertainment Applications of Virtual Environments," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1143–1166.

Ballas, J. A., Brock, D., Stroup, J., and Fouad, H. (2001), "The Effect of Auditory Rendering on Perceived Movement: Loudspeaker Density and HRTF," in *Proceedings of the 2001 International Conference on Auditory Display*, Espoo, Finland, July 29–August 1, pp. 235–238.

Barba, C., Deaton, J. E., Santarelli, T., Knerr, B., Singer, M., and Belanich, J. (2006), "Virtual Environment Composable Training for Operational Readiness (VECTOR)," in *Proceedings of the Proceedings of the 25th Army Science Conference,* "*Transformational Army Science and Technology—Charting the Future of S&T for the Soldier*," Orlando, FL, November 27–30.

Bartle, R. (2003), *Designing Virtual Worlds*, New Riders, Berkeley, CA.

Barton, M. (2007), "The History of Computer Role-Playing Games Part III: The Platinum and Modern Ages (1994–2004)," *Gamasutra*, available: http://www.gamasutra.com/view/feature/1571/the_history_of_computer_.php, accessed October 31, 2011.

Basdogan, C., and Loftin, B., (2008), "Multimodal Display Systems: Haptic, Olfactory, Gustatory, and Vestibular," in *The Handbook of Virtual Environment Training: Understanding, Predicting and Implementing Effective Training Solutions for Accelerated and Experiential Learning*, Vol. 2, D. Schmorrow, J. Cohn, and D. Nicholson, Eds., Praeger Security International, Westport, CN, pp. 116–135.

Basdogan, C., Laycock, S. D., Day, A. M., Patoglu, V., and Gillespie, R. B. (2008), "3-DoF Haptic Rendering," in *Haptic Rendering*, M. C. Lin and M. Otaduy, Eds., A. K. Peters. Wellesley, MA, pp. 311–331.

Begault, D. (1994), *3-D Sound for Virtual Reality and Multimedia*, Academic, Boston.

Biggs, S. J., and Srinivasan, M. A. (2002), "Haptic Interfaces," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 93–116.

Bordnick, P. S., Traylor, A. C., Graap, K. M., Copp, H. L., and Brooks, J. (2005), "Virtual Reality Cue Reactivity Assessment: A Case Study in a Teen Smoker," *Applied Psychophysiology and Biofeedback*, Vol. 30, No. 3, pp. 187–193.

Bordnick, P. S., Traylor, A., Copp, H. L., Graap, K. M., Carter, B., Ferrer, M., and Walton, A. P. (2008), "Assessing Reactivity to Virtual Reality Alcohol Based Cues," *Addictive Behaviors*, Vol. 33, pp. 743–756.

Boreham, N. C. (1985), "Transfer of Training in the Generation of Diagnostic Hypotheses: The Effect of Lowering Fidelity of Simulation," *British Journal of Educational Psychology*, Vol. 55, pp. 213–223.

Botella, C., et al. (2007), "Virtual Reality Exposure in the Treatment of Panic Disorder and Agoraphobia: A Controlled Study," *Clinical Psychology and Psychotherapy*, Vol. 14, pp. 164–175.

Bowman, D., Gabbard, J., and Hix, D. A. (2002), "Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods," *Presence: Teleoperators and Virtual Environments*, Vol. 11, No. 4, pp. 404–424.

Brashers-Krug, T., Shadmehr, R., and Bizzi, E. (1996), "Consolidation in Human Motor Memory," *Nature*, Vol. 382, No. 6588, pp. 252–255.

Brooks, B. M., and Rose, F. D. (2003), "The Use of Virtual Reality in Memory Rehabilitation: Current Findings and Future Directions," *NeuroRehabilitation*, Vol. 18, No. 2, pp. 147–157.

Brooks, B. M., Rose, F. D., Potter, E. A., Jayawardena, S., and Morling, A. (2004), "Assessing Stroke Patients' Prospective Memory Using Virtual Reality," *Brain Injury*, Vol. 18, No. 4, pp. 391–401.

Bruder, G., Steinicke, F., and Hinrichs, K. H. (2009), "Arch-Explore: A Natural User Interface for Immersive Architectural Walkthroughs," in *Proceedings of IEEE Symposium on 3D User Interfaces* (3 *DUI*), March 14–15, Lafayette, LA., K. Kiyokawa, Ed., IEEE Press, Los Alamitos, CA, pp. 75–82.

Bryanton, C., Bossé, J., Brien, M., McLean, J., McCormick, A., and Sveistrup, H. (2006), "Feasibility, Motivation, and Selective Motor Control: Virtual Reality Compared to Conventional Home Exercise in Children with Cerebral Palsy," *CyberPsychology and Behavior*, Vol. 9, pp. 123–128.

Bullinger, H.-J., Breining, R., and Braun, M. (2001), "Virtual Reality for Industrial Engineering: Applications for Immersive Virtual Environments," in *Handbook of Industrial Engineering: Technology and Operations Management*, 3rd ed., G. Salvendy, Ed., Wiley, New York, pp. 2496–2520.

Bungert, C. (2007), "HMD/Headset/VR-Helmet Comparison Chart," available: http://www.stereo3d.com/hmd.htm, accessed May 22, 2010.

Burdea, G. C., and Coiffet, P. (2003). *Virtual Reality Technology*, 2nd ed., John Wiley & Sons, Hoboken, NJ.

Burigat, S., and Chittaro, L. (2007), "Navigation in 3D Virtual Environments: Effects of User Experience and Location-Pointing Navigation Aids," *International Journal of Human-Computer Studies*, Vol. 65, No. 11, pp. 945–958.

Butler, R. A. (1987), "An Analysis of the Monaural Displacement of Sound in Space," *Perception and Psychophysics*, Vol. 41, pp. 1–7.

Calvert, S. L. (2002), "The Social Impact of Virtual Environment Technology," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 663–680.

Calvert, S. L., and Tan, S. L. (1994), "Impact of Virtual Reality on Young Adult's Physiological Arousal and Aggressive Thoughts: Interaction Versus Observation," *Journal of Applied Developmental Psychology*, Vol. 15, pp. 125–139.

Carbonaro, M. et al. (2008), "Interactive Story Authoring: A Viable Form of Creative Expression for the Classroom," *Computers and Education*, Vol. 51, No. 2, pp. 687–707.

Carpenter, A., Johnston, M., Kokini, C, and Hale, K. (2010), "CogGauge: A Game-Based Cognitive Assessment Tool," Proceedings of the *International Conference on Human-Computer Interaction in Aerospace HCI-Aero 2010* (Cape Canaveral, 3–5 November 2010), ACM Press, New York.

Carretta, T. R., and Ree, M. J. (1993), "Basic Attributes Test (BAT): Psychometric Equating of a Computer-Based Test," *International Journal of Aviation Psychology*, Vol. 3, pp. 189–201.

Carretta, T. R., and Ree, M. J. (1995), "Air Force Officer Qualifying Test Validity for Predicting Pilot Training Performance," *Journal of Business and Psychology,* Vol. 9, pp. 379–388.

Carroll, M., Fuchs, S., Carpenter, A., Hale, K., Abbott, R. G., and Bolton, A. (2010a), "Development of an Autodiagnostic Adaptive Precision Trainer for Decision Making (ADAPT-DM)," *International Test and Evaluation Journal*, Vol. 31, No. 2, pp. 247–263.

Carroll, M., Fuchs, S., Hale, K., Dargue, B., and Buck, B. (2010b), "Advanced Training Evaluation System: Leveraging Neuro-Physiological Measurement to Individualize Training," Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) Annual Meeting, Orlando, FL.

Carroll, M. B., Kokini, C., and Moss, J. (2010c), *Multi-Axis Performance Interpretation Tool (MAPIT) Field Training Effectiveness Evaluation*, Program Interim Report, Contract No. W911QY-07-C-0084, Office of Naval Research, Arlington, VA.

Carson, D. (2000a), "Environmental Storytelling, Part 1: Creating Immersive 3D Worlds Using Lessons Learned from the Theme Park Industry," available: http://www.gamasutra.com/features/20000301/carson_pfv.htm, accessed May 24, 2010.

Carson, D. (2000b), "Environmental Storytelling, Part 2: Bringing Theme Park Environment Design Techniques Lessons to the Virtual World," available: http://www.gamasutra.com/features/20000405/carson_pfv.htm, accessed May 24, 2010.

Centers for Disease Control and Prevention (CDC) (2010, January 11), "CDC Virtual World Requirements and Best Practices," available: http://www.cdc.gov/SocialMedia/Tools/guidelines/pdf/virtualworld.pdf, accessed June 1, 2010.

Chatfield, T., (2010), *Fun Inc.: Why Games are the 21st Century's Most Serious Business*, Virgin Books.

Chen, J. L., and Stanney, K. M. (1999), "A Theoretical Model of Wayfinding in Virtual Environments: Proposed

Strategies for Navigational Aiding," *Presence: Teleoperators and Virtual Environments*, Vol. 8, No. 6, pp. 671–685.

Chertoff, D. B., Schatz, S. L., McDaniel, R., and Bowers, C. A. (2008). "Improving Presence Theory through Experiential Design," *Presence: Teleoperators and Virtual Environments,*, Vol. 17, No. 4, pp. 405–413.

Childress, M. F., and Braswell, R. (2006), "Using Massively Multiplayer Online Role-Playing Games for Online Learning," *Distance Education,* Vol. 27, No. 2, pp. 187–196.

Chua, P. T., Crivella, R., Daly, B., Hu, N., Schaaf, R., Ventura, D., et al. (2003), "Training for Physical Tasks in Virtual Environments: Tai Chi," *Proceedings of IEEE Virtual Reality Conference 2003*, Los Angeles, CA, March 22–26, 2003, IEEE Press, Alamitos, CA, pp. 87–97.

Coelho, C. M., Santos, J. A., Silvério, J., and Silva, C. F. (2006), "Virtual Reality and Acrophobia: One-Year Follow-Up and Case Study," *CyberPsychology and Behavior*, Vol. 9, pp. 336–341.

Cohen, M. (1992), "Integrating Graphic and Audio Windows," *Presence: Teleoperators and Virtual Environments*, Vol. 1, No. 4, pp. 468–481.

Cohn, J., DiZio, P., & Lackner J. R. (2000), "Reaching During Virtual Rotation: Context-Specific Compensation for Expected Coriolis Forces," *Journal of Neurophysiology*, Vol. 83, No. 6, pp. 3230–3240.

Cohn, J. V., Stanney, K. M., Milham, L. M., Jones, D. L., Hale, K. S., Darken, R. P., and Sullivan, J. A. (2007), "Training Evaluation of Virtual Environments," in *Assessment of Problem Solving Using Simulations*, E. L. Baker, J. Dickieson, W. Wulfeck, and H. O'Neil, Eds., Lawrence Erlbaum, Mahwah, NJ, pp. 81–105.

Cote, S., and Bouchard, S. (2005), "Documenting the Efficacy of Virtual Reality Exposure with Psychophysiological and Information Processing Measures," *Applied Psychophysiology and Biofeedback*, Vol. 30, pp. 217–232.

Cruz-Neira, C., Sandin, D. J., and DeFanti, T. A. (1993), "Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE," *ACM Computer Graphics*, Vol. 27, No. 2, pp. 135–142.

Dahlquist, L. M., Weiss, K. E., Law, E. F., Soumitri, S., Herbert, L. J., Horn, S. B., Wohlheiter, K., and Ackerman C. S. (2010), "Effects of Videogame Distraction and a Virtual Reality Type Head-Mounted Display Helmet on Cold Pressor Pain in Young Elementary School-Aged Children," *Journal of Pediatric Psychology*, Vol. 35, No. 6, pp. 617–625.

Darken, R. P., and Peterson, B. (2002), "Spatial Orientation, Wayfinding, and Representation," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 493–518.

Deaton, J. E., et al. (2005), "Virtual Environment Cultural Training for Operational Readiness (VECTOR)," *Virtual Reality*, Vol. 8, pp. 156–167.

Deutsch, J. E., and Mirelman, A. (2007), "Virtual Reality-Based Approaches to Enable Walking for People Poststroke," *Topics in Stroke Rehabilitation*, Vol. 14, pp. 45–53.

Dickey, M. D. (2005), "Three-Dimensional Virtual Worlds and Distance Learning: Two Case Studies of Active Worlds as a Medium for Distance Education," *British Journal of Educational Technology* Vol. 36, No, 3, pp. 439–451.

DiZio, P., and Lackner, J. R. (1995), "Motor Adaptation to Coriolis Force Perturbations of Reaching Movements: Endpoint but not Trajectory Adaptation Transfers to the Non-Exposed Arm," *Journal of Neurophysiology*, Vol. 74, No. 4, pp. 1787–1792.

Draper, M. H., Prothero, J. D., and Viirre, E. S. (1997), "Physiological Adaptations to Virtual Interfaces: Results of Initial Explorations," in *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, p. 1393.

Driskell, J. E., and Johnston, J. H. (1998), "Stress Exposure Training," in *Making Decisions under Stress: Implications for Individual and Team Training*, J. A. Cannon-Bowers & E. Salas, Eds., APA Press, Washington, DC, pp. 191–217.

Durlach, B. N. I., and Mavor, A. S. (1995), *Virtual Reality: Scientific and Technological Challenges*, National Academy Press, Washington, DC.

Eastgate, R. M., et al. (2006), "Modified Virtual Reality for Treatment of Amblyopia," *Eye*, Vol. 20, pp. 370–374.

Egner, J. (2009, May 29), "One Story with 1700 Authors," *Current Magazine*, Current.org. available: http://current .org/tech/tech0709itvsgame.shtml, accessed October 28, 2011.

Emmelkamp, P. M. G., Bruynzeel, M., Drost, L., and van der Mast, C. A. P. G., (2001), "Virtual Reality Treatment in Acrophobia: A Comparison with Exposure in Vivo," *Cyberpsychology & Behavior*, Vol. 4, No. 3, pp. 335–339.

Erol, A., Bebis, G., Nicolescu, M., Boyle, D., and Twombly, X. (2007), "Vision-Based Hand Pose Estimation: A Review," *Computer Vision and Image Understanding*, Vol. 108, pp. 52–73.

Ersner-Hershfield, H., Bailenson, J., and Carstensen, L. L. (2008), "Feeling More Connected to Your Future Self: Using Immersive Virtual Reality to Increase Retirement Saving," poster presented at the Association for Psychological Science Annual Convention, Chicago, IL.

Eustace, K., Lee, M., Fellows, G., Byteway, A., and Irving, L. (2004), "The Application of Massively Multiplayer Online Role Playing Games to Collaborative Learning and Teaching Practice in Schools," in *Beyond the Comfort Zone: Proceedings of the 21st ASCILITE Conference*, R. Atkinson, C. McBeath, D. Jonas-Dwyer, and R. Phillips, Eds., p. 263, Perth, 5–8 December, available: http://www.ascilite.org.au/conferences/perth04/procs/ eustace-poster.html, accessed October 21, 2011.

Favalora, G. E. (2005), "Volumetric 3D Displays and Application Infrastructure," *Computer*, Vol. 38, No. 8, pp. 37–44.

Feltovich, P. J., Eccles, D. W., and Hoffman, R. R. (2004), "Implications of the Reductive Tendency for the Redesign of Complex Sociotechnical Systems," Report, Institute for Human and Machine Cognition, to the Advanced Decision Architectures Collaborative Technology Alliance. Sponsored by the US Army Research Laboratory under cooperative agreement DAAD19-01-2-0009.

Forrest, F. C., Taylor, M. C., Postlethwaite, K. C., and Aspinall, R. (2002), "Use of a High Fidelity Simulator to Develop Testing of the Technical Performance of Novice Anaesthetists," *British Journal of Anaesthesia*, Vol. 88, No. 3, pp. 338–344.

Fouad, H. (2004), "Ambient Synthesis with Random Sound Fields," in *Audio Anecdotes: Tools, Tips, and Techniques for Digital Audio*, K. Greenebaum, Ed., A. K. Peters, Natick, MA.

Fouad, H., and Ballas, J. (2000, April), "An Extensible Toolkit for Creating Virtual Sonic Environments," paper presented at the International Conference on Auditory Displays, ICAD 2000, Atlanta, GA.

Fox, J., and Bailenson, J. N. (2009), "Virtual Self-Modeling: The Effects of Vicarious Reinforcement and Identification on Exercise Behaviors," *Media Psychology*, Vol. 12, pp. 1–25.

Fox, J., Arena, D., and Bailenson, J. N. (2009a), "Virtual Reality: A Survival Guide for the Social Scientist," *Journal of Media Psychology*, Vol. 21, No. 3, pp. 95–113.

Fox, J., Bailenson, J. N., and Binney, J. (2009b), "Virtual Experiences, Physical Behaviors: The Effect of Presence on Imitation of an Eating Avatar," *Presence: Teleoperators and Virtual Environments*, Vol. 18, No. 4, pp. 294–303.

Foxlin, E. (2002), "Motion Tracking Requirements and Technologies," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 163–210.

Fried, M. P, et al. (2010), "From Virtual Reality to the Operating Room: The Endoscopic Sinus Surgery Simulator Experiment," *Otolaryngology—Head and Neck Surgery*, Vol. 142, No. 2, pp. 202–207.

Fuchs, S., Johnston, M., Hale, K. S., and Axelsson, P. (2008), "Results from Pilot Testing of a System for Tactile Reception of Advanced Patterns (STRAP)," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, New York, NY, September 22–26.

Gabbard, J. L., Hix, D., and Swan, J. E. II (1999), "User-Centered Design and Evaluation of Virtual Environments," *IEEE Computer Graphics and Applications*, Vol. 19, No. 6, pp. 51–59.

Gee, J., (2008), *Getting Over the Slump: Innovation Strategies to Promote Children's Learning*, Joan Ganz Cooney Center, New York, NY, available: http://www.joanganzcooneycenter.org/upload_kits/1_1.pdf, accessed October 21, 2011.

Gold, J. I., Belmont, K. A., and Thomas, D. A. (2007), "The Neurobiology of Virtual Reality Pain Attenuation," *Cyberpsychology & Behavior*, Vol. 10, No. 4, pp. 536–544.

Grantcharov, T. P., Kristiansen, V. B., Bendix, J., Bardram, L., Rosenberg, J., and Funch-Jensen, P. (2004), "Randomized Clinical Trial of Virtual Reality Simulation for Laparoscopic Skills Training," *British Journal of Surgery*, Vol. 91, No. 2, pp. 146–150.

Gregg, L., and Tarrier, N. (2007), "Virtual Reality in Mental Health: A Review of the Literature," *Social Psychiatry and Psychiatric Epidemiology*, Vol. 42, pp. 343–354.

Gress, W., and Willkomm, B. (1996), "Simulator-based Test Systems as a Measure to Improve the Prognostic Value of Aircrew Selection," *Selection and Training Advances in Aviation: Advisory Group for Aerospace Research and Development Conference Proceedings*, Prague, Czech Republic, Vol. 588, pp. 15-1–15-4.

Gross, M., Wurmlin, S., Naef, M., Lamboray, E., Spagno, C., Kunz, A., Koller-Meier, E., Svoboda, T., Gool, L. V., Lang, S., Strehlke, K., Moere, A.V., and Staadt, O. (2003), "Blue-C: A Spatially Immersive Display and 3D Video Portal for Telepresence," *ACM Transactions on Graphics*, Vol. 22, No. 3, pp. 819–827.

Guest, T. (2008), *Second Lives: A Journey through Virtual Worlds,* Random House, New York.

Gunter, G. A., Kenny, R. F., and Vick, E. H. (2008), "Taking Educational Games Seriously: Using the RETAIN Model to Design Endogenous Fantasy into Standalone Educational Games," *Educational Technology Research and Development*, Vol. 56 No. 5–6, pp. 511–537.

Gutierrez-Maldonado J, Ferrer-Garcia M, Caqueo-Urizar A, and Letosa-Porta A. (2006), "Assessment of Emotional Reactivity Produced by Exposure to Virtual Environments in Patients with Eating Disorders," *Cyberpsychology and Behavior*, Vol. 9, No. 5, pp. 507–513.

Gutierrez-Osuna, R. (2004), "Olfactory Interaction," in *Berkshire Encyclopedia of Human-Computer Interaction*, W. S. Bainbride, Ed., Berkshire Publishing, Great Barrington, MA, pp. 507–511.

Hale, K. S., Fuchs, S., Axelsson, P., Baskin, A., and Jones, D. (2007), "Determining Gaze Parameters to Guide EEG/ERP Evaluation of Imagery Analysis," in *Foundations of Augmented Cognition*, 4th ed., D. D. Schmorrow, D. M. Nicholson, J. M. Drexler, and L. M. Reeves, Eds., Strategic Analysis Inc., Arlington, VA, pp. 33–40.

Hale, K. S., and Stanney, K. M. (2004), "Deriving Haptic Design Guidelines from Human Physiological, Psychophysical, and Neurological Foundation," *IEEE Computer Graphics and Applications*, Vol. 24, No. 2, pp. 33–39.

Hale, K. S., Stanney, K. M., Milham, L. M., Bell-Carroll, M. A., and Jones, D. L. (2009), "Multimodal Sensory Information Requirements for Enhancing Situation Awareness and Training Effectiveness," *Theoretical Issues in Ergonomics Science*, Vol. 10, No. 3, pp. 245–266.

Harris, S. R., Kemmerling, R. L., and North, M. M. (2002), "Brief Virtual Reality Therapy for Public Speaking Anxiety," *CyberPsychology and Behavior*, Vol. 5, pp. 543–550.

Hassinger, J. P., Dozois, E. J., Holubar, J. D., Pawlina, W., Pendlimari, R., Fidler, J. L., Holmes, D. R., and Robb, R. A. (2010), "Virtual Pelvic Anatomy Simulator Improved Medical Student Comprehension of Pelvic Anatomy," *FASEB Journal*, Vol. 24, 825.3.

Hayes, E. R. (2006), "Situated Learning in Virtual Worlds: The Learning Ecology of Second Life," *Proceedings of the 47th annual Adult Education Research Conference*, University of Minnesota, Minneapolis, Minnesota, available: http://www.adulterc.org/Proceedings/2006/Proceedings/Hayes.pdf, accessed October 21 2011.

Hayward, V. (2008), "A Brief Taxonomy of Tactile Illusions and Demonstrations That Can Be Done in a Hardware Store," *Brain Research Bulletin*, Vol. 75, No. 6, pp. 742–752.

Hetherington, I. L. (2007), "PocketSUMMIT: Small Footprint Continuous Speech Recognition," in *Proceedings of International Conference on Spoken Language Processing (Interspeech 2007 - ICSLP)*, Antwerp, Belgium, August 27–31 IEEE Press, Los Alamitos, CA, pp. 1465–1468.

Hettinger, L. J. (2002), "Illusory Self-Motion in Virtual Environments," in *Handbook of Virtual Environments:*

*Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 471–492.

Heylen, D., Van Es, I., Nijholt, A., and Van Dijk, B. (2008), "Chapter 1: Controlling the Gaze of Conversational Agents," available: http://en.scientificcommons.org/43376669, accessed June 2, 2010.

Hill, R. W., Jr., Gratch, J., Marsella, S., Rickel, J., Swartout, W., and Traum, D. (2003), "Virtual Humans in the Mission Rehearsal Exercise System," *Künstliche Intelligenz*, Vol. 17, pp. 5–12.

Hix, D., and Gabbard, J. L. (2002), "Usability Engineering of Virtual Environments," in *Handbook of Virtual environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 681–699.

Hoffman, H. G., Patterson, D. R., Seibel, E., Soltani, M., Jewett-Leahy, L., and Sharar, S. R. (2008), "Virtual Reality Pain Control during Burn Wound Debridement in the Hydrotank," *Clinical Journal of Pain*, Vol. 24, No. 4, pp. 299–304.

Holden, M. (2005), "Virtual Environments for Motor Rehabilitation: A Review," *CyberPsychology & Behavior*, Vol. 8, No. 3, pp. 187–211.

Hollerbach, J. (2002), "Locomotion Interfaces," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 239–254.

Holmes, D. W., Williams, J. R., and Tilke, P. (2010), "An Events Based Algorithm for Distributing Concurrent Tasks on Multi-Core Architectures," *Computer Physics Communications*, Vol. 181, No. 2, pp. 341–354.

Huggins-Daines, D., Kumar, M., Chan, A., Black, A. W., Ravishankar, M., and Rudnicky, A. I. (2006), "PocketSphinx: A Free Real-Time Continuous Speech Recognition System for Hand-Held Devices," in *Proceedings of 31st International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006)*, Toulouse, France, May 14–19, IEEE Press, Los Alamitos, CA, pp. 185–188.

Isdale, J. (2000, April), "Motion Platform Systems," *VR News: April Tech Review*, available: http://vr.isdale.com/vrTechReviews/MotionLinks_2000.html, accessed May 21, 2010.

Isdale, J., Fencott, C., Heim, M., and Daly, L. (2002), "Content Design for Virtual Environments," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 519–532.

Jedrzejewski, M. (2004), "Computation of Room Acoustics on Programmable Video Hardware," Master's Thesis, Polish-Japanese Institute of Information Technology, Warsaw, Poland.

Johnsen, K., et al. (2005), "Experiences in Using Immersive Virtual Characters to Educate Medical Communication Skills," in *Proceedings of 2005 IEEE Conference on Virtual Reality*, IEEE CS Press, pp. 179–186.

Johnston, M. R., Carpenter, A. C., and Hale, K. (2011), "Test-Retest Reliability of CogGauge: A Cognitive Assessment Tool for Spaceflight," in *Engineering Psychology and Cognitive Ergonomics: Lecture Notes in Computer Science*, D. Harris, Ed., Springer, Heidelberg, Vol. 6781, pp. 565–571.

Johnston, M., Stanney, K., Hale, K., and Kennedy, R. S. (2010), "A Framework for Improving Situation Awareness of the UAS Operator through Integration of Tactile Cues," in *Proceedings of the 3rd Applied Human Factors and Ergonomics (AHFE) International Conference 2010*, Miami, FL, July 17–20.

Jones, A., McDowall, I., Yamada, H., Bolas, M., and Debevec, P. (2007, July), "Rendering an Interactive 360° Light Field Display," *ACM Transactions on Graphics*, Vol. 26, No. 3, Article 40.

Jones, D. L., Del Guidice, K., Moss, J., Hale, K. (2010), *Medic Vehicle Environment/Task Trainer (M-VETT): Phase I Process and Results*, Final Technical Report submitted to Office of the Secretary of Defense under contract W81XWH-10-C-0198.

Jones, D., Stanney, K., and Fouad, H. (2005), "An Optimized Spatial Audio System for Virtual Training Simulations: Design and Evaluation," in *Proceedings of the International Conference on Auditory Display*, Limerick, Ireland, July 6–9, 2005.

Jones, K. (1990), "General Activity for Management Education–A Deliberate Ambivalence," *Simulation/Games for Learning*, Vol. 20, No. 2, pp. 142–151.

Jones, L., Bowers, C. A., Washburn, D., Cortes, A., and Vijaya Satya, R. (2004), "The Effect of Olfaction on Immersion into Virtual Environments," in *Human Performance, Situation Awareness and Automation: Current Research and Trends*, Vol. 2, D. A. Vincenzi, M. Mouloua, and P. A. Hancock, Eds., Lawrence Erlbaum, Mahwah, NJ, pp. 282–285.

Kalawsky, R. S. (1993), *The Science of Virtual Reality and Virtual Environments*, Addison-Wesley, Wokingham, England.

Kallman, E. A. (1993), "Ethical Evaluation: A Necessary Element in Virtual Environment Research," *Presence: Teleoperators and Virtual Environments*, Vol. 2, No. 2, pp. 143–146.

Karim, M. S., Karim, A. M. L., Ahmed, E., and Rokonuzzaman, M. (2003), "Scene Graph Management for OpenGL Based 3D Graphics Engine," in *Proceedings of the International Conference on Computer and Information Technology (ICCIT 2003)*, Vol. 1, pp. 395–400.

Katz, N., Ring, H., Naveh, Y., Kizony, R., Feintuch, U., and Weiss, P. L. (2005), "Interactive Virtual Environment Training for Safe Street Crossing of Right Hemisphere Stroke Patients with Unilateral Spatial Neglect," *Disability and Rehabilitation*, Vol. 29, No. 2, pp. 177–181.

Kaur, K. (1999), "Designing Virtual Environments for Usability," unpublished doctoral dissertation, City University, London.

Kelley, A. I., Orgel, R. F., and Baer, D. M. (1985), "Seven Strategies That Guarantee Training Transfer," *Training and Development Journal*, Vol. 39, No. 11, pp. 78–82.

Kennedy, R. S., and Fowlkes, J. E. (1992), "Simulator Sickness Is Polygenic and Polysymptomatic: Implications for Research," *International Journal of Aviation Psychology*, Vol. 2, No. 1, pp. 23–38.

Kennedy, R. S., and Graybiel, A. (1965), "The Dial test: A Standardized Procedure for the Experimental Production of Canal Sickness Symptomatology in a Rotating Environment," Report No. 113, NSAM 930, Naval School of Aerospace Medicine, Pensacola, FL.

Kennedy, R. S., and Stanney, K. M. (1996), "Virtual Reality Systems and Products Liability," *Journal of Medicine and Virtual Reality*, Vol. 1, No. 2, pp. 60–64.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., and Lilienthal, M. G. (1993), "Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness," *International Journal of Aviation Psychology*, Vol. 3, No. 3, pp. 203–220.

Kennedy, R. S., Stanney, K. M., Ordy, J. M., and Dunlap, W. P. (1997), "Virtual Reality Effects Produced by Head-Mounted Display (HMD) on Human Eye-Hand Coordination, Postural Equilibrium, and Symptoms of Cybersickness," *Society for Neuroscience Abstracts*, Vol. 23, 772.

Kennedy, R. S., Stanney, K. M., and Dunlap, W. P. (2000), "Duration and Exposure to Virtual Environments: Sickness Curves during and across Sessions," *Presence: Teleoperators and Virtual Environments*, Vol. 9, No. 5, pp. 466–475.

Kennedy, R. S., Kennedy, K. E., and Bartlett, K. M. (2002), "Virtual Environments and Products Liability," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 543–554.

Kennedy, R. S., Drexler, J. M., Jones, M. B., Compton, D. E., and Ordy, J. M. (2005), "Quantifying Human Information Processing (QHIP): Can Practice Effects Alleviate Bottlenecks?," in *Quantifying Human Information Processing*, D. K. McBride and D. Schmorrow, Eds., Lexington Books, Lanham, MD, pp. 63–122.

Kessler, G. D. (2002), "Virtual Environment Models," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 255–276.

Khatib, F., et al. (2011), "Crystal Structure of a Monomeric Retroviral Protease Solved by Protein Folding Game Players," *Nature Structural & Molecular Biology*, Vol. 18, pp. 1175–1177.

King, M. (2009), *Military Simulation and Virtual Training Market 2009–2019*. Companiesandmarkets.com. London, England.

Kim, J. Y., Allen, J. P., and Lee, E. (2008), "Alternate Reality Gaming," *Communications of the ACM*, Vol. 51, No. 2, pp. 36–42.

Kitazaki, M., Onimaru, S., and Sato, T. (2010), "Vection and Action Are Incompatible," in *Proceedings of the 2nd IEEE VR 2010 Workshop on Perceptual Illusions in Virtual Environments (PIVE 2010)*, Waltham, MA, March 21, 2010, F. Steinicke and P. Willemsen, Eds., pp. 22–23, available: http://pive.uni-muenster.de/paper/PIVE_proceedings2010.pdf, accessed May 31, 2010.

Knerr, B. W., Breaux, R., Goldberg, S. L., and Thurman, R. A. (2002), "National Defense," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 857–872.

Koltko-Rivera, M. E. (2005), "The Potential Societal Impact of Virtual Reality," in *Advances in Virtual Environments Technology: Musings on Design, Evaluation, and Applications*, K. M. Stanney and M. Zyda, Eds., HCI International 2005: 11th International Conference on Human-Computer Interaction, CD-ROM, Vol. 9, Lawrence Erlbaum Associates, Mahwah, NJ.

Konrad, J., and Halle, M. (2007, November), "3-D Displays and Signal Processing—An Answer to 3-D Ills?," *IEEE Signal Processing Magazine*, Vol. 24, No. 6, pp. 97–111.

Kuntze, M. F., Stoermer, R., Mager, R., Roessler, A., Mueller-Spahn, F., and Bullinger, A. H. (2001), "Immersive Virtual Environments in Cue Exposure," *Cyberpsychology and Behavior*, Vol. 4, No. 4, pp. 497–501.

Lackner, J. R., and DiZio, P. (1994), "Rapid Adaptation to Coriolis Force Perturbations of Arm Trajectory," *Journal of Neurophysiology*, Vol. 72, No. 1, pp. 299–313.

Laughlin, D., Roper, M., and Howell, K. (2007, April), "NASA eEducation Roadmap: Research Challenges in the Design of Persistent Immersive Synthetic Environments for Education and Training," Federation of American Scientists, Washington, DC, available: http://www.fas.org/programs/ltp/publications/NASA%20eEducation%20Roadmap.pdf, accessed July 28, 2010.

Lawson, B. D., Sides, S. A., and Hickinbotham, K. A. (2002), "User Requirements for Perceiving Body Acceleration," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 135–161.

Leonetti, C., and Bailenson, J. N. (2010), "High-Tech View: The Use of Immersive Virtual Environments in Jury Trials," *Marquette Law Review*, Vol. 93, No. 3, pp. 1073–1120.

Maekawa, T., Itoh, Y., Takamoto, K., Tamada, K., Maeda, T., Kitamura, Y., and Kishino, F. (2009), "Tearable: Haptic Display That Presents a Sense of Tearing Real Paper. Virtual Reality Software and Technology," in *Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology*, Association for Computing Machinery, New York, pp. 27–30.

Majumder, A. (2003), "A Practical Framework to Achieve Perceptually Seamless Multi-Projector Displays Ph.D Dissertation, University of North Carolina at Chapel Hill, available: http://www.cs.unc.edu/~welch/media/pdf/dissertation_majumder.pdf, accessed May 24, 2010.

Mantovani, F., and Castelnuovo, G. (2003), "Sense of Presence in Virtual Training: Enhancing Skills Acquisition and Transfer of Knowledge through Learning Experience in Virtual Environments," in *Being There: Concepts, Effects and Measurement*, G. Riva, F. Davide, and W. A. IJsselsteijn, Eds., Ios Press, Amsterdam, The Netherlands, pp. 167–181.

Maran, N. J., and Glavin, R. J. (2003), "Low- to High- Fidelity Simulation–A Continuum of Medical Education?," *Medical Education*, Vol. 37, No. S1, pp. 22–28.

Martinussen, M. (1996), "Psychological Measures as Predictors of Pilot Performance: A Meta-analysis," *International Journal of Aviation Psychology*, Vol. 6, pp. 1–20.

Massimino, M., and Sheridan, T. (1993), "Sensory Substitution for Force Feedback in Teleoperation," *Presence: Teleoperators and Virtual Environments*, Vol. 2, No. 4, pp. 344–352.

McCauley, M. E., and Sharkey, T. J. (1992), "Cybersickness: Perception of Self-Motion in Virtual Environments," *Presence: Teleoperators and Virtual Environments*, Vol. 1, No. 3, pp. 311–318.

McCauley-Bell, P. R. (2002), "Ergonomics in Virtual Environments," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 807–826.

McQuaide, S. C., Seibel, E. J., Kelly, J. P., Schowengerdt, B. T., and Furness, T. A. III (2003), "A Retinal Scanning Display System That Produces Multiple Focal Planes with a Deformable Membrane Mirror," *Displays*, Vol. 24, No. 2, pp. 65–72

Melzer, J. E., and Rash, C. E. (2009), "The Potential of an Interactive HMD," in *Helmet-Mounted Displays:*

*Sensation, Perception, and Cognition Issues*, C. E. Rash, M. Russo, T. Letowski, and E. Schmeisser, Eds., U.S. Army Aeromedical Research Laboratory, Fort Rucker, AL, pp. 877–898.

Melzer, J. E., Brozoski, F. T., Letowski, T. R., Harding, T. H., and Rash, C. E. (2009), "Guidelines for HMD Design," in *Helmet-Mounted Displays: Sensation, Perception, and Cognition Issues*, C. E. Rash, M. Russo, T. Letowski, and E. Schmeisser, Eds., U.S. Army Aeromedical Research Laboratory, Fort Rucker, AL, pp. 805–848.

Menzies, D. (2002, July), "Scene Management for Modelled Audio Objects in Interactive Worlds," in *Proceedings of the 8th International Conference on Auditory Displays*, Kyoto, Japan.

Minamizawa, K., Kamuro, S., Kawakami, N., and Tachisuggest, S. (2008), "A Palm-Worn Haptic Display for Bimanual Operations in Virtual Environments," in *Euro-Haptics 2008*, M. Ferre, Ed., Springer-Verlag, Berlin, pp. 458–463.

Moline, J. (1995), *Virtual Environments for Health Care. White Paper for the Advanced Technology Program (ATP).* Retrieved October 20, 2011, from the National Institute of Standards and Technology website: http://www.itl.nist.gov/iaui/ovrt/projects/health/vr-envir.htm.

Mouba, J. (2009), "Performance of Source Spatialization and Source Localization Algorithms Using Conjoint Models of Interaural Level and Time Cues," in *Proceedings of the Twelfth International Conference on Digital Audio Effects (DAFx-09)*, Como, Italy, September 1–4.

Mulgund, S., Stokes, J., Turieo, M., and Devine, M. (2002), "Human/Machine Interface Modalities for Soldier Systems Technologies," Final Report No. 71950-00, TIAX, LLC, Cambridge, MA.

Murray, J. H. (1997), *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*, Free Press, New York.

Nakatsu, R., Rauterberg, M., and Vorderer, P. (2005), "A New Framework for Entertainment Computing: From Passive to Active Experience," in *Proceedings of the 4th International Conference on Entertainment Computing (ICEC 2005)*, Lecture Notes in Computer Science, Vol. 3711, F. Kishino, Y. Kitamura, H. Kato, and N. Nagata, Eds., Springer Verlag, Berlin and Heidelberg, pp. 1–12.

North, M. M., North, S. M., and Coble, J. R. (2002), "Virtual Reality Therapy: An Effective Treatment for Psychological Disorders," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1065–1078.

Optale, G., Capodieci, S., Pinelli, P., Zara, D., Gamberini, L., and Riva, G. (2001), "Music-Enhanced Immersive Virtual Reality in the Rehabilitation of Memory Related Cognitive Processes and Functional Abilities: A Case Report," *Presence: Teleoperators and Virtual Environments*, Vol. 10, No. 4, pp. 450–462.

Osgood, C. E. (1949), "The Similarity Paradox in Human Learning: A Resolution," *Psychological Review*, Vol. 56, pp. 132–143.

Pappo, H. A. (1998), *Simulations for Skills Training*, Educational Technology Publications, Englewood Cliffs, NJ.

Parsons, T. D., and Rizzo, A. A. (2008), "Affective Outcomes of Virtual Reality Exposure Therapy for Anxiety and Specific Phobias: A Meta-Analysis," *Journal of Behavior Therapy and Experimental Psychiatry*, Vol. 39, pp. 250–261.

Parsons, T. D., Bowerly, T., Buckwalter, J. G., and Rizzo, A. A. (2007), "A Controlled Clinical Comparison of Attention Performance in Children with ADHD in a Virtual Reality Classroom Compared to Standard Neuropsychological Methods," *Child Neuropsychology*, Vol. 13, No. 4, pp. 363–381.

Patrick, J. (1992), *Training: Research and Practice*, Academic Press, San Diego, CA.

Popescu, G. V., Burdea, G. C., and Trefftz, H. (2002), "Multimodal Inter-Action Modeling," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 435–454.

Portillo-Rodriguez, O., Avizzano, C. A., Bergamasco, M., and Robles-De-La-Torre, G. (2006), "Haptic Rendering of Sharp Objects Using Lateral Forces," in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (ROMAN06)*, Hatfield, United Kingdom, September 6–8, pp. 431–436.

Powers, M. B., and Emmelkamp, P. M. G. (2008), "Virtual Reality Exposure Therapy for Anxiety Disorders: A Meta-Analysis," *Journal of Anxiety Disorders*, Vol. 22, pp. 561–569.

Pulkki, V. (1997), "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *Journal of the Audio Engineering Society*, Vol. 45, No. 6, pp. 456–466.

Rauterberg, M. (2009), "Entertainment Computing, Social Transformation and the Quantum Field," in *Proceedings of Intelligent Technologies for Interactive Entertainment (INTETAIN 2009)*, Lecture Notes of the Institute for Computer Sciences, Vol. 9, A. Nijholt, D. Reidsma, and H. Hondorp, Eds., Springer Verlag, Berlin and Heidelberg, pp. 1–8.

Radoff, J. (2011), *Game On: Energize Your Business with Social Media Games*, Wiley Publishing, Indianapolis, IN.

Reeves, L. M., Lai, J. C., Larson, J. A., Oviatt, S. L., Balaji, T. S., Buisine, S., Collings, P., Cohen, P. R., Kraal, B., Martin, J. C., McTear, M. F., Raman, T. V., Stanney, K. M., Su, H., and Wang, Q. Y. (2004), "Guidelines for Multimodal User Interface Design," *Communications of the ACM*, Vol. 47, No. 1, pp. 57–59.

Reger, G. M., and Gahm, G. A. (2008), "Virtual Reality Exposure Therapy for Active Duty Soldiers," *Journal of Clinical Psychology*, Vol. 64, pp. 940–946.

Regian, J. W., Shebilske, W. L., and Monk, J. M. (1992), "Virtual Reality: An Instructional Medium for Visual-Spatial Tasks," *Journal of Communication*, Vol. 42, pp. 136–149.

Rideout, V. J., Foehr, U. G., and Roberts D. F. (2010), *Generation $M^2$: Media in the Lives of 8–18 Year-Olds*, Kaiser Family Foundation, Menlo Park, CA, available: http://www.kff.org/entmedia/upload/8010.pdf, accessed October 31, 2011.

Riecke, B. E., Väljamäe, A., and Schulte-Pelkum, J. (2009), "Moving Sounds Enhance the Visually-Induced Self-Motion Illusion (Circular Vection) in Virtual Reality," *ACM Transactions on Applied Perception*, Vol. 6, No. 2, Article 7, pp. 7:1–7:27.

Rizzo, A., Bowerly, T., Buckwalter, J., Klimchuk, D., Mitura, R., and Parsons, T. D. (2006), "A Virtual Reality Scenario for All Seasons: The Virtual Classroom," *CNS Spectrums*, Vol. 11, No. 1, pp. 35–44.

Rizzo, S., Rothbaum, B., and Graap, K. (2006), "Virtual Reality Applications for the Treatment of Combat-Related PTSD," in *Combat Stress Injury Theory, Research, and Management*, C. R. Figley and W. P. Nash, Eds., pp. 295–329, Routledge, New York, NY.

Robles-De-La-Torre, G., and Hayward, V. (2001), "Force Can Overcome Object Geometry in the Perception of Shape through Active Touch," *Nature*, Vol. 412, No. 6845, pp. 445–448.

Rolland, J. P., Biocca, F. A., Barlow, T., and Kancherla, A. (1995), "Quantification of Adaptation to Virtual-Eye Location in See-Thru Head-Mounted Displays," in *Proceedings of the IEEE Virtual Reality Annual International-Symposium '95*, IEEE Computer Society Press, Los Alimitos CA, pp. 56–66.

Rolston, M. (2010), "Your Computer in 2020," in Your Life in 2020, N. Perlroth, Ed., available: http://www.forbes.com/2010/04/08/3d-computers-2020-technology-data-companies-10-frog.html, accessed May 25, 2010.

Romoser, M. R. E., and Fisher, D. L. (2009), "The Effect of Active Versus Passive Training Strategies on Improving Older Driver's Scanning in Intersections," *Human Factors*, Vol. 51, No. 5, pp. 652–668.

Roscoe, S. N. (1982), *Aviation Psychology*. Iowa State University Press, Ames, IA.

Rothbaum, B. O., Hodges, L., Smith, S., Lee, J. H., and Price, L. (2000), "A Controlled Study of Virtual Reality Exposure Therapy for the Fear of Flying," *Journal of Consulting and Clinical Psychology*, Vol. 68, pp. 1020–1026.

Rothbaum, B. O., Hodges, L., Watson, B. A., Kessler, G. D., and Opdyke, D. (1996), "Virtual Reality Exposure Therapy in the Treatment of Fear of Flying: A Case Report," *Behaviour Research & Therapy*, Vol. 34, pp. 477–481.

Rouse, W. B. (1982), "A Mixed-Fidelity Approach to Technical Training," *Journal of Educational Technology Systems*, Vol. 11, pp. 103–115.

Roy, S., Klinger, E., Legeron, P., Lauer, F., Chemin, I., and Nugues, P. (2003), "Definition of a VR-Based Protocol to Treat Social Phobia," *CyberPsychology and Behavior*, Vol. 6, pp. 411–420.

Sadowski, W., and Stanney, K. (2002), "Presence in Virtual Environments," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 791–806.

Satava, R., and Jones, S. B. (2002), "Medical Applications of Virtual Environments," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Inc., Mahwah, NJ, pp. 93–116.

Seeber, B. U., and Fastl, H. (2003), "Subjective Selection of Non-Individual Head-Related Transfer Function," in *Proceeding of the 2003 International Conference on Auditory Display*, Boston University, Boston, MA, July 6–9, pp. 259–262.

Segal, D., and Fernandez, R. L. (2009, August), "Teaching Physician Decision Making in a Technical Age," *Virtual Mentor*, Vol. 11, No. 8, pp. 607–610.

Serenko, A., and Detlor, B. (2004), "Intelligent Agents as Innovations," *Artificial Intelligence and Society*, Vol. 18, No. 4, pp. 364–381.

Sheridan, T. B. (1993), "My Anxieties About Virtual Environments," *Presence: Teleoperators and Virtual Environments*, Vol. 2, No. 2, pp. 141–142.

Shilling, R. D., and Shinn-Cunningham, B. (2002), "Virtual Auditory Displays," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 65–92.

Siem, F. M., Carretta, T. R., and Mercatante, T. A. (1988), *Personality, Attitudes, and Pilot Training Performance: Preliminary Analysis,* Tech. Report. N. AFHRL-TP-87-62, Brooks Air Force Base, TX: AFHRL, Manpower and Personnel Division.

Slater, M., and Steed, A. (2000), "A Virtual Presence Counter," *Presence: Teleoperators and Virtual Environments*, Vol. 9, No. 5, pp. 413–434.

Spors, S., and Ahrens, J. (2010), "Analysis and Improvement of Pre-Equalization In 2.5-Dimensional Wave Field Synthesis," in *Proceedings of the 128th Audio Engineering Society (AES) Convention, P21—Multichannel and Spatial Audio: Part 2 Session (P21-3 )*, London, May 25.

Squire, K., (2005), "Changing the Game: What Happens When Video Games Enter the Classroom," *Innovate: Journal of Online Education*, Vol. 1, No. 6, available: http://innovateonline.info/pdf/vol1_issue6/Changing_the_Game_What_Happens_When_Video_Games_Enter_the_Classroom_.pdf, accessed October 28, 2011.

Stanney, K. M., and Cohn, J. V. (2012), "Virtual Environments," in *Handbook of Human–Computer Interaction (3rd Edition)*, J. Jacko, Ed., Taylor & Francis, Hampshire, United Kingdom.

Stanney, K. M., Graeber, D. A., and Kennedy, R. S. (2005), "Virtual Environment Usage Protocols," in *Handbook of Standards and Guidelines in Ergonomics and Human Factors*, W. Karwowski, Ed., Lawrence Erlbaum, Mahwah, NJ, pp. 381–398.

Stanney, K. M., and Kennedy, R. S. (1998, October), "Aftereffects from Virtual Environment Exposure: How Long Do They Last?," in *Proceedings of the 42nd Annual Human Factors and Ergonomics Society Meeting*, Chicago, IL, pp. 1476–1480.

Stanney, K. M., and Kennedy, R. (2008), "Simulator Sickness," in *Human Factors in Simulation and Training*, D. Vincenzi, J. A. Wise, M. Mouloua, and P. A. Hancock, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 117–128,

Stanney, K. M., Kingdon, K., Nahmens, I., and Kennedy, R. S. (2003), "What to Expect from Immersive Virtual Environment Exposure: Influences of Gender, Body Mass Index, and Past Experience," *Human Factors*, Vol. 45, No. 3, pp. 504–522.

Stanney, K., Kokini, C., Fuchs, S., Axelsson, P., and Phillips, C. (2010), "Auto-Diagnostic Adaptive Precision Training—Human Terrain (ADAPT-HT): A Conceptual Framework for Cross-Cultural Skills Training," in *Proceedings of the 3rd Applied Human Factors and Ergonomics (AHFE) International Conference 2010*, Miami, FL, July 17–20.

Stanney, K. M., Mollaghasemi, M., and Reeves, L. (2000), "Development of MAUVE, the Multi-Criteria Assessment of Usability for Virtual Environments System," Final Report, Contract No. N61339-99-C-0098, Naval Air Warfare Center, Training Systems Division, Orlando, FL.

Stanney, K. M., Mourant, R., and Kennedy, R. S. (1998a), "Human Factors Issues in Virtual Environments: A Review of the Literature," *Presence: Teleoperators and Virtual Environments*, Vol. 7, No. 4, pp. 327–351.

Stanney, K. M., et al. (1998b), "Aftereffects and Sense of Presence in Virtual Environments: Formulation of a Research and Development Agenda (Report sponsored by the Life Sciences Division at NASA Headquarters)," *International Journal of Human-Computer Interaction*, Vol. 10, No. 2, pp. 135–187.

Stanney, K. M., Schmorrow, D. D., Johnston, M., Fuchs, S., Jones, D., Hale, K., Ahmad, A., and Young, P. (2009), "Augmented Cognition: An Overview," in *Reviews of Human Factors and Ergonomics*, Vol. 5, F. T. Durso, Ed., Human Factors and Ergonomics Society, Santa Monica, CA, pp. 195–224.

Stanney, K., et al. (2004), "A Paradigm Shift in Interactive Computing: Deriving Multimodal Design Principles from Behavioral and Neurological Foundations," *International Journal of Human-Computer Interaction*, Vol. 17, No. 2, pp. 229–257.

Stanney, K. M., and Zyda, M. (2002), "Virtual Environments in the 21st Century," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1–14.

Steinicke, F., and Willemsen, P., Eds. (2010), *Proceedings of the 2nd IEEE VR 2010 Workshop on Perceptual Illusions in Virtual Environments* (*PIVE 2010*), Waltham, MA, March 21, 2010, available: http://pive.uni-muenster.de/paper/PIVE_proceedings2010.pdf, accessed May 31, 2010.

Sterling, G. C., Magee, L. E., and Wallace, P. (2000, March), "Virtual Reality Training—A Consideration for Australian Helicopter Training Needs?," paper presented at Simulation Technology and Training (SimTecT2000), Sydney Australia, March 2000.

Stoffregen, T., Bardy, B. G., Smart, L. J., and Pagulayan, R. (2003), "On the Nature and Evaluation of Fidelity in Virtual Environments," in *Virtual and Adaptive Environments: Applications, Implications, and Human Performance Issues*, L. J. Hettinger and M. W. Haas, Eds., Lawrence Erlbaum Associates, Publishers, Mahwah, NJ, pp. 111–128.

Storms, R. L. (2002), "Auditory-Visual Cross-Modality Interaction and Illusions," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 455–470.

Strickland, D., Hodges, L., North, M., and Weghorst, S. (1997), "Overcoming Phobias by Virtual Exposure," *Communications of the ACM*, Vol. 40, No. 8, pp. 34–39.

Sullivan, A. (2004), "DepthCube Solid-State 3D Volumetric Display," in *Proceedings of the SPIE Stereoscopic Displays and Virtual Reality Systems*, Vol. 5291, pp. 279–284.

Suma, E. A., Clark, S., Finkelstein, S. L., and Wartell, Z. (2010), "Leveraging Change Blindness for Walking in Virtual Environments," in *Proceedings of the 2nd IEEE VR 2010 Workshop on Perceptual Illusions in Virtual Environments* (*PIVE 2010*), Waltham, MA, March 21, 2010, F. Steinicke and P. Willemsen, Eds., p. 10, available: http://pive.uni-muenster.de/paper/PIVE_proceedings2010.pdf, accessed May 31, 2010.

Swartz, K. O. (2003), "Virtual Environment Usability Assessment Methods Based on a Framework of Usability Characteristics," Master's Thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Thorndike, E. L., and Woodworth, R. S. (1901), "The Influence of Improvement of One Mental Function Upon the Efficiency of the Other Functions," *Physiological Review*, Vol. 8, No. 3, pp. 247–261.

Tippett, W. J., Lee, J-H., Zakzanis, K. K., Black, S. E., Mraz, R., and Graham, S. J. (2009), "Visually Navigating a Virtual World with Real-World Impairments: A Study of Visually and Spatially Guided Performance in Individuals with Mild Cognitive Impairments," *Journal of Clinical and Experimental Neuropsychology*, Vol. 31, No. 4, pp. 447–454.

Turk, M. (2002), "Gesture Recognition," in *Handbook of Virtual Environments: Design, Implementation, and Applications*, K. M. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 223–238.

Ultimate 3D Links (2010), "Commercial 3D Software," available: http://www.3dlinks.com/links.cfm?categoryid=1&subcategoryid=1#, accessed May 25, 2010.

Van Merriënboer, J. J. G., and Kester, L. (2005), "The Four-Component Instructional Design Model: Multimedia Principles in Environments for Complex Learning," In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning*, Cambridge University Press, New York, pp. 71–93.

Vertanen, K., and Kristensson, P. O. (2009), "Parakeet: A Continuous Speech Recognition System for Mobile Touch-Screen Devices," in *Proceedings of the 13th International Conference on Intelligent User Interfaces*, Sanibel Island, FL. February 9–11, pp. 237–246.

Vince, J. (2004), *Introduction to Virtual Reality*, 2nd ed., Springer-Verlag, Berlin.

Vozenilek, J., Huff, J. S., Reznek, M., and Gordon, J. A. (2004), "See One, Do One, Teach One: Advanced Technology in Medical Education," *Academic Emergency Medicine*, Vol. 11 No. 11, pp. 1149–1154.

Washburn, D. A., and Jones, L. M. (2004), "Could Olfactory Displays Improve Data Visualization?," *Computing in Science and Engineering*, Vol. 6, No. 6, pp. 80–83.

Washburn, D. A., Jones, L. M., Satya, R. V., Bowers, C. A., and Cortes, A. (2003), "Olfactory Use in Virtual Environment Training," *Modeling and Simulation*, Vol. 2, No. 3, pp. 19–25.

Welch, R. B. (1978), *Perceptual Modification: Adapting to Altered Sensory Environments*, Academic, New York.

Welch, R. B. (1997), "The Presence of Aftereffects," in *Design of Computing Systems: Cognitive Considerations*, G. Salvendy, M. Smith, and R. Koubek, Eds., Elsevier Science Publishers, Amsterdam, The Netherlands, pp. 273–276.

Wiecha, J., Heyden, R., Sternthal, E., and Merialdi, M. (2010), "Learning in a Virtual World: Experience with Using Second Life for Medical Education," *Journal of Medical Internet Research*, Vol. 12, No. 1, p. e1, available: http://www.jmir.org/2010/1/e1/, accessed May 20, 2010.

Witmer, B., and Singer, M. (1998), "Measuring Presence in Virtual Environments: A Presence Questionnaire," *Presence: Teleoperators and Virtual Environments*, Vol. 7, No. 3, pp. 225–240.

Yee, N., Bailenson, J. N., and Ducheneaut, N. (2009), "The Proteus Effect: Implications of Transformed Digital Self-Representation on Online and Offline Behavior," *Communication Research*, Vol. 36, No. 2, pp. 285–312.

Yeh, S-C., Parsons, T. D., McLaughlin, M., and Rizzo, A. A. (2007), "Virtual Reality Upper Extremity Motor Training for Post-Stroke Rehabilitation," *Journal of the International Neuropsychological Society*, Vol. 13, Suppl. S1, p. 58.

Zotkin, D. N., Duraiswami, R., Grassi, E., and Gumerov N. A. (2006), "Fast Head Related Transfer Function Measurement Via Reciprocity," *Journal of the Acoustical Society of America*, Vol. 120, No. 4, pp. 2202–2215.

# CHAPTER 37

# NEUROERGONOMICS IN HUMAN–SYSTEM INTERACTION

**Dylan Schmorrow**
**U.S. Navy**
**Arlington, Virginia**

**Kay M. Stanney, Kelly S. Hale, and Sven Fuchs**[*]
**Design Interactive, Inc.**
**Oviedo, Florida**

**Glenn Wilson**
**Air Force Research Laboratory**
**Wright-Patterson Air Force Base, Ohio**
**and**
**Physiometrex**
**Keizer, Oregon**

**Peter Young**
**Colorado State University**
**Fort Collins, Colorado**

*Despite moments of insight and even genius, the human mind often seems to fall far below its full potential. The level of human thought varies greatly in awareness, efficiency, creativity, and accuracy. Our physical and sensory capabilities are limited . . . [Moreover], our tools are difficult to handle, rather than being natural extensions of our capabilities. In the coming decades, however, converging technologies promise to increase significantly our level of understanding, transform human sensory and physical capabilities, and improve interactions between mind and tool.*

Roco and Bainbridge (2002, p. 4)

## 1 INTRODUCTION

Realizing the full potential of the human mind is arguably the goal of neuroergonomics. This burgeoning field evolved from a need to increase the intimacy of the coupling between human and system in an effort to alleviate costs that less sensitive approaches to augmenting human performance have imposed. Our implements are intended to be intuitive, natural extensions of human capabilities; yet more often than not support solutions fall far short of this goal. For example,

---

[*] Current affiliation: Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE, Wachtberg, Germany.

in the modern era, as tools evolved, they brought to light the physical limitations of the human and, in turn, the field of ergonomics emerged to address these limitations (Stanney, in press). Consequently, with proper ergonomic design, an anatomical fit between human and system is achieved. With the information age, massive data revealed the cognitive limitations of the human and in response the field of cognitive ergonomics (also called human–computer interaction, usability, human–systems integration) arose and provided for a better cognitive fit between human and system (Hoc, 2008). Yet, even with the gains provided by these fields, the systems of today often stretch human capabilities, rendering the solutions far less powerful than intended. Current predictions are grim, with some forecasting that by "2030 machine capabilities will have increased to the point that humans will have become the weakest component in a wide array of systems and processes" (Dahm, 2010, p. x).

In an effort to avoid this inauspicious outcome and, in turn, realize the human potential, the nature of human–system couplings is evolving. Whereas in the past efforts to resolve incompatibilities between human and system have been tackled from the "outside in," providing support via better cognitive and physical design, current efforts are delving inward peering into the brain and translating neural signals into models that can precisely support the current cognitive and physical state of the human. Thus, the third ergonomic era has emerged—that of neural ergonomics (also called neuroergonomics; Parasuraman 2003; Parasuraman and Wilson, 2008), which seeks to direct and adapt human–system interaction based on real-time knowledge of human brain function, thereby achieving precise optimization of the fit between human and system. This era aims to overcome the constraints that human cognitive and physical limitations present and add the brain as another physical structure around which ergonomic principles can be applied for optimizing interactive system design (Hancock and Szalma, 2004; Fafrowicz and Marek, 2007). The result is the design of a neuro-sustainable environment that embraces the human potential rather than one that creates meaningless conflict with our fastidious, built-in human software (Tolja, 2010).

While much work still needs to be done to formalize the neuroergonomics field of study, Parasuraman and Hancock (2004, p. 4) have suggested that the tenets of this field include the view that "(a) the human brain implements cognition and action, (b) the brain is itself shaped by the physical environment, and (c) both brain and behavior (e.g., action) must be examined in order to understand fully how human cognition and action are coordinated with the world of artifacts." Neuroergonomics is seen as an expansion of the communication channels between human and system (Hancock and Szalma, 2004), one that is based on a deep knowledge of the brain in its situated environment. The power of neuroergonomics thus lies in the validity of its measurement of cognitive activity, established by correlating neurophysiological signatures to cognitive measures (e.g., workload, engagement, distraction) and associating these indicators with a given environmental context

(Hettinger et al., 2003). This, in turn, allows for adaptation of the environment to better fit the capabilities of the human. One can easily see the value of such neuroadaptive systems in the context of a monitoring task, such as airport baggage screening (Carpenter et al., 2010), where a change in brain state indicating inattentiveness to a monitoring event could be detected and the screener's attention could be appropriately redirected to an item of interest. Such adaptations are not limited to the cognitive domain, as possible applications of neuroadaptive systems extend to physical activity, where sensing equipment could be used to detect physical limitations (e.g., muscular fatigue, musculoskeletal injury, physical disability) and adapt human input requirements accordingly (Karwowski et al., 2003). In such cases, an individual would no longer have to use muscular activity (e.g., press a key, move a mouse) to interact with the environment but could instead direct their mental activity and use unique brain electrical "signatures" to control external devices (Parasuraman and Wilson, 2008). Such synergy between human and system has brought about ethical concerns. Specifically, as the coupling between human and system becomes more intimate and the number of direct links between brain activity and system activity increases, the boundaries of human identity blur; such philosophical matters have been discussed in detail elsewhere (Hancock and Szalma, 2004, 2006; Keebler et al., 2010).

Neuroadaptive systems thus promise to provide a very tightly coupled feedback loop in which the system modifies its behavior to effectively accommodate meaningful variations in the state of the user, thereby remediating the fundamental communications disconnect that so often exists between humans and their support systems (Hettinger et al., 2003). Such systems aim to achieve highly synergistic communication between human and system by directly detecting and recognizing situations in which the user's cognitive and/or physical state has changed in a manner that affects system interaction and then balancing information flow and interaction demands on a moment-by-moment basis to precisely match the dynamic state of the user. This dynamic synergy may ultimately bring us closer to realizing the full potential of the human mind. To realize this goal, it is essential to first understand human information processing and how best to dynamically adapt information flow and interaction demands given the human's capabilities and limitations.

## 1.1 Human Information-Processing Limitations

Current understanding of human information processing suggests that information is perceived through multiple sensory processors. This information is then perceptually encoded (i.e., stimulus is identified and recognized), processed by a working memory (WM) subsystem that is regulated and controlled by attention via the executive function (EF), which may be supported by long-term memory (LTM), to arrive at a decision, which in turn triggers a human response (Baddeley, 1986, 1990, 2000; Wickens, 1992). Within human information processing there are several "bottlenecks" or points of limited

processing capacity, including sensory memory, WM, attention, executive function, and response execution.

### 1.1.1 Sensory Memory Bottleneck

Sensory memory is responsible for encoding information and converting it to a usable mental form (Atkinson and Shiffrin, 1968, 1971). Research suggests there may be different sensory memory system for each of the human senses, including visual, auditory, tactile (haptic), olfactory, and gustatory. Behavioral studies suggest that human information processing begins with information being perceived on average in about 100 ms (Cheatham and White, 1954; Harter, 1967) by one of the sensory processors. The visual *iconic* sensory memory modality has been suggested to have an average capacity of about 17 items, and this iconic percept is fleeting, decaying completely, on average, in about 200 ms if it does not transfer to WM (Sperling, 1960, 1963; Averbach and Coriell, 1961; Neisser, 1967). Audition, or *echoic* sensory memory, is suggested to have an average capacity of five items and is a bit more persistent, with the "internal echo" lasting an average of about 1.5 s (Neisser, 1967; Darwin et al., 1972). *Haptic* sensory memory is very limited in terms of capacity (Watkins and Watkins, 1974; Mahrer and Miles, 2002) and has a decay rate between 2 and 8 s (Bliss et al., 1966; Posner and Konick, 1966; Lachman et al., 1979). Little is known about olfactory and gustatory sensory memories. In general, a considerable amount of information can be perceived if it is allocated across *multiple* sensory systems. Thus, *given the limited capacity of sensory memory, neuroergonomics seeks to enhance sensory perception by exploiting multiple sensory channels for increased input capacity* (see Table 1). Sensory stimuli that have passed the sensory memory bottleneck and are rapidly decaying must then compete for the drastically limited resources of WM and attention.

### 1.1.2 Working Memory Bottleneck

Working memory allows people to maintain and manipulate information that has been perceived by sensory memory and is currently available in a short-term memory store. In general, WM is described as a functional component of cognition "that allows humans to comprehend and mentally represent their immediate environment, to retain information about their immediate past experience, to support the acquisition of new knowledge, to solve problems, and to formulate, relate, and act on current goals" (Baddeley and Logie, 1999, p. 29). It is considered a temporary active storage area where information is manipulated and maintained for executing simple and complex tasks (e.g., serial recall, problem solving). Working memory is divided into separate processes that are required for short-term storage [according to Baddeley and Logie's (1999) model, these include the phonological loop and visuospatial sketchpad] and for allocating attention and coordinating maintained information (i.e., the executive function).

Working memory is still being defined, and research has suggested dissociations in both the phonological loop (i.e., phonological store vs. articulatory rehearsal mechanism) (Baddeley and Logie, 1999) and visuospatial sketchpad (visual form and color recognition vs. localization) (Carlesimo et al., 2001; Mendez, 2001; Pickering, 2001). Further definition of the visual component of working memory has recently suggested that only one *high-resolution* object representation can be maintained in visual working memory (while many high-resolution representations can be maintained in iconic sensory memory and visual short-term memory; Sligte et al., 2010). Such capacity has recently been found to depend on the informational maximum rather than an item number maximum (Alvarez and Cavanaugh, 2004; Luria et al., 2010; Sligte et al., 2010). In general, WM is said to have a limited capacity of about seven chunks, a rapid decay rate of about 200 ms, and a recognize–act processing time of 70 ms, on average (Miller, 1956; Card et al., 1983). Research suggests, however, that presenting information multimodally can in fact enhance human information processing via an increase in WM capacity, with gains on the order of three times Miller's (1956) "magical number" of 7 being realized in one such study (Samman et al., 2004). These gains could be tempered if the costs for modality switching are high; this is discussed in the next two sections.

**Table 1 Neuroergonomics Approaches to Overcoming Human Information-Processing Bottlenecks**

| Human Information- Processing Bottleneck | Objective |
| --- | --- |
| Sensory memory | Enhance sensory perception by exploiting multiple sensory channels for increased input capacity |
| Working memory | Support simultaneous processing of competing tasks by allocating data streams strategically to various multimodal sensory systems while maintaining multimodal information demands within working memory capacity |
| Attention | Equip computers such that they become aware of subtle cues emanating from humans indicating how they are prioritizing incoming information (i.e., directing attention) and capitalize on these cues to enhance human information processing |
| Executive function | Enhance information processing by directing the recall of contextual information that cues the optimal interpretation of incoming information and moderates the effects of modality switching |
| Response execution | Expand available response modalities by allowing for direct brain–computer interaction |

Given separable WM components and WM capacity enhancements based on modality, Wickens's (1984) multiple resource theory (MRT) can be expanded to suggest that *modality-based* resources can be utilized strategically at different points in user interaction to streamline a user's cognitive load (Stanney et al., 2004). In such a case, total WM capacity will depend on how dissimilar streams of information are in terms of modality. An expanded MRT would address how to allocate multimodal WM resources, particularly during multitasking, in such a way as to allow attention to be time shared among various tasks. Thus, *neuroergonomics seeks to support simultaneous processing of competing tasks by strategically allocating data streams to various multimodal sensory systems while maintaining multimodal information demands within WM capacity* (see Table 1).

### 1.1.3 Attention Bottleneck

Three general categories of attention theories can be found in the literature: (1) "cause" theories, in which attention is suggested to modulate information processing (e.g., via a spotlight that functions as a serial scanning mechanism or via limited resource pools); (2) "effect" theories, in which attention is suggested to be a by-product of information processing among multiple systems (e.g., stimulus representations compete for neuronal activation); and (3) hybrids that combine cause-and-effect theories (Fernandez-Duque and Johnson, 2002). In general, attention is suggested to be a selective process via which stimulus representations are transferred between sensory memory and WM and then contribute to the processing of information once in working memory. Attention improves human performance on a wide range of tasks, minimizes distractions, and facilitates access to awareness (i.e., focused attention). In the best case, attention helps to filter out irrelevant multimodal stimuli. In the worst case, critical information is lost due to overload of incoming information, stimulus competition, or distractions. Thus, if one were to try to enhance WM via multimodal interaction, such stimulation would impose a trade-off between the benefits of incorporating additional sensory systems and the costs associated with dividing attention between various sensory modalities (Makovski and Jiang, 2007). Thus, while directed attention can modulate maintenance of specific representations in WM and help define the interplay between attention and WM (Lepsien and Nobre, 2007), attention must be moderated judiciously if it is to support enhanced human performance. Further, "executive" attention (McCabe et al., 2010), also known as executive control (Logan, 2003), attentional control (Balota et al., 1999), controlled attention (Engle et al., 1999), cognitive control (Depue et al., 2006; Jacoby et al., 2005), and inhibitory control (Hasher et al., 2007), among others, is thought to be *the* cognitive ability underlying performance on complex cognitive tasks, and thus its correct modulation is essential if a high level of human–computer symbiosis is to be achieved. Specifically, neuroergonomics seeks to "build systems that sense, and share with users, natural signals about attention to support … fluid mixed-initiative collaboration with computers … an assessment of a user's current

and future attention (could thus) be employed to triage computational resources" (Horvitz et al., 2003, p. 52). Thus, *with neuroergonomics, computers will become aware of subtle cues emanating from humans indicating how they are prioritizing incoming information (i.e., directing attention) and will capitalize on these cues to enhance human information processing* (see Table 1).

### 1.1.4 Executive Function Bottleneck

The EF system is suggested to be responsible for selection, initiation, and termination of human information-processing routines (e.g., encoding, storing, and retrieving) (Matlin, 1998; Baddeley, 2003). It controls (i.e., focuses, divides, and switches) attention, integrates information from WM subcomponents, and connects WM with contextually triggered information from LTM. The EF is thus associated with regulatory processes underlying the control of human information processing and sheds light on operational costs associated with these control activities (Zakay and Block, 2004). The EF is thought to be especially active in handling novel situations (i.e., those with contextual ambiguity), such as those involving planning or decision making, error correction or troubleshooting, novel sequences of actions or responses, danger or technical difficulty, or the need to overcome habitual responses (Norman and Shallice, 1980; Shallice, 1982). When a person faces such contextual ambiguity during human information processing, high-level control functions of the EF become engaged. During such processing, a person will retrieve the multiple interpretations associated with a given uncertain situation, choose the more likely interpretation based on context and frequency of occurrence, discard alternative interpretations, and mark that point in their information representation as a choice point (Zakay and Block, 2004). Reducing contextual ambiguity, and thus effortful EF processing, would involve easing selection among multiple interpretations by increasing the number of contextual cues associated with any given alternative.

As indicated previously, frequent switching between one modality or task and another will incur a cost of switching that will be associated with inhibitions of responses to the previous modality stimuli or task, selection and activation of the response best associated with the new modality or task context, and resequencing of these stimuli. Since more frequent switching may entail greater contextual changes, it is expected to engage effortful EF processing. Thus, it is important during modality switching to consider the cost of such contextual changes. However, recent research has demonstrated that executive control is plastic and adaptive in terms of its interference resolution process and thus increased efficiency of this process can be obtained through training and could be used to support such modality switching (Persson and Reuter-Lorenz, 2008). Such gains could prove particularly helpful during high-load information processing. *Neuroergonomics seeks to enhance information processing by directing recall of contextual information that cues optimal interpretation of incoming information and moderates the effects of modality switching* (see Table 1).

### 1.1.5 Response Execution Bottleneck

At a choice point, if the EF selects an action, decision information moves into the response execution stage where action commences via a response modality (manual, vocal, eye gaze). Stimulus–central processing–response (SCR) compatibility theory suggests that multimodal stimuli are processed in modality-specific codes and that information will be processed faster if a stimulus (S) is matched with an appropriate response (R) modality (Wickens, 1992). Specifically, tasks that demand verbal WM, such as interpretation of system status, are thought to be best presented via speech and require a *verbal* response, while those that require visual WM, such as remembering the location of a threat on a radar screen, are best presented visually and responded to *manually* (Wickens and Hollands, 2000). However, these mappings may not be as straightforward during multitasking, as response performance (e.g., time, accuracy) tends to break down during complex, high-workload task conditions (Jones et al., 2004). Thus, there is a need to expand response modalities and provide control mechanisms that are effortless and intuitive for use in multitasking environments. Neuroergonomics provides the potential for such a solution by allowing control of interactive systems via neural signals [e.g., electroencephalography (EEG) signals, event-related potentials (ERPs)]. Such neurally based response systems, called brain–computer interfaces, accept commands directly from the human brain without requiring a physical action from the user and use these commands to operate and direct interactive technologies (Hettinger et al., 2003). In its ultimate instantiation, where technology can reliably recognize a multitude of human thought patterns, the interactive system becomes "an extension of the mind itself" (Lusted and Knapp, 1996, p. 82). Such systems could open up the world of computing for those with physical disabilities and limitations (Karwowski et al., 2003). *Neuroergonomics thus seeks to expand the available response modalities by allowing for direct brain–computer interaction* (see Table 1).

## 2 COGNITIVE STATE ASSESSORS

Neuroergonomics seeks to enhance human–system interaction substantially by adopting a paradigm shift from primarily passive systems dependent on user input to proactive systems that gauge and detect, via diagnostic psychophysiological sensors, human information-processing bottlenecks and then employing augmentation strategies to overcome these limitations. To realize this paradigm shift, one must first be able to characterize *cognitive state* such that the noted bottlenecks can be monitored and regulated appropriately. Research in psychophysiology, principally through brain-imaging techniques, has established a correspondence between cognitive processors and particular brain structures that have an identifiable locus in the brain. This allows use of neural signals from those structures as a diagnostic tool of *cognitive load*, which can be measured in real time while a person is engaged with an interactive

system. Such psychophysiological data streams can be used to characterize cognitive state, specifically current load on information-processing bottlenecks.

## 2.1 Psychophysiological Techniques for Capturing a Cognitive State

Many human–system interactive situations do not provide sufficient human performance information that can be used to infer cognitive state or what shall herein be called an operator's functional state (OFS). This is especially true of highly automated systems, which for the most part put the human in a monitoring role (Byrne and Parasuraman, 1996). Because system monitoring does not require overt behavioral responses, it is difficult to assess user state. Thus, a user may not be in an optimal state at all times, and system corrections or malfunctions may not be detected and responded to correctly. A methodology is needed that provides accurate assessment of OFS in the absence of overt performance data and to provide additional information when performance data are available. Psychophysiological measures have been suggested to fill this role.

Psychophysiological signals are always present and can often be collected unobtrusively, thereby providing a source of uninterrupted information about user state (Kramer, 1991; Wilson and Eggemeier, 1991; Scerbo et al., 2001; Wilson, 2002a; Gratton et al., 2008). Correlations between psychophysiological measures and OFS have been described (Wilson and Schlegel, 2003). Although these correlations do not prove causality, they do suggest that psychophysiological measures can be used to assess OFS and, further, that this information can be used to modify system parameters to meet the momentary needs of users (i.e., cognitive augmentation via adaptive aiding). Of the several criteria for implementation of OFS-driven adaptive aiding, three crucial ones are that (1) significant and meaningful system performance improvements must be demonstrated; (2) the sensors used must be nonintrusive to a user's primary task, as this would hinder human–system performance; and (3) their use must be acceptable to users.

For widespread adoption, it must be demonstrated that OFS assessment and aiding either (1) improve human performance and enhance job success for work-related applications or (2) enhance the interactive experience for entertainment-based or other such experiential applications. An example of a successful application of adaptive aiding is the use of antigravity (anti-*g*) suits, which require wearing additional gear that inflates at predetermined *g*-levels. These suits have been proven to save lives because they can prevent *g*-induced loss of consciousness in jet pilots and have therefore met with wide acceptance.

### 2.1.1 Current Status

In the past, the typical approach when using psychophysiological measures to assess OFS was to collect one or more measures and demonstrate that statistically significant differences exist between at least two levels of task demand or human state such as fatigue. Most

of this research has been conducted in the laboratory. However, a growing body of research is expanding into operational environments. Psychophysiological measures have been applied successfully in driving, flight, and other test and evaluation environments (Wilson, 2002a). For example, heart rate has been shown to be increased significantly under high-mental-workload conditions compared to low-mental-workload conditions during flight (Wilson, 2002b; Kobus et al., 2005). EEG, a physiological measure of the momentary functional state of cerebral structures, provides useful information about both high cognitive workload and inattention (Kramer, 1991; Wilson and Eggemeier, 1991; Gundel and Wilson, 1992; Sterman and Mann, 1995; Gratton et al., 2008). Specifically, theta-band EEG activity has been reported to increase with increased task demands (Gundel and Wilson, 1992; Gevins et al., 1998; Hankins and Wilson, 1998). While much work has been done on cognitive state sensors, the current maturity of neuroergonomic measures varies widely (Fidopiastis and Wiederhold, 2008; Hale et al., 2012). Some measures are still in technology development, such as functional near-infrared imaging (fNIR), posture tracking, and pupilometry. Others are in technology demonstrations, such as electrodermal response (EDR), EEG, electromyography (EMG), and galvanic skin response (Kobus et al., 2005). Still others have made their way into the operational system and/or subsystem, such as electrocardiography (ECG; Schnell et al., 2008a, 2008b) and eye/gaze tracking (Carroll et al., 2009; Carroll, Fuchs, Hale et al., 2010). Several efforts have demonstrated real-time state measures over the past 10 years that meet the requirements of (1) sensitivity to different brain states and/or processes, (2) reliability, and (3) practicality in fielded use (Stanney et al., 2010). The current challenge with cognitive state measurement is in developing means to substantially advance real-time data fusion and classifier construction techniques (Fidopiastis and Wiederhold, 2008; Hale et al., 2012).

### 2.1.2 Current Technology for Recording Psychophysiological Data

Numerous psychophysiological measures have been shown to provide valuable information concerning OFS in real-world operational environments (Wilson, 2002a; Wilson and Schlegel, 2003; Gratton et al., 2008). Because of the restrictions of the operational environment, some psychophysiological methods cannot be used. For example, positron emission tomography (PET), functional magnetic resonance imaging (fMRI), and magnetoencephalography (MEG) are not practical OFS gauges because the associated recording equipment is too restrictive, too large, and requires special shielding, among other prohibiting conditions. Even those measures that are less prohibitive have drawbacks. Almost all currently available, operationally useful psychophysiological sensors require contact with a user's body and use some form of electrolyte sensors. This is the case for EEG, ECG, EMG, and electrooculography (EOG). Users typically do not like to wear such sensors and associated equipment. Further, the sensors are usually attached to the skin with some type of adhesive, and repeated application in a day-to-day operational environment may cause skin irritation. There are less invasive options, such as pupillometry and eye point of regard, which are making their way into current applications (Hale et al., 2012).

An area of concern for all psychophysiological measures is that of artifacts. This is especially the case in operational environments where operators move, talk, walk, and engage in other activities that produce artifacts in recorded data. Techniques are available for detecting and removing certain artifacts, for example, EOG contamination of EEG (He et al., 2007). Other artifacts are difficult to remove, such as EMG in EEG, and may require the removal of the contaminated data. Artifact-free data must be provided in real time, which requires immediate detection and removal of the offending artifacts.

### 2.1.3 New Sensor Technologies

New sensor technologies promise to provide users with more acceptable recording methods and valuable OFS data. Sensors that require only "dry" (no electrolyte or adhesive) contact with the skin have been developed (Kingsley et al., 2002; Trejo et al., 2003) and tested (Christensen et al., 2009, 2010b). Two approaches that are being explored for dry EEG sensors are capacitive coupled and optical sensors. These technologies can also be used to record ECG, EMG, and EOG. Currently, dry-sensor EEG can be recorded from non–hairy skin areas such as the forehead, and low-cost consumer-grade devices have emerged that utilize this technology for entertainment purposes like gaming, but these solutions have also been used in research studies (cf. Rebolledo-Mendez et al., 2009). For most research applications, however, larger numbers of sensors may be necessary, and the goal is to be able to record EEG from anywhere on the scalp using these sensors. A step towards this is products that use moist pads instead of conductive gel, such as the headset described in Campbell et al. (2010). Eye activity can be recorded using video cameras that image the face from a distance, requiring no actual contact with users (Carroll et al., 2009). Recent advances include tracking devices that feature motorized cameras that follow the eyes of the user to better compensate for head movements (http://www.interactive-minds.com/en/eye-tracker/eyefollower). Additionally, sensor technology has been developed that provides measures of brain activity using blood flow technology. For example, fNIR sensors provide information about brain oxygen levels, cortical blood volume, and neuronal activity (Izzetoglu et al., 2003, 2005; Wildey et al., 2010).

### 2.1.4 Functional Near-Infrared Sensors

Using near-infrared light emitters, near-infrared energy can be directed through the scalp and skull and reflected from underlying cerebral tissue. Two types of cerebral information can be obtained from fNIR. The first type is hemodynamic response, reflecting oxyhemoglobin and deoxyhemoglobin concentrations in the brain. The consensus is that increased brain activity results in
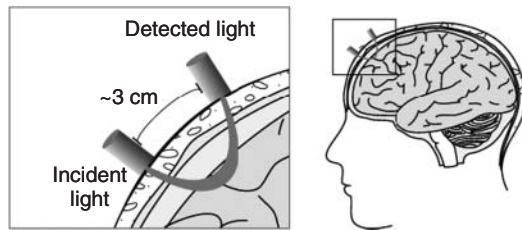
**Figure 1** Locations of the infrared emitter and detector area important to ensure that cortical tissue is imaged. (From Downs and Downs, 2004.)

increased levels of local oxyhemoglobin and decreased levels of deoxyhemoglobin (Gratton and Fabiani, 2001; Gratton et al., 2008). These responses have been used to investigate cognitive activity (Hock et al., 1997; Villringer and Chance, 1997; Takeuchi, 1999; Izzetoglu et al., 2003, 2005). The second type of information that can be obtained from fNIR is to detect changes in the optical characteristics in brain tissue that are related to neuronal activity (Gratton and Fabiani, 2001; Gratton et al., 2008). The exact cause of these optical changes is not totally understood. This latter method is said to provide millisecond temporal resolution; the first method is much slower. For either procedure the infrared emitters and sensors have only to touch the scalp rather than being affixed to it (see Figure 1). The emitter–sensor unit can be held in place using a strap or cap arrangement. Conventional fNIR systems may be impaired by absorption from the subject's hair and thus they are most often applied to nonhairy regions, such as the forehead. Recent work in fNIR hairbrush sensors has improved the sensitivity over hairy regions by redesigning the optrode with fiber tips designed to thread through the hair, which provides better scalp contact (Wildey et al., 2010). These systems function on hairy areas of the scalp and so are not restricted to the forehead region. This developing technology holds a great deal of promise for advancing our understanding of cognition and may be used more readily in operational environments than sensor technologies that require adhesives.

## 2.2 Transforming Sensors into Cognitive State Gauges

To be useful, real-time assessment of cognitive activity by means of psychophysiological measures must be transformed from individual measures to cognitive gauges. Whereas consideration of individual measures provides valuable information, neuroergonomics requires gauges that are composite estimates characterizing the functional state of a user (such as those to gauge load on the human information-processing bottlenecks, as well as others, such as Kolmogorov entropy of EEG signals and task load, which are mentioned in Sections 3.1.3 and 3.1.4). Given the complexity inherent to most operational environments, it is not sufficient simply to be aware that statistical changes exist in several measures. Measures or gauges must be able to characterize

the functional state of a user such that this information can be used to implement adaptive aiding (i.e., triggering of augmentation strategies) in real time in real-world situations. In 2003, the U.S. Defense Department Defense Advanced Research Project Agency (DARPA) conducted a technology integration experiment (TIE) with various psychophysiological sensors [i.e., EEG, event-related potential, fNIR, pupil dilation, heart rate variability (HRV), arousal, galvanic skin response] to demonstrate the feasibility of simultaneous data collection (Morrison et al., 2003). The TIE demonstrated that real-time computation of sensor data to produce online gauge information was feasible and further confirmed that several sensor technologies could be combined with minimal interference. However, substantial variability between human participants in gauge sensitivity suggested the need for additional research. The unique, individual participant psychophysiological response characteristics are well known and no doubt contribute to this variability. Additional research has thus focused on how to transform sensors to specific OFS gauges. One such example is the composite stress gauge (Raj et al., 2003), which uses a weighted average of pupillometry ECG and electrodermal response (EDR) to detect a participant's response to changes in cognitive load. The New Workload Assessment Monitor (NuWAM) sensor suite is another example and is based on an artificial neural network (ANN) cognitive state classifier and combines EEG, ECG, and EOG sensors (Krizo et al., 2005; Wilson and Russell, 2006). The Cognitive Cockpit (CogPit) sensor suite is yet another example; this suite gauges cognitive–affective status in near-real time as derived from four main sensor sources, including behavioral measures, EEG, subjective measures, and contextual information (Dickson, 2005). More recently, there have been a number of systems that have integrated eye tracking and EEG (Tucker and Luu, 2009). One such example is the sensor suite integrated into the Auto-Diagnostic Adaptive Precision Training (ADAPT) framework, which combines EEG, eye tracking, heart rate, and behavioral measures to capture cognitive state on a second-by-second basis and evaluate an individual as they progress from novice to expert (Carroll et al., 2010a).

*Thus, neuroergonomics seeks to leverage a set of psychophysiological gauges that allow for real-time assessment of cognitive state, particularly current load on information-processing bottlenecks, which can then be transformed directly into computer control commands for triggering implementation of augmentation strategies. A further goal is to develop procedures that provide predictions of future bottlenecks so that corrective actions can be taken before system performance degradation. Reactive systems, which detect and respond to current bottleneck conditions, will be very useful but may result in uneven system performance because they react to already poor OFS. Systems that can accurately predict upcoming bottlenecks will be able to maintain continuous optimal system performance without the possible perturbations of a reactive system. These systems will be able to detect trends in the physiological signals that can predict OFS breakdowns.*

## 3  HUMAN−SYSTEM AUGMENTATION

In Section 1, various human information-processing bottlenecks were discussed (i.e., sensory memory, WM, EF, attention). In Section 2, means of gauging the current cognitive load on a person were considered. Neuroergonomics seeks to overcome the noted points of limited capacity processing through the utilization of human−system augmentation strategies, which will be triggered by cognitive state gauges. It is suggested that through dynamic augmentation strategies the cost of these bottlenecks (e.g., degraded human performance due to overload, underload, stress, losses in situational awareness, or emotional state) can be overcome.

When designing dynamic augmentation strategies, designers should target individualized approaches for their selection and configuration. To optimize the effectiveness of augmentations and develop trusted, adaptable, and flexible task allocation schemes and seamless and elegant interruption management strategies, augmentations should consider not only the cognitive state of the user but also individual learning/operating styles and preferences that may develop over time during system use. The relationship between the human user and an intelligent augmentation system should be viewed as a team environment; in turn, a "shared mental model" (Cannon-Bowers et al., 1993) should be instilled, including the system's understanding of the user's styles, preferences, capabilities, and intent. Shared mental model components (cf. Mathieu et al., 2000) that current adaptive systems often do not account for but that could be considered for inclusion in neuroadaptive systems include interaction patterns, communications channels, role interdependencies, and task strategies, as well as the knowledge, attitudes, preferences, and tendencies of teammates. Thus, to optimally augment human−system interaction, the system must be able to detect and learn (1) its user's intent and information needs at any given time, (2) the individual styles and behavioral patterns exhibited by the user and his or her teammates, (3) which cognitive state patterns are likely to impact performance in specific users, and (4) the cognitive limitations and preferences of user and teammates, including individual cognitive capacity, information-processing styles, attention span, and other relevant attributes.

## 3.1  Augmentation Strategies

In conventional human−system interaction, an excessive amount of cognitively demanding tasks can be imposed on a user. In such situations, human information processing can break down at any of the identified bottlenecks. Instead of overloading users, interactive systems should seek to achieve cognitive congeniality (Kirsh, 1996) by (1) *presenting* an optimal level of task-relevant information and ensuring that it is readily perceived, (2) optimizing cognitive load on WM by *sequencing* and *pacing* tasks appropriately, and (3) reducing the number and cost of mental computations required for task success by *delegating* tasks when appropriate. Taken together, these strategies should increase the speed, accuracy, and robustness of human−system interaction. Each of

these augmentation strategies (i.e., task presentation, sequencing, pacing, and delegation) is discussed below. It should be noted that other such strategies can and should be identified. Additional augmentation strategies to consider include but are not limited to techniques for supporting information filtering and triage, multitasking, mixed-initiative interaction, and context-sensitive interaction (Horvitz et al., 2003).

### 3.1.1  Task Presentation

When designing interactive systems, a central question is which information should be conveyed via which modality. Conventional interactive systems present information to users primarily via visual cues, sometimes offering auditory accessories. Yet to optimize sensory processing, thereby relieving the sensory memory bottleneck, one should consider the types of information each modality is particularly suited to display. Table 2 presents theorized suitability of sensory modalities for conveying various information sources. In addition to suitability, one must consider capacity. As aforementioned, Samman et al. (2004) demonstrated that multimodal WM capacity can reach levels nearly three times that of Miller's (1956) magical number 7. Thus, rather than overloading a single modality, by distributing information across multiple modalities, the WM bottleneck can be relieved. Table 3 represents the WM capacity of various modalities based on several studies (Bliss et al., 1966; Sullivan and Turvey, 1974; Smyth and Pendleton, 1990; Keller et al., 1995; Livermore and Laing, 1996; Woodin and Heil, 1996; Feyereisen and Van der Linden, 1997; Matsuda, 1998; Jinks and Laing, 1999; Laska and Teubner, 1999; Frenchman et al., 2003). The numbers in Table 3 suggest the upper limit on the number of items that should be presented via each modality, as individual modality capacity tends to decline during multimodal multitasking even though overall capacity increases (Samman et al., 2004). Thus, with knowledge of the information sources constituting a given application, a determination of optimal modalities can be made to direct multimodal task presentation. More specifically, after characterizing a given application's information sources via a task analysis, first a matching to the optimal modality can be determined using Table 2. Then, given the outcome of the related OFS gauges (i.e., current load on sensory and WM bottlenecks), a determination of reserve capacity can be estimated using Table 3 and a selection of the optimal modality made (i.e., the one with the best match from Table 2 and adequate reserve capacity). If real-time (e.g., physiologically based) workload measures from the user are available, these estimates can be compared to actual onset of an overload condition. Additionally, the effectiveness of changes in task presentation can be evaluated with respect to specific modalities. The system could then dynamically adjust the augmentation parameters and thresholds to account for the individual's cognitive capacity and preferences. The applied implication is that in cognitively demanding task environments, the information should be presented not only in a modality that is most suitable but also in one that is not currently fully loaded, thereby easing the sensory memory and WM bottlenecks. Thus, the first augmentation strategy is to *identify the optimal modality*

**Table 2  Theorized Suitability of Modalities for Conveying Various Information Sources**[a]

| Information Source | Sensory Modality | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| | Visual | Verbal | Tactile | Kinesthetic | Tonal | Olfactory |
| Spatial acuity (size, distance, position) | + + | □ | + | + | | – – |
| 2D localization (absolute/relative location in 2D) | + + | + | + | □ | + | – – |
| 3D localization (absolute/relative location in 3D) | | □ | + | □ | + | – – |
| Change over time | + + | + | + | | □ | – – |
| Persistent attention | + + | – – | + + | | – – | |
| Absolute quantitative parameters | + + | + | – – | – – | – – | – – |
| Temporal (e.g., duration, interval, rhythm) | □ | | + | + | + + | – – |
| Instructions | + | + + | | | | |
| Rapid cuing (e.g., alerts, warning) | + | + | + + | | + + | |
| Surface characteristics (e.g., roughness, texture) | + | | + + | | | |
| Hand–eye coordination (e.g., object manipulation) | | | + + | | | |
| Memory aid (e.g., recognition of a formerly perceived object) | + | + | – | – | + + | + + |
| Affective or ambient information | □ | □ | | | + | + + |

[a]*Key:* + +, best modality; +, next best; □, neutral; –, not well suited, but possible; – –, unsuitable.
*Source*: Adapted from European Telecommunications Standards Institute (ETSI, 2002).

**Table 3  WM Capacity of Various Sensory Modalities**

| WM Subsystem | Capacity |
| --- | --- |
| Visual | 2–5 |
| Verbal | 4–7 |
| Spatial | 5–7 |
| Tactile | 3–5 |
| Kinesthetic | 3–5 |
| Tonal | 4–6 |
| Olfactory | 3–4 |

*by which to present information based on consideration of suitability principles as well as current psychophysiological measures of cognitive load and demonstrated effectiveness of modality augmentations for a given operator* (see Table 4).

### 3.1.2  Task Sequencing

Once the modality by which to present an information source is determined, the information event can be scheduled. The MRT (Wickens, 1984) suggests that people are more efficient in time-sharing tasks when different resources are utilized in terms of sensory stimuli modality (e.g., visual, auditory), WM processing codes (e.g., spatial, verbal), and response modality (e.g., vocal, manual). For example, various studies have suggested that a person can recall more in two tasks with different types of modalities combined than in a single task, especially if the modalities or types of representation are very different (Klapp and Netick, 1988; Penney, 1989; Baddeley, 1990; Cowan, 2001; Sulzen, 2001). More recent MRT efforts have suggested that task interference can be minimized by leveraging opposite ends of four task dimensions, including processing stages (perception, cognition, response): perceptual (sensory), modality (visual, verbal, spatial, tactile, kinesthetic,

tonal, olfactory), visual processing channels (focal, ambient), and WM processing codes (spatial, verbal) (Wickens, 2002). An applied implication of this theory is that time sharing of tasks should be more effective with cross-modal as compared to intramodal information displays. Thus, through systematic sequencing of tasks, simultaneous processing of competing tasks can be allocated strategically across various multimodal sensory systems in an effort to maintain multimodal information demands within WM capacity. Beyond addressing the WM bottleneck, this augmentation strategy can assist in prioritizing incoming information by sequencing cues according to priority, thereby directing attention. When applying this strategy, it is essential to ensure that there is a means to avoid the adaptive state from oscillating too frequently. This can be done through the application of robust controllers (see Section 4). Through systematic control of the adaptive state, this strategy also addresses the EF bottleneck by moderating the effects of modality switching.

To determine task sequencing (i.e., ordering and combining of tasks), a conflict matrix could be calculated following Wickens's (2002) approach, in which the amount of conflict between resource pairs for task couplings is determined. This calculation factors in both conflict and task difficulty (i.e., resource demands), resulting in a task interference value. This could be done in conjunction with a timeline analysis (Sarno and Wickens, 1995), which calculates resource demand levels of time-shared tasks over the time during which the tasks are to be performed. In allocating resources, these principles could be coupled with a scheme of task priorities (as derived through an a priori task analysis), which taken together could guide task ordering and combining given current resource constraints (i.e., task interference values and OFS gauge outputs from all four bottlenecks). Such a system could be further optimized by accounting for individual cognitive capacities and preferences of the user that are measured

**Table 4  Augmentation Strategies**

| Augmentation Strategy | Description | Human Information-Processing Bottleneck Addressed |
|---|---|---|
| Task presentation | Identify optimal modality by which to present information based on consideration of suitability principles, current psychophysiological measures of cognitive load, and demonstrated effectiveness of modality augmentations for a given user | Sensory and working memory |
| Task sequencing | Assign modalities to information sources and schedule them, considering priority, such that they minimize interference over the performance period while leveraging robust controllers to moderate effects of modality switching | Sensory and working memory, attention, executive function |
| Task pacing | Provide external pacing of tasks, which could be achieved by monitoring behavioral entropy, while accounting for individual strategies, preferences, and tendencies | Working memory, attention, executive function |
| Task delegation | Direct assisted explicit task delegation based on psychophysiological indexes of task load and an individual user's knowledge, preferences, and performance | Attention, executive function |

and interpreted by the system during use. The second augmentation strategy is thus to *assign modalities to information sources and then schedule them, considering priority, such that they minimize interference over the performance period while leveraging robust controllers to moderate the effects of modality switching* (see Table 4). This should help relieve the sensory, WM, attention, and EF bottlenecks.

### 3.1.3  Task Pacing

Time management is an essential component of many dynamic task situations (and is also critical to feedback stability of closed-loop systems; see Section 4). Yet, in cognitively demanding task environments, pacing skills can decline rapidly, as temporal judgments depend on the amount of attentional resources allocated to a temporal processor (Casini and Macar, 1999). Further, internal (self) pacing has been shown via EEG signals to impose higher human information-processing demands compared to externally (e.g., via metronome) paced tasks (Gerloff et al., 1998). Disruption of an orderly rhythm is thought to increase the entropy of the human information-processing system, thereby increasing information content due purely to asynchronous pacing of a task. Such disruption can occur when a person becomes overloaded with information, as this often results in delayed event detection and more corrective responses (Boer, 2001). Interestingly, Boer (2001) developed a simple but highly predictive linear model based on Wickens and Hollands's (2000) MRT, which predicted the effect of various tasks on steering entropy and driver performance. The model demonstrated that steering entropy was affected primarily by loading of spatial tasks, as would be predicted by MRT because driving is a highly spatial task. Thus, to achieve effective time management, a potential augmentation strategy would be to *provide external pacing of tasks, which could be achieved by monitoring behavioral entropy, while accounting for individual strategies, preferences, and tendencies* (see Table 4). Specifically, the Kolmogorov entropy (K-entropy) of EEG signals can be

used to assess information flow (Pravitha et al., 2003). K-entropy is proportional to the rate at which information about the state of a dynamical system is lost in the course of time. This entropy index has been shown to fluctuate with changes in the complexity of human information processing, such as that imposed by fatigue (leading to a lesser extent of information flow through particular brain regions) (Rekha et al., 2003) or information overload (King, 1991) while remaining quite stable during performance of demanding cognitive tasks (Pravitha et al., 2003). Thus, using K-entropy of EEG signals to direct task pacing should help relieve the WM, attention, and EF bottlenecks, as it could help optimize the pace of the processing of incoming information and minimize disruptions.

### 3.1.4  Task Delegation

In the context of neuroergonomics, the purpose of dynamic task delegation would be to increase information throughput by balancing the utilization of human resources across a network of users. Task delegation allows for distribution of task demands across individuals as well as coordination between humans and automated systems. In task delegation, certain actions required by a particular task performer are delegated to another performer or back to the system itself once task load gets above some threshold or other types of cognitive breakdowns (e.g., loss of situation awareness, critical performance decrements) are detected (Dearden et al., 2000; Hoc, 2001; Debernard et al., 2002). Such handing off can be implicit (i.e., imposing an allocation based on current OFS load predictions or real-time measurements) or explicit, in that it requires an action from the task performer prior to allocation. Although it has been shown to lead to better performance than explicit allocation, implicit allocation does not always meet with user acceptance, as humans like to maintain control of dynamic task situations and become anxious when they lose control (Hoc et al., 2001). This, in turn, could affect behavioral entropy, thereby affecting system pacing. Taken together, this could affect system

stability properties negatively (see Section 4). Assisted explicit allocation is a compromise, where after detecting an overload using an OFS gauge of task load, such as the task engagement index used by Prinzel et al. (2000, 2003), the interactive system would make an allocation proposal which the human would be able to veto but would not be in charge of allocating. This cooperative task allocation strategy generally leads to effective performance while avoiding complacency by requiring the human to cooperate in the allocation process. Alternatively, the consequences and effects of task delegation could be measured to gauge user acceptance in certain situations, so that such strategies can be used in contexts where no detrimental effects are expected. Thus, a fourth potential augmentation strategy would be to *direct assisted explicit task delegation based on psychophysiological indexes of task load and an individual user's knowledge, preferences, and performance* (see Table 4). This should help relieve the attention and EF bottlenecks, as it eases the need to determine what to attend to.

## 4 ROBUST CONTROLLERS

Although augmentation strategies have the potential to enhance human performance through reducing the load on human information-processing bottlenecks, they could also lead to an adaptive state that oscillates too frequently, thereby destabilizing human–system interaction over time. Thus, there is a need to identify techniques for ensuring that changes requested through the augmentation strategies are implemented so as to maintain system stability and enhance human performance. Mathematical system theory deals with the modeling, analysis, and design of complex dynamic systems. Robust control theory is a discipline of mathematical system theory that is concerned with the analysis and design of feedback controllers for situations where there is only partial or incomplete knowledge of the underlying system dynamics. In the work discussed in this chapter, whereby a user's display/input is adapted based on his or her measured cognitive load, it is important to note that a feedback loop is being *closed* around the human. Moreover, since the underlying system dynamics involve the human, it is certainly true that only partial knowledge concerning a user's state will be available, hence the need for this section on robust control.

### 4.1 Control System Models

Recent developments in the field of cognitive neuroscience have heralded a great deal of change in what is known about human mental operations (Posner and DiGirolamo, 2000; McCabe et al., 2010). As has been discussed, these advances have the potential to allow psychophysiological indicators to direct human–system interaction (Farwell and Donchin, 1988). The ability to use sensors to measure the cognitive performance of a user immediately through psychophysiological characteristics, and virtually instantly adapt a system to meet user needs, presents an exciting new paradigm in interactive systems. The introduction of such real-time adaptive aiding offers the prospect of radically altering how

humans interact with computer technology. However, one important aspect of such a potential change in the nature of human–system interaction is the inherent difference between open- and closed-loop systems.

Even well-understood, stable open-loop systems will show very different performance under closed-loop operation. A simple example of this effect can be seen when bringing a speaker and a microphone (connected to each other) too close together. A well-known audio feedback effect occurs as the signal from the speaker runs through the microphone, back out of the speaker, back into the microphone, and so on. The resulting feedback loop is (typically) unstable and produces a familiar (and unpleasant) sound. The volume of this sound may grow or decay (corresponding to unstable and stable feedback systems, respectively), depending on the proximity of the microphone to the speaker (which implicitly sets the loop gain in the feedback system). Thus, two perfectly well-behaved open-loop systems (speaker and microphone) may or may not be closed-loop stable, depending on how feedback is applied. A more precise quantitative example of such behavior for a neuroergonomic system will be provided later, where it is shown that a stable open-loop system may generate a stable or unstable closed-loop system, depending on how feedback is designed.

Although a great deal about human performance may be understood, the nature of the shift from an open- to a closed-loop system is a unique type of change. As a result, many standard predictable aspects of cognitive and motor performance may operate in drastically different ways in closed-loop systems. A prime candidate for understanding such closed-loop circumstances is through the use of engineering control systems theory. [For a discussion of the pros and cons of various types of models, see Baron et al. (1990).] Control systems theory deals with fundamental properties of systems as described (typically) by mathematical models. It provides a framework and tools for analyzing fundamental system properties, such as performance, noise rejection, and stability, and offers systematic approaches for designing systems with these desired properties.

The idea of applying control theory to humans has some history, with Wiener (1948) widely considered to be the first person to draw parallels between control systems in machines and the organization present within some living systems. However, few attempts have been made to apply control systems theory to human–system interaction (Flach, 1999; Jagacinski and Flach, 2003; Young et al., 2004), and thus this is an exciting area of research where much remains to be done. One notable exception that the current effort draws from is Card et al.'s (1983) model human processor (MHP). The MHP is a human information-processing model consisting of a basic block diagram interconnect model of a human, with an associated estimate of the time taken by each processing stage to process relevant data. For neuroergonomic purposes, the three most relevant stages (i.e., blocks) are probably the perceptual, cognitive, and motor processors. This is illustrated in Figure 2, which shows a human operator piloting a vehicle. In this example, information from the operator's system display
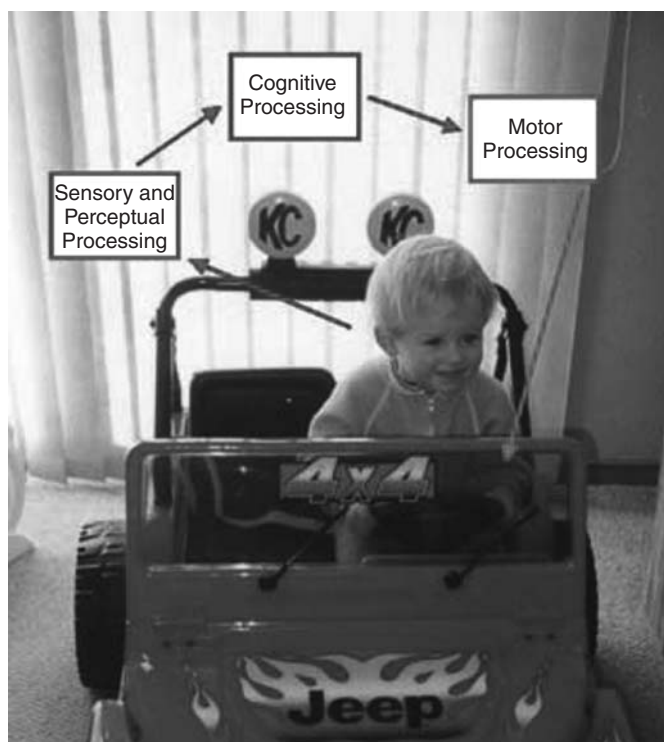
**Figure 2**  Human information-processing model.

would first pass through the operator's perceptual (i.e., sensory) processor, being perceived, on average, in about 100 ms (Cheatham and White, 1954; Harter, 1967). Perceived information would then be available to the cognitive processor, which has an average cycle time of 70 ms. The cognitive processor would then make a decision, and that decision would be implemented by the motor processor, which has an average cycle time of 70 ms, with a resulting action on the vehicle controls. Note that these three blocks provide an internal model of the operator's interaction with the external vehicle displays and controls. This block diagram model not only characterizes the flow of information and commands between the vehicle and operator but also enables us to access the internal state of the operator at various stages in the process. This allows modeling of what a neuroergonomic system might have access to (internal to the human; e.g., load on human information-processing bottlenecks) and how those data might be used to direct closed-loop human–system interaction.

If one considers a control systems model incorporating the flow of human information processing, the time taken by each block adds time delay to the model. However, it does much more than that. As indicated in the early discussion on bottlenecks, it also implies a certain *bandwidth* for the system, both in terms of channel capacity and because signals that vary more rapidly than the time constant of the system (i.e., high-frequency signals) do not pass through it. Hence the processing blocks act as low-pass filters, only allowing through signals that
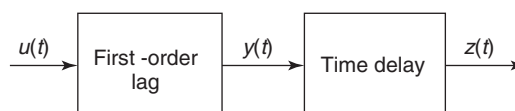


**Figure 3**  Block model for each component.

are below the system bandwidth. For example, humans do not generally perceive the flicker on a computer monitor because it typically occurs at a frequency (100 Hz) higher than that of the perceptual processor's bandwidth of only about 10 Hz. As a first attempt at modeling such a phenomenon, the effects of time lags in human perceptual, cognitive, and motor processing blocks are considered. This results in a dynamic model of the form shown in Figure 3.

Note that the setup depicted in Figure 3 is a generic dynamic model of any one of the MHP components (perceptual, cognitive, motor) shown in Figure 2 (although the model parameters will be different for each). The dynamic models associated with each MHP component ("first-order lag" and "time delay") of the block model are given, respectively, in the time domain (i.e., convolution representation) as

$$y(t) = \frac{1}{\tau} \int_0^t e^{-(t-\gamma)/\tau} u(\gamma)\, d\gamma$$
$$z(t) = y(t - \tau)$$

for each processing block [with overall input $u(t)$ and output $z(t)$], with the time constant $\tau$ taken from the relevant processing time in the MHP model. The first-order lag models the dynamic relationship between input and output signals, which captures the bandwidth effect described earlier. This is most easily seen using the Laplace transform to transform this model from the time domain to an equivalent frequency-domain representation:

$$Y(s) = G(s)U(s)$$

where the function $G(s)$ is given as

$$G(s) = \frac{1}{1 + s\tau}$$

This is known as the *transfer function* of the system. [See Phillips and Parr (1999) for an overview of transform methods for signals and systems; see Ogata (2002) for an overview of the application of these techniques to dynamic systems and feedback control.] A key point is that the time-domain convolution operator has been transformed into a simple multiplication operator in the frequency domain. That multiplication operator, $G(s)$, is both complex valued and frequency varying. The function $G(s)$ captures the frequency response of the system in both magnitude and phase.

To see this, one can evaluate the transfer function along the imaginary axis, that is, substitute $s = j\omega$ into the model (equivalent to specializing the Laplace transform to a Fourier transform) to yield

$$G(j\omega) = \frac{1}{1 + j\omega\tau} = \frac{1}{\sqrt{1 + (\omega\tau)^2}} e^{-j\tan^{-1}\omega\tau}$$

which is the frequency response of the system (with $\omega$ the real-valued frequency). This has the desired low-pass frequency response. Low-frequency (slowly varying) signals pass through almost unattenuated, but higher frequency (rapidly varying) signals are more and more attenuated until hardly any of the signal passes through the system at all. This variation of the magnitude response with frequency in the first-order lag block is what accounts for the computer monitor effect (i.e., lack of perceiving flicker) described earlier (one could not account for this effect with a time delay block alone because the frequency response of a pure time delay is flat, i.e., no variation of magnitude with frequency).

Note that this magnitude response comes with an associated phase response. Low-frequency signals pass though this system with almost undistorted phase. However, as frequency increases, the signals start to incur phase lag, which ultimately reaches $90°$ at high frequency. Phase lag has a destabilizing effect on closed-loop feedback systems, so understanding the relationship between the magnitude and phase of different frequency signals as they pass through the system is of crucial importance in designing any feedback control system.

These various steps have provided the separate pieces necessary to build a model of an entire open-loop system. Since transfer functions operate by multiplication, models for the individual blocks can be cascaded. These are linear models and therefore they commute, so the order of cascade can be changed, and hence time delays can be accumulated into a single block if desired. This now provides a quantitative dynamic model for the human as illustrated in Figure 4. Note that, as discussed above, this model captures the gain–phase relationship with frequency, which is crucial if the model is to be used in a feedback control loop.

This model should allow accurate predictions of open-loop performance and other properties of the system to be made. However, it is important to note that this control theory–based model is in a form that will also allow for prediction of how performance and properties are modified when transforming to a closed-loop setup, which is described in the following section.

### 4.1.1 Closed-Loop Models

Neuroergonomics aims to provide display and information systems that take measurements from OFS gauges,
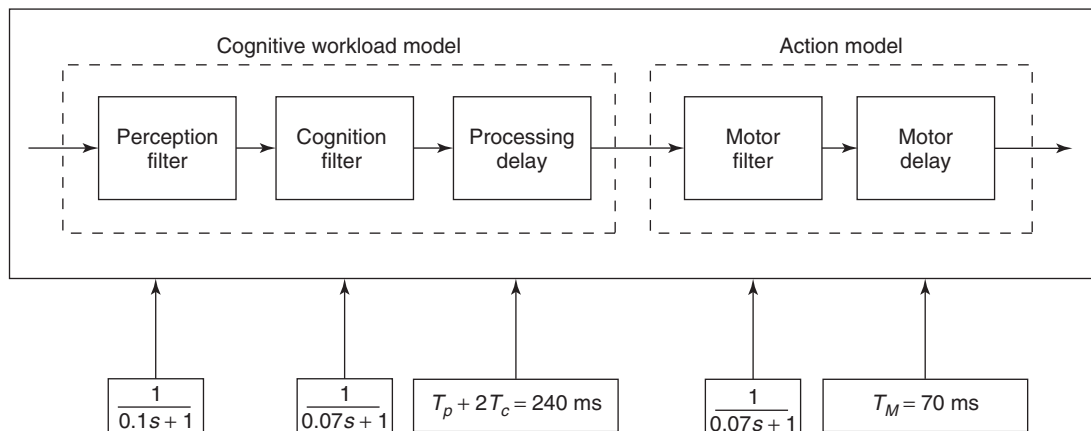


**Figure 4** Dynamic control system model of the human.

such as those described in Section 2, and use these data to dynamically adapt human–system interaction. The sensor dynamics of any future OFS gauges are still to be determined, so as a starting point such sensors are modeled here as simple first-order lags with a time constant $\tau = 1$ s. The sensor data would be used to dynamically change inputs to a user by directing instantiation of augmentation strategies, such as those described in Section 3. As an example, consider an application where workload is reduced via the *task delegation* augmentation strategy (Wickens et al., 1998). In such an application, using OFS gauges to detect cognitive overload (e.g., through a EEG-derived index of task engagement) (Prinzel et al., 2003), lower priority tasks would be offloaded to automated agents, with the goal of maintaining users working at their maximum capacity. Such a closed-loop human–system interaction model was implemented in the Matlab/Simulink simulation environment, which is illustrated in Figure 5.

Various pieces of a neuroergonomic system can be seen in the model in Figure 5, including the human perceptual, cognitive, and motor processors. Note both the OFS gauge that detects the state of the human user (i.e., cognitive work overload measurement) and the augmentation strategies [i.e., within the proportional–integral–derivative (PID) controller] that will alter the input to the human. The rest of the model contains task inputs to the system, displayed outputs at various points (e.g., actual vs. measured cognitive workload), and a simple model of performance errors resulting from cognitive overload. The feedback loop being closed is now apparent in this simulation model, which drives the need for a systematic control theory approach.

### 4.2 Controller Analysis and Design

Even this simple model has already produced some important findings. In particular, one major finding from initial efforts with the model is to show how dynamic instability can result from introducing feedback

within the system. That is to say that rapid detection of cognitive state under high workload might result in input being removed, which would reduce workload and hence information would be added, which would once again result in high workload, and the cycle repeats. This simple illustration indicates how users might find their display cycling rapidly through cluttered and decluttered states as a result of changes detected in workload. Control theory offers a means to remove such instability and optimize performance.

Figure 6 shows results from three simulations of a task overload situation. The input to each of these simulations is the same: Initially, the user is fully loaded (and making no errors), and then a step increase in workload is introduced 1 s into the simulation. This results in task overload from that point on, with subsequent performance errors. Note that each of these simulations uses the same system model, so the only difference is how (or if) the feedback control (i.e., augmentation strategy) is applied.

Starting from the left, the first plot of Figure 6 shows the resulting performance errors for an open-loop simulation (i.e., with the neuroergonomic system disabled). As the workload of the task increases, the plot shows how the number of errors quickly rises to a certain level and stays there. The next panel shows a poorly designed neuroergonomic system. This system utilizes simple proportional control; that is, the control action $c(t)$ that reduces task workload to the user is just directly proportional to measured overload $m(t)$. Thus, the controller is of the form

$$c(t) = Km(t)$$

and the control designer simply chooses the proportionality constant or controller gain $K$, which determines when an augmentation strategy (i.e., task delegation in this case) is to be implemented. High-gain controllers, with large $K$ values, use a high-magnitude feedback
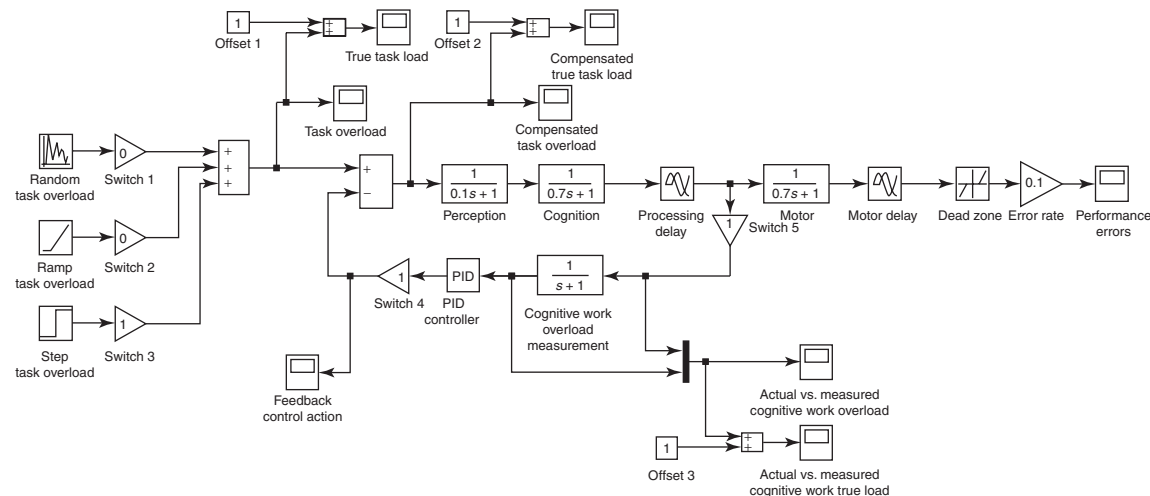


**Figure 5**  Matlab/Simulink model of closed-loop human–system interaction.
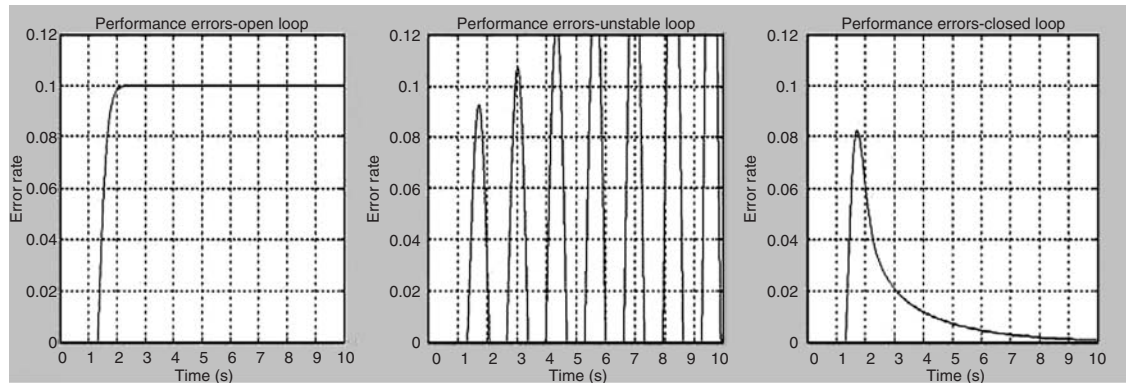
**Figure 6** Simulation results for a closed-loop dynamic model.

signal that tries aggressively to drive the control loop to the desired point (for fast or high performance). If $K$ is chosen too aggressively, however, the closed-loop system will approach (or even exceed) stability margins. In this example, the gain $K$ is chosen poorly, resulting in instability of the type described above, with the input being reduced rapidly and then increased, resulting in highly fluctuating performance from the user. Note that the precise values of $K$ that drive the system into instability depend on the specific problem (and can be predicted accurately with control theory methods), but they can certainly occur at plausible real-world values (in this example, $K = 2.8$).

Proportional control is what people often think of when they consider feedback. A simple version is the cruise control in a car, which moves the gas pedal in a manner proportional to the difference between the desired and actual speeds. However, this simple control strategy can deliver only limited performance improvements, even when designed correctly. For instance, one could never get steady-state errors down to zero with this type of control. This approach is limited because it utilizes the same gain for all frequencies (and hence all signals), so one does not have sufficient degrees of freedom to exploit any trade-offs in the design. A very common type of controller used in engineering applications is the PID controller. This generates a corrective action from a measurement of the form

$$c(t) = K_P m(t) + K_I \int_0^t m(\tau)d\tau + K_D \frac{dm(t)}{dt}$$

There are now three constants to be chosen (designed), $K_P, K_I$, and $K_D$, which correspond, respectively, to the amounts of proportional, integral, and derivative feedback used in the closed loop. Note that the integral action effectively includes memory and thus allows better compensation at low frequency and hence improved steady-state performance. The derivative action essentially includes anticipation, which allows for improved high-frequency performance, resulting in better transient response and improved stability properties. The overall controller has frequency-varying

gain, which allows design trade-offs to be exploited more properly. The right panel of Figure 6 shows a functional closed-loop system using a well-designed PID controller to deliver closed-loop stability and good performance. It is clear that even maximum errors never reach the level of the open-loop (automation-free) system and that they quickly drop to minimal levels (asymptotically approaching zero) without any undesirable oscillatory transient response.

### 4.2.1 Human Dynamics and Achievable Performance

The first benefit of the modeling approach described above is that it provides some *proof of concept* for the neuroergonomic concept: namely, to show precisely how an integrated system of OFS gauges, augmentation strategies, and robust controllers can combine to augment performance. The caveat from this work, however, is to note that such systems need to be designed carefully, with a systematic control theory approach rather than simple heuristic tuning, else neuroergonomics may fail to fulfill its potential.

Fortunately, systematic modeling can offer assistance in terms of determining the nature of information required and parameters necessary for driving specific OFS gauges. The types of questions that could be addressed by this type of analysis include:

- What time constant/bandwidth is necessary for a particular OFS gauge to have a significant useful effect (i.e., how fast)?
- What resolution is required of the OFS gauge (i.e., how accurate)?
- How much noise can reasonably be tolerated on any given measurement?
- What would additional measurements/gauges offer?
- What performance level could be achieved (given the above)?

These questions should be addressed in future work in the area of modeling and analyses. Note that both

qualitative and quantitative analyses can be carried out, and both have their uses (e.g., qualitative analysis might steer one toward a particular technology, whereas quantitative analysis might allow one to design and implement it accurately). Note also that specific scenarios can be carried out in a simulation, which would allow one to test out certain strategies repeatedly and reliably before going to the expense of constructing an experimental setup, including low-probability events that might not occur in an experimental setting. Furthermore, control theory includes powerful analysis tools that go well beyond simple simulation to address fundamental trade-offs and limitations inherent in any feedback loop (Doyle et al., 1992).

Ultimately, the modeling strategies described in this section would aim to predict the impact on human–system performance of various augmentation strategies for changing how information is provided to a user. In addition, they have the potential to highlight areas that would receive particular benefit from such augmentation. Thus, overall, this work can provide the basis for future systematic closed-loop analysis and controller design, bringing to bear powerful tools from engineering control theory. The power of such analysis tools is demonstrated in the next section.

### 4.2.2 Individual Differences and Robustness Analysis

Preliminary robustness analysis was conducted on the closed-loop system with the PID controller modeled in Figure 5. The theory of robust control deals with systems subject to uncertainty such as any closed-loop neuroergonomic system would be subject to due to individual differences (Parasuraman, 2011; Stevens et al., 2007) (among other reasons, as noted above). Control theory provides a means of examining what performance on such a system will be, rather than just an idealized simulation. It also allows for examination of variation between users, since relevant parameters in the model can be varied (e.g., speed of the MHP processors) to determine to what extent a given control scheme is robust against such variations.

The theoretical tools used here to model individual differences were based on the structured singular value (SSV), or $\mu$, and its extensions to handle real parametric uncertainty (Young, 2001). The idea is that one first has to use linear fractional transformations (LFTs) to rearrange the problem into canonical $M - \Delta$ form, as illustrated in Figure 7. Here $M(s)$ collects all the known dynamics of the (closed-loop) system, and $\Delta$ is a (block) diagonal structured perturbation, which in the case of individual differences analysis will consist of real parametric uncertainty representing variation in the parameters of the model. Thus, this approach handles LFT (block diagram) perturbations rather than handing perturbed coefficients directly in a (transfer function) model, but this apparent limitation is readily overcome, as we illustrate below.

Each differences analysis considered variations in two time constants (i.e., speed of the perceptual and cognitive processors). These could arise due to variations among users but could also be introduced through
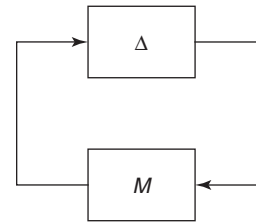


**Figure 7**   Canonical form for SSV analysis.
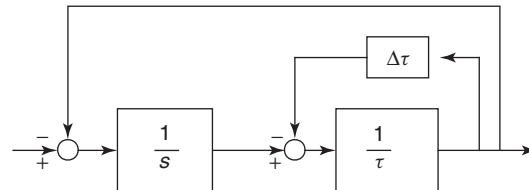


**Figure 8**   Variation in time constant as an LFT perturbation.

inaccuracies in the modeling approach. To realize this analysis, these variations were cast as a block diagram perturbation. This can be done by noting the interconnect in Figure 8, which shows an example of rearranging parametric uncertainty as an LFT (block diagram).

Mathematically straightforward block diagram calculations now reveal that the transfer function in Figure 8 is represented by

$$\frac{1}{1 + s(\tau + \Delta\tau)}$$

so that the block diagram perturbation in Figure 8 actually becomes a perturbed coefficient in the transfer function model, and to be specific it represents a perturbation in the time constant of the first-order lag model.

This approach was applied to the closed-loop neuroergonomic system model represented in Figure 5, considering a parametric variation in the time constants of the first-order lag models of the perceptual and cognitive processor blocks. Note that the motor processor block is not in the feedback loop in this scenario, so it does not affect stability and hence was not included in the robustness analysis presented here. The LFT and $\mu$ analysis machinery could then be applied to this block diagram. The mathematics of this approach is quite involved and uses computational complexity theory, complex analysis, and linear algebra among others (Young, 2001). Space constraints prevent going into any kind of detailed explanation here, but the end result of this analysis was to give a parameter range over which (robust) stability is *guaranteed*. This means that no parameter combination in the allowed range can cause instability. For example, in this case one could guarantee that no person with a combination of processing time constants for the perceptual and cognitive processors in the ranges specified would cause the closed-loop neuroergonomic system in

Figure 5 to go unstable. It is important to note the power of this *guarantee*, because one cannot get such guarantees from any amount of exhaustive simulation or testing (it is always possible that a parameter combination is missed, which causes a problem no matter how many variations are tried).

The results of this analysis showed that both the perceptual and cognitive processor time constants could be reduced to very small numbers (practically all the way down to zero), indicating that faster user response than predicted was no problem. The upper limits for the perceptual and cognitive processor time constants were found to be 3.7 and 2.6 s, respectively. Thus, it is possible for the system to go unstable with slower users. However, the degree of robustness afforded by a PID controller is huge. Specifically, in this example the time constants are a factor of more than 37 times greater than those nominally assumed (e.g., the MHP's "slowman" to "fastman" range for the perceptual processor is 150 ms and for the cognitive processor is 145 ms) (Card et al., 1983), meaning that tremendous variability in the perceptual and cognitive processor time constants can be tolerated between users and tasks. These robustness analysis results were also confirmed by a simulation model which showed stable behavior for all parameter variations in the range allowed but which could be driven unstable by parameter combinations outside these ranges (Young et al., 2004). This individual-differences analysis serves to illustrate what could be done when a neuroergonomic system is coupled with a systematic control-theoretic approach.

### 4.2.3 Robust Controller Synthesis

The control theory methods reviewed above can facilitate the design of high-performance closed-loop systems, even for systems whose dynamics are only partially known (Packard and Doyle, 1993; Zhou et al., 1996; Young, 2001). This is not achieved by optimizing nominal performance measures as in classical optimal control techniques such as linear quadratic Gaussian/linear quadratic regulator control (Ogata, 2002). Rather, these new approaches attempt to optimize robust performance measures utilizing techniques such as $\mu$-synthesis (Packard and Doyle, 1993). In this way, systems can be designed which are insensitive (or robust) to variations in the system that are naturally occurring but hard to predict a priori (e.g., differences between users). The mathematical machinery underlying such techniques is quite involved, and the associated optimization problems can be nonconvex and even NP hard. At first sight, such problems may appear to be intractable, and indeed, global minima usually cannot be guaranteed. Nevertheless, practical computation schemes have been developed using approximation schemes such as upper and lower bounds. These schemes are capable of finding very good approximate solutions in a reasonable amount of time. Moreover, there are numerically efficient implementations available of the associated algorithms, usually in convenient Matlab form (Balas et al., 1991), so such designs can readily be carried out (with the appropriate software) in a reasonable time using current computer hardware.

All this adds up to the fact that developers of neuroergonomic systems have at their disposal a number of powerful tools for robust controller analysis and synthesis. These theoretical techniques offer the potential of safely optimizing performance in a neuroergonomic system while maintaining guaranteed closed-loop stability.

## 5 APPLICATION DOMAINS

As with any scientific discovery or technical innovation, there are multiple paths upon which technology components will advance. Identifying specific applications at the dawn of a significant advance in our understanding of a field of study is problematic in that the assumptions on which hypothesized applications are based are very likely to be flawed. The assumptions that must be made (and that are likely to be wrong) include:

- Who will take advantage of emergent technology components?
- What components of the emergent technology will ultimately prove most useful and robust?
- Where will the emergent technology components be found to be most useful?
- When will the various components of the emergent technology be validated sufficiently for incorporation into real-world systems?
- Why will the emergent technology components be seen as beneficial?
- How will the emergent technology components be used?

Given the challenges inherent in answering these questions, a good starting point is to identify potential general application domains and then extract examples from these domains in hopes of describing potential uses of emergent technology components. The general application domains likely to be affected most by neuroergonomic technology components include operational domains that involve massive data and/or vigilance and thus would benefit from real-time cognitive readiness and assessment capabilities; educational domains, such as a scenario-based training system that can adapt in real time to trainee performance, as assessed by both overt behavior and cognitive state (as captured by OFS gauges); and clinical domains such as medical applications, where, for example, the real-time attention processes of children with attention-deficit hyperactivity disorder (ADHD) are monitored and reinforcement interventions for "paying attention" are employed. The value of neuroergonomics to the operational, educational, and clinical application domains should be noted, but the specific examples reviewed below should not be assumed to be predictions of actual application areas.

### 5.1 Operational

A primary objective of neuroergonomics is to use real-time assessment of OFS to optimize cognitive and physical performance during operational activities. This will be particularly important in high-complexity,

information-rich operational environments, such as command and control or stock market analysis, or high-vigilance task environments, such as baggage screening, long-haul truck driving, or power plant operations. The latter can allow for predicting physiologically based variations in sleepiness versus alertness during operational performance and mitigating the associated risks (Warm et al., 2008).

Early examples of neuroergonomics in operational environments focused on adaptive automation, where an operator's state was monitored and when overload was detected, tasks were offloaded. For example, Pope et al. (1995) developed a closed-loop system that leveraged an EEG-based index of task engagement to determine optimal task allocation schemes. Freeman et al. (1999) extended this work by incorporating absolute measures of EEG task engagement, monitoring operator performance, and switching the task to an automatic mode during periods of high task engagement. Prinzel et al. (2003) were one of the first to use multiple physiological indicators (i.e., EEG, ERPs, HRV) to direct task automation in a closed-loop system. Wilson and Russell (2004) demonstrated marked improvements in operator performance by reducing task demands during periods of high operator workload. Similarly, Wilson and Russell (2007) provided physiologically adaptive aiding (i.e., automating some critical task functions) during unmanned vehicle operations and demonstrated a 25% performance improvement as compared to unaided performance. In addition, Berka et al. (2004) derived task allocation schemes between humans and autonomous agents in a naval command-and-control simulation environment by using real-time EEG-based assessment of alertness and memory states. Shaw et al. (2010) tried to use transcranial Doppler sonography (TCD), which measures cerebral blood flow velocity, to trigger adaptive automation in a command-and-control task; it was found to be effective in gauging only the first task workload transitions but not subsequent transitions. Thus, TCD's effectiveness in supporting physiologically driven adaptive automation is uncertain.

Beyond task allocation schemes, neuroergonomics has also been applied to image analysis. For example, ERP has been used to classify when targets are detected, missed, correctly rejected, and/or responded to with false alarms (Cowell, et al., 2007; Fuchs et al., 2007). Similarly, Mathan et al. (2006, 2008) used ERPs and overt physical responses to detect targets within satellite images to achieve a fivefold reduction in target detection time at high accuracy levels, as compared with conventional broad area image analysis. In addition, Touryan et al. (2010) used individualized EEG classifier models to distinguish between targets and clutter in visual images. In addition, real-time measures of cognitive states have been associated with performance while flying an aircraft (Schnell et al., 2006, 2008a, 2008b). Numerous studies in flight and other operational domains have shown that psychophysiological data provide accurate measures of operator fatigue and mental workload (Wilson and Schlegel, 2003; Wilson, 2001, 2002b). Neuroergonomics has also been conceptualized for use in decision support systems (Carroll et al.,

2010a), for air traffic control (Ayaz et al., 2011), and for mitigating the effects of stress on human performance (Hancock and Szalma, 2007; Haufler and Hatfield, 2010).

Neuroergonomic applications extend to the physical domain, as well, where sensing equipment provides a means to extend the manner in which an individual interacts with computing technology by facilitating "direct interaction"... According to Bonadio et al. (2002, p. 181):

> The communication among people and between people and machines or tools has not been fully realized because of the indirect interactions. The external tools need to be manipulated as an independent extension of one's body in order to achieve the desired goal. If machines and devices could be incorporated into the "neural space" as an extension of one's muscles or senses, they could lead to unprecedented augmentation in human sensory, motor, cognitive, and communication performance.

There are many such examples of physically based neuroergonomic applications. For example, brain potentials have been coupled with external devices controlled by the physically handicapped (Farwell and Donchin, 1988), including individuals with little or no motor function (Pfurtscheller et al., 2000). In addition, Felton et al. (2005) developed an electrocorticogram-controlled brain–computer interface that allowed paraplegics to compose letters on a computer. Other such brain–computer interfaces have been used to operate voice synthesizers and move robotic arms (Birbaumer, 2006; Mussa-Ivaldi et al., 2007). Thus, in the psychomotor domain, neuroergonomics expands the available response modalities by allowing for direct brain–computer interaction.

In general, from an operational perspective, neuroergonomics promises to support human performance in such a way that substantially more of the human potential is tapped. However, despite the apparent advantages of physiologically aided operational performance, the interactive relationship between operator and aiding system is a complex one that evolves over time and thus must be supple in design to achieve maximal effectiveness (Christensen et al., 2010a).

## 5.2 Educational

Learning may benefit greatly from the application of neuroergonomics. Specifically, as an individual acquires a new skill, there are associated functional changes in the brain that occur which can be monitored in real time by OFS physiological gauges and then precisely adapted to. Understanding these brain changes could provide a means to achieve "precision training"—where an individual is provided with precisely the right material at precisely the right time in precisely the right format based on an individual's current proficiency level. Such a system would identify a person's current level of expertise and would allow the person to be guided rapidly to heightened levels of sustained performance in a context-independent fashion. Additionally, a person's cognitive performance during training could be

periodically or continuously assessed to ensure that their training was proceeding appropriately. These techniques could be incorporated into a multilevel approach to training, one that capitalizes on being able to observe patterns both at the overt behavioral level and at a deeper structure neuroimaging level, thereby distinguishing between novice and expert behaviors and tracking progression toward "expert" neural activation over time. An application that could characterize expert performance, identify where in the novice–expert continuum a trainee's performance lies, and then mold the trainee's patterns to more closely reflect an expert's would revolutionize training (Stanney et al., 2010).

In terms of educational application examples, neuroergonomics has been integrated into perceptual skills training (Carroll, 2010; Carroll et al., 2008, 2009), instrument flight panel training (Carroll et al., 2010b), cultural communications skills training (Oden et al., 2010; Palmer and Kobus, 2010), and baggage screening training (Carpenter et al., 2010) as well as conceptualized for use in unmanned air vehicle (UAV) operations training (Baldwin et al., 2010; Fuchs et al., 2008; Sibley et al., 2010) and forward air controller training (Fuchs et al., 2009) and for assessing individual differences in human performance to support the development of better selection and training methodologies (Parasuraman & Jiang, 2011), among other examples. Stevens et al. (2007) have pointed out that individual differences in EEG patterns during learning will require incorporation of the individual student's profile into augmentation algorithms.

The use of such physiologically-based adaptive training systems is expected to substantially improve the efficiency and effectiveness of simulation-based training. One example demonstrated an approximate 20% increase in search performance effectiveness when trainees were presented with both expert visual scan patterns and scan feedback strategy based on their own visual scan patterns during search and detect task training (Carroll et al., 2010c). In general, neuroergonomic training applications can be expected to provide a reduction in training cycle time and increased quality of training outcome of at least 30% each [Fletcher (2001) suggests that, compared to conventional classroom instruction, instructional technology reduces time to reach learning objectives by about one-third and increases trainee achievement by about one-third; also see Woolf and Regian (2000)]. Thus, the impacts of neurergonomics on education could include more efficient and effective training, with the anticipated result being substantially deeper learning and more rapid attainment of expertise.

### 5.3 Clinical

Within the clinical domain one can imagine that, by leveraging neuroergonomic technology, clinicians could be better trained in clinical decision making related to the diagnosis and treatment of medical disorders (Rizzo and Parasuraman, 2007). For example, neuroergonomic applications have been conceptualized for use in error monitoring during medical diagnosis (Fedota and Parasuraman, 2010), epilepsy and sleep disorder diagnosis (Casson et al., 2010), mild traumatic brain injury treatment (Stanney et al., 2009), and stroke rehabilitation (Moller and Mikulis, 2007) and as neuroprostheses for individuals with various disabilities (Riener, 2007). The target identification research discussed in Section 5.1 suggests that neuroergonomic procedures could be useful for detecting cancers in diagnostic images. Such levels of improved target identification would be highly useful in medical diagnostics. The possible utility of neuroergomonic procedures in helping paralyzed patients was also mentioned earlier (Karwowski et al., 2003; Parasuraman and Wilson, 2008). To date, developments of emergent neuroergonomic technology components have been least aligned with such potential clinical applications. However, successes in the operational and training domains will likely accelerate application developments in the clinical domain.

In summary, applications of neuroergonomics are in their infancy. Examples of possible applications of underlying technology components can readily be imagined, and some are starting to emerge, mostly in the training community. There is significant evidence to suggest that the technology components are ready for insertion into mature applications and that the operational, educational, and clinical domains have capability gaps that call for technology solutions offered by the field of neuroergonomics. Thus, although mature neuroergonomic science and technology components have now embarked on the path toward application, the only certainty along this journey is that the applications developed will be like no others that have come before them.

## 6 CONCLUSIONS

Neuroergonomics seeks to create a new level of communication between human and support system, where human brains and computing machines are tightly coupled, thereby achieving a partnership that surpasses the information-handling capacity of either entity alone. Such improvement in human–system capability is clearly a worthy goal, whether the context is clinical restoration of function, educational applications, market-based improvements in worker efficiency, or warfighting superiority. Neuroergonomics is an attempt to realize a revolutionary paradigm shift in interactive computing, not by optimizing the *friendliness* of connections between human and computer but by achieving a *symbiotic dyad* of silicon- and carbon-based enterprises, thereby achieving a neurosustainable environment that maximizes the human potential.

## 7 ACKNOWLEDGMENTS

## REFERENCES

Alvarez, G. A., and Cavanagh, P. (2004), "The Capacity of Visual Short-Term Memory Is Set Both by Visual Information Load and by Number of Objects," *Psychological Science*, Vol. 15, No. 2, pp. 106–111.

Atkinson, R. C., and Shiffrin, R. M. (1968), "Human Memory: A Proposed System and Its Control Processes," in *The Psychology of Learning and Motivation*, K. W. Spence and J. T. Spence, Eds., Academic, New York.

Atkinson, R. C., and Shiffrin, R. M. (1971), "The Control of Short Term Memory," *Scientific American*, Vol. 225, No. 2, pp. 82–90.

Averbach, E., and Coriell, A. S. (1961), "Short-Term Memory in Vision," *Bell System Technical Journal*, Vol. 40, pp. 309–328.

Ayaz, H., Willems, B., Bunce, S., Shewokis, P. A., Izzetoglu, K., Hah, S., Deshmukh, A. R., and Onarul, B. (2011), "Estimation of Cognitive Workload During Simulated Air Traffic Control Using Optical Brain Imaging Sensors," in *AC'11 Proceedings of the 6th International Conference on Foundations of Augmented Cognition: Directing the Future of Adaptive Systems*, D. Schmorrow and C. Fidopiastis, Eds., Springer-Verlag, Berlin, Heidelberg, pp. 549–558.

Baddeley, A. (1986), *Working Memory*, Oxford University Press, New York.

Baddeley, A. (1990), *Human Memory: Theory and Practice*, Allyn & Bacon, Boston.

Baddeley, A. (2000), "Short-Term and Working Memory," in *The Oxford Handbook of Memory*, E. Tulving and F. Craik, Eds., Oxford University Press, New York, pp. 77–92.

Baddeley, A. (2003), "Working Memory: Looking Back and Looking Forward," *Nature Reviews: Neuroscience*, Vol. 4, pp. 829–839.

Baddeley, A., and Logie, R. (1999), "Working Memory: The Multiple Component Model," in *Models of Working Memory*, A. Miyake and P. Shah, Eds., Cambridge University Press, New York, pp. 28–61.

Balas, G., Doyle, J., Glover, K., Packard, A., and Smith, R. (1991), *The μ Analysis and Synthesis Toolbox*, MathWorks and MUSYN, Natick, MA.

Baldwin, C., Coyne, J. T., Roberts, D. M., Barrow, J. H., Cole, A., Sibley, C., Taylor, B., and Buzzell, G. (2010), "Prestimulus Alpha as a Precursor to Errors in a UAV Target Orientation Detection Task," in *Advances in Understanding Human Performance: Neuroergonomics, Human Factors Design, and Special Populations*, T. Marek, W. Karwowski, and V. Rice, Ed., Taylor & Francis/CRC Press, Boca Raton, FL.

Balota, D. A., Cortese, M. J., Duchek, J. M., Adams, D., Roediger, H. L., McDermott, K. B., and Yerys, B. E. (1999), "Veridical and False Memories in Healthy Older Adults and in Dementia of the Alzheimer's Type," *Cognitive Neuropsychology*, Vol. 16, pp. 361–384.

Baron, S., Kruser, D. S., and Messick, B. (1990), *Quantitative Modeling of Human Performance in Complex, Dynamic Systems*, National Academy Press, Washington, DC.

Berka, C., Levendowski, D. J., Cvetinovic, M. M., Petrovic, M. M., Davis, G., Lumicao, M. N., Zivkovic, V. T., Popovic, M. V., and Olmstead, R. (2004), "Real-Time Analysis of EEG Indexes of Alertness, Cognition, and Memory Acquired With a Wireless EEG Headset," *International Journal of Human-Computer Interaction*, Vol. 17, pp. 151–170.

Birbaumer, N. (2006), "Breaking the Silence: Brain-Computer Interfaces (BCI) for Communication and Motor Control," *Psychophysiology*, Vol. 43, pp. 517–532.

Bliss, J. C., Crane, H. D., Mansfield, P. K., and Townsend, J. T. (1966), "Information Available in Brief Tactile Presentations," *Perception and Psychophysics*, Vol. 1, pp. 273–283.

Boer, E. R. (2001), "Behavioral Entropy as a Measure of Driving Performance," keynote address at Driving Assessment 2001, Aspen, CO, August 14–17, available: http://www.ppc.uiowa.edu/driving-assessment/2001/Summaries/Driving%20Assessment%20Papers/44_boer_edwin.pdf, accessed February 6, 2004.

Bonadio, J. et al. (2002), "Improving Human Health and Physical Capabilities: Theme C Summary," in *Converging Technologies for Improving Human Performance*, M. C. Roco and W. S. Bainbridge, Eds., NSF/DOC-sponsored report, available: http://www.wtec.org/ConvergingTechnologies/Report/NBIC_report.pdf, accessed September 28, 2010.

Byrne, E. A., and Parasuraman, R. (1996), "Psychophysiology and Adaptive Automation," *Biological Psychology*, Vol. 42, pp. 249–268.

Campbell, A. T., Choudhury, T., Hu, S., Lu, H., Mukerjee, M. K., Rabbi, M., and Raizada, R. D. S. (2010), "NeuroPhone: Brain-Mobile Phone Interface Using a Wireless EEG Headset," paper presented at MobiHeld 2010, New Delhi, India, August 30, 2010, available: http://sensorlab.cs.dartmouth.edu/pubs/neurophone.pdf, accessed 11/9/2010.

Cannon-Bowers, J. A., Salas, E., and Converse, S. A. (1993), "Shared Mental Models In Expert Team Decision Making," in *Current Issues in Individual and Group Decision Making*, N. J. Castellan, Jr., Ed., Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 221–246.

Card, S. K., Moran, T. P., and Newell, A. (1983), *The Psychology of the Human–Computer Interaction*, Lawrence Erlbaum Associates, Mahwah, NJ.

Carlesimo, G. A., Perri, R., Turriziani, P., Tomaiuolo, F., and Caltagirone, C. (2001), "Remembering What But Not Where: Independence of Spatial and Visual Working Memory in the Human Brain," *Cortex*, Vol. 37, pp. 519–537.

Carpenter, A., Fuchs, S., Whitlow, S., Fiore, S., and Hale, K. (2010), *Auto-Diagnostic Adaptive Precision Training for Baggage Screeners (Screen-ADAPT)*, final technical report delivered to the Department of Homeland Security Science and Technology Directorate, August 13, 2010.

Carroll, M., Milham, L., and Champney, R. (2009), "Military Observation: Perceptual Skills Training Strategies," in *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) Annual Meeting*, National Training Systems Association, Arlington, VA.

Carroll, M., Fuchs, S., Carpenter, A., Hale, K., and Abbott, R. G. (2010a), "Development of an Auto-Diagnostic Adaptive Precision Trainer for Decision Making (ADAPT-DM)," *International Test and Evaluation Association Journal*, Vol. 31, pp. 247–263.

Carroll, M., Fuchs, S., Hale, K., Dargue, B., and Buck, B. (2010b), "Advanced Training Evaluation System (ATES): Leveraging Neuro-Physiological Measurement to Individualize Training," in *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) Annual Meeting*, National Training Systems Association, Arlington, VA.

Carroll, M., Kokini, C., and Moss, J. (2010c), *Multi-Axis Performance Interpretation Tool (MAPIT) Training Effectiveness Evaluation Report*, report submitted under BAA Contract No. N00014-10-C-0091 to Office of Naval Research, Code 30, Arlington, VA.

Carroll, M. B. (2010), "Empirical Evaluation of the Effectiveness of Eye Tracking-Based Search Performance Diagnosis and Feedback Methods," Ph.D. Dissertation, University of Central Florida, Orlando, FL.

Carroll, M. B., Champney, R. K., Milham, L. M., Jones, D. L., Chang, D., and Martin, G. (2008), "Development of an Instructor Aid to Diagnose Performance," in *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) Annual Meeting*, National Training Systems Association, Arlington, VA.

Casini, L., and Macar, F. (1999), "Multiple Approaches to Investigate the Existence of an Internal Clock Using Attentional Resources," *Behavioural Processes*, Vol. 45, Nos. 1–3, pp. 73–85.

Casson, A. J., Yates, D. C., Smith, S. J. M., Duncan, J. S., and Rodriguez-Villegas, E. (2010), "Wearable Electroencephalogram: What Is It, Why Is It Needed, and What Does It Entail?" *IEEE Engineering in Medicine and Biology Magazine*, Vol. 29, No. 3, pp. 44–56.

Cheatham, P. G., and White, C. T. (1954), "Temporal Numerosity, III: Auditory Perception of Number," *Journal of Experimental Psychology*, Vol. 47, pp. 425–428.

Christensen, J. C., Davis, I. E., Estepp, J. R, Monnin, J., and Wilson, G. F. (2009), "Validation of a Dry Electrode System for EEG," paper presented at The 53rd[h] Annual Human Factors and Ergonomics Society Meeting, San Antonio, TX, October 19–23, 2009, pp. 1171–1175.

Christensen, J. C., Estepp, J. R., Davis, I. E., and Wilson, G. F. (2010a), "Learning Effects in Physiologically Activated Adaptive Aiding," in *Proceedings of the 3rd Applied Human Factors and Ergonomics (AHFE) International Conference 2010*, Miami, FL, July 17–20.

Christensen, J. C., Estepp, J. R., Monnin, J., and Wilson, G. F. (2010b), "Evaluation of a Dry Electrode System for Electroencephalography: Applications for Psychophysiological Cognitive Workload Assessment," paper presented at The 54th[h] Annual Human Factors and Ergonomics Society Meeting, September 27-October 1, 2010, San Francisco CA, pp. 210–214.

Cowan, N. (2001), "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity," *Behavioral and Brain Sciences*, Vol. 24, pp. 87–114.

Cowell, A., et al. (2007), "Construction and Validation of a Neurophysio-Technological Framework for Imagery Analysis," paper presented at the 12th International Conference of Human-Computer Interaction, Beijing, People's Republic of China, July 22–27.

Dahm, W. J. A. (2010), *Technology Horizons: A Vision for Air Force Science & Technology during 2010–2030*, Vol. 1, AF/ST-TR-10-01-PR 15 May 2010, available: http://www.af.mil/shared/media/document/AFD-100727-053.pdf.

Darwin, C. J., Turvey, M. T., and Crowder, R. G. (1972), "An Auditory Analogue of the Sperling Partial Report Procedure: Evidence for Brief Auditory Storage," *Cognitive Psychology*, Vol. 3, pp. 255–267.

Dearden, A., Harrison, M., and Wright, P. (2000), "Allocation of Function: Scenarios, Context and the Economics of Effort," *International Journal of Human–Machine Studies*, Vol. 52, pp. 289–318.

Debernard, S., Cathelain, S., Crevits, I., and Poulain, T. (2002), "AMANDA Project: Delegation of Tasks in the Air Traffic Control Domain," paper presented at the 5th International Conference on the Design of Cooperative Systems, COOP'02, Saint Raphael, France, June 4–7.

Depue, B. E., Banich, M. T., and Curran, T. (2006), "Suppression of Emotional and Non-Emotional Content in Memory: Effects of Repetition on Cognitive Control," *Psychological Science*, Vol. 17, pp. 441–447.

Dickson, B. T. (2005), "The Cognitive Cockpit—A Testbed for Augmented Cognition," in CD *Proceedings of the First International Conference on Augmented Cognition*, Vol. 11, Las Vegas, NV, pp. 479–488.

Downs, T., and Downs, H. (2004), "Hairy Situations: fNIR Technology and Novasol," paper presented at the DARPA PI Meeting, Augmented Cognition: Improving Warfighter Information Under Stress, Orlando, FL, January 6–8.

Doyle, J. C., Francis, B. A., and Tannenbaum, A. R. (1992), *Feedback Control Theory*, Macmillan, New York.

Engelbart, D. C. (1963), "A Conceptual Framework for the Augmentation of Man's Intellect," in *Vistas in Information Handling*, P. W. Howerton, Ed., Spartan Books, Washington, DC, pp. 1–29.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., and Conway, A. R. A. (1999), "Working Memory, Short-Term Memory and General Fluid Intelligence: A Latent Variable Approach," *Journal of Experimental Psychology: General*, Vol. 128, pp. 309–331.

European Telecommunications Standards Institute (ETSI) (2002), *Human factors (HF): Guidelines on the Multimodality of Icons, Symbols, and Pictograms*, Report No. ETSI EG 202 048 v 1.1.1 (2002–08), ETSI, Sophia Antipolis, France.

Fafrowicz, M., and Marek, T. (2007), "Quo Vadis, Neuroergonomics?" *Ergonomics*, Vol. 50, No. 11, pp. 1941–1949.

Farwell, L. A., and Donchin, E. (1988), "Talking Off the Top of Your Head: Toward a Mental Prosthesis Utilizing Event-Related Brain Potentials," *Electroencephalography and Clinical Neurophysiology,* Vol. 70, No. 6, pp. 510–523.

Fedota, J. R., and Parasuraman, R. (2010), "Neuroergonomics and Human Error," *Theoretical Issues in Ergonomics Science*, Vol. 11, No. 5, pp. 402–421.

Felton, E. A., Wilson, J. A., Radwin, R. G., Williams, J. C., and Garell, P. C. (2005), "Electrocorticogram-Controlled Brain-Computer Interfaces in Patients with Temporary Subdural Electrode Implants," *Neurosurgery*, Vol. 57, pp. 425.

Fernandez-Duque, D., and Johnson, M. L. (2002), "Cause and Effect Theories of Attention: The Role of Conceptual Metaphors," *Review of General Psychology*, Vol. 6, No. 2, pp. 153–165.

Feyereisen, P., and Van der Linden, M. (1997), "Immediate Memory for Different Kinds of Gestures in Younger and Older Adults," *Current Psychology of Cognition*, Vol. 16, pp. 519–533.

Fidopiastis, C. M., and Weiderhold, M. (2008), "Mindscape Retuning and Brain Reorganization with Hybrid Universes: The Future of Virtual Rehabilitation," in *The PSI Handbook of Virtual Environments for Training & Education: Developments for the Military and Beyond*, D. Schmorrow, J. Cohn, and D. Nicholson, Eds., Praeger Security International, Westport, CT, pp. 427–434.

Flach, J. M. (1999), "Beyond Error: The Language of Coordination and Stability," in *Human Performance and Ergonomics*, P. A. Hancock, Ed., Academic, San Diego, CA, pp. 109–128.

Fletcher, J. D. (2001), "What Do Sharable Instructional Objects Have To Do with Intelligent Tutoring Systems, and Vice Versa?" *International Journal of Cognitive Ergonomics*, Vol. 5, No. 3, pp. 317–333.

Freeman, F. G., Mikulka, P. J., Prinzel, L. J., and Scerbo, M. W. (1999), "Evaluation of an Adaptive Automation System Using Three EEG Indices with a Visual Tracking System," *Biological Psychology*, Vol. 50, pp. 61–76.

Frenchman, K. A. R., Fox, A. M., and Maybery, M. T. (2003), "The Hand Movement Test as a Tool in Neuropsychological Assessment: Interpretation within a Working Memory Theoretical Framework," *Journal of the International Neuropsychological Society*, Vol. 9, pp. 633–641.

Fuchs, S., Hale, K. S., Stanney, K. M., Juhnke, J., and Schmorrow, D. (2007), "Enhancing Mitigation in Augmented Cognition," *Journal of Cognitive Engineering and Decision Making*, Vol. 1, No. 3, pp. 309–326.

Fuchs, S., Jones, D.L., Stensrud, B. and Hale, K. S. (2008), *Integration of Psychophysiological and Performance Measures into an Adaptive Aiding System: Development of a "Proactive Aiding in Command and Control Environments" System (PACES)*, final technical report delivered to the Department of the Air Force under contract FA8650-08-M-6826, October 28, 2008.

Fuchs, S., Hale, K., Baskin, A., Lawson, M., and Whitlow, S. (2009), *Development of an Auto-Diagnostic Adaptive Precision Training for Perceptual Skills (ADAPT-PS) to Support Real-Time Simulation-Based Training and Adaptation*. final technical report delivered to the Office of Naval Research under contract N00014-09-M-0140, August 18, 2009.

Gerloff, C., Richard, J., Hadley, J., Schulman, A. E., Honda, M., and Hallett, M. (1998), "Functional Coupling and Regional Activation of Human Cortical Motor Areas during Simple, Internally Paced and Externally Paced Finger Movements," *Brain*, Vol. 121, pp. 1513–1531.

Gevins, A., Smith, M. E., Leong, H., McEvoy, L, Whitfield, S., Du, R., and Rush, G. (1998), "Monitoring Working Memory Load during Computer-Based Tasks with EEG Pattern Recognition Methods," *Human Factors*, Vol. 40, pp. 79–91.

Gratton, G., and Fabiani, M. (2001), "The Event-Related Optical Signal: A New Tool for Studying Brain Function," *International Journal of Psychophysiology*, Vol. 42, pp. 109–121.

Gratton, G., Kramer, A. F., and Fabiani, M. (2008), "Brain Sensors and Measures," in D. D. Schmorrow and K. M. Stanney, Eds., *Augmented Cognition: A Practitioner's Guide*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1–26.

Gundel, A., and Wilson, G. F. (1992), "Topographical Changes in the Ongoing EEG Related to the Difficulty of Mental Task," *Brain Topography*, Vol. 5, pp. 17–25.

Hale, K. S., Stanney, K. M., and Schmorrow, D. D. (2012), "Augmenting Cognition in HCI: 21st Century Adaptive System Science and Technology," in *Handbook of Human-Computer Interaction*, 3rd ed., J. Jacko and A. Sears, Eds., Taylor & Francis/CRC Press, Boca Raton, FL.

Hancock, P. A., and Szalma, J. L. (2003), "The Future of Neuroergonomics," *Theoretical Issues in Ergonomics Science*, Vol. 4, pp. 238–249.

Hancock, P. A., and Szalma, J. L. (2004), " On the Relevance of Qualitative Methods for Ergonomics," *Theoretical Issues in Ergonomics Science*, Vol. 5, pp. 499–506.

Hancock, P. A., and Szalma, J. L. (2007), "Stress and Neuroergonomics," in *Neuroergonomics: The Brain at Work*, R. Parasuraman and M. Rizzo, Eds., Oxford University Press, New York, pp. 195–206.

Hankins, T. C., and Wilson, G. F. (1998), "A Comparison of Heart Rate, Eye Activity, EEG and Subjective Measures of Pilot Mental Workload during Flight," *Aviation, Space and Environmental Medicine*, Vol. 69, pp. 360–367.

Harter, M. R. (1967), "Excitability and Cortical Scanning: A Review of Two Hypotheses of Central Intermittency in Perception," *Psychological Bulletin*, Vol. 68, pp. 47–58.

Hasher, L., Lustig, C., and Zacks, R. T. (2007), Inhibitory Mechanisms And The Control Of Attention. In A. Conway, C. Jarrold, M. Kane, A. Miyake, & J. Towse (Eds.), *Variation in working memory*, New York: Oxford University Press

Haufler, A. J., and Hatfield, B. D. (2010), Performance Under Pressure: A Cognitive Neuroscience Approach To Emotion Regulation, Psychomotor Performance And Stress. *Proceedings of the 3rd Applied Human Factors and Ergonomics (AHFE) International Conference 2010*, July 17–20, Miami, FL.

He, P, Wilson, G. F., and Russell, C. A. (2007), "Removal of Ocular Artifacts From The EEG: A Comparison Between Time-Domain Regression Method And Adaptive Filtering Method Using Simulated Data," *Medical and Biological Engineering and Computing,* Vol. 45, pp. 495–503.

Heilmeier, G. H. (1999), "1999 Woodruff Distinguished Lecture Transcript: From POTS to PANS.com—Transitions in the World of Telecommunications for the Late 20th Century and Beyond," available: http://www.me.gatech.edu/me/publicat/99trans.html, accessed September 15, 2004.

Hettinger, L. J., Branco, P., Encarnacao, M., and Bonato, P. (2003), "Neuroadaptive Technologies: Applying Neuroergonomics to the Design of Advanced Interfaces," *Theoretical Issues in Ergonomic Science*, Vol. 4, pp. 220–237.

Hoc, J. M. (2001), "Towards a Cognitive Approach to Human–Machine Cooperation in Dynamic Situations," *International Journal of Human–Computer Studies*, Vol. 54, pp. 509–540.

Hoc, J.-M. (2008), "Cognitive Ergonomics: A Multidisciplinary Venture," *Ergonomics*, Vol. 51, No. 1, pp. 71–75.

Hock, C., et al. (1997), "Decrease in Parietal Cerebral Hemoglobin Oxygenation during Performance of a Verbal Fluency Task: Inpatients with Alzheimer's Disease Monitored by Means of Near-Infrared Spectroscopy (NIRS), Correlation with Simultaneous rCBF-PET Measurements," *Brain Research*, Vol. 755, pp. 293–303.

Horvitz, E., Kadie, C. M., Paek, T., and Hovel, D. (2003), "Models of Attention in Computing and Communications: From Principles to Applications," *Communications of the ACM*, Vol. 46, No. 3, pp. 52–59.

Izzetoglu, K., Bunce, S., Onaral, B., Pourrezaei, K., and Chance, B. (2003), "Functional Optical Brain Imaging Using Near-Infrared during Cognitive Tasks," *International Journal of Human-Computer Interaction*, Vol. 17, No. 2, pp. 211–227.

Izzetoglu, M., et al. (2005), "Functional Near-Infrared Neuroimaging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 13, pp. 153–159.

Jacoby, L. L., Bishara, A. J., Hessels, S., and Toth, J. P. (2005), "Aging, Subjective Experience, and Cognitive Control: Dramatic False Remembering by Older Adults," *Journal of Experimental Psychology: General*, Vol. 134, pp. 131–148.

Jagacinski, R. J., and Flach, J. M. (2003), *Control Theory for Humans: Quantitative Approaches to Modeling Performance*, Lawrence Erlbaum Associates, Mahwah, NJ.

Jinks, A., and Laing, D. G. (1999), "Temporal Processing Reveals a Mechanism for Limiting the Capacity of Humans to Analyze Odor Mixtures," *Cognitive Brain Research*, Vol. 8, No. 3, pp. 311–325.

Jones, D., Ahmad, A., Stanney, K. M., and Bowers, C. (2004), "Multimodal, Multitask Interaction Design: A Follow Up Study to Challenge Unimodal Design Assumptions," paper presented at the 48th Annual Human Factors and Ergonomics Society Meeting, pp. 421–425, New Orleans, LA, September 20–24, 2004.

Karwowski, W., Siemionow, W., and Gielo-Perczak, K. (2003), "Physical Neuroergonomics: The Human Brain in Control of Physical Work Activities," *Theoretical Issues in Ergonomic Science*, Vol. 4, pp. 175–199.

Keebler, J. R., Ososky, S., Taylor, G. S., Sciarini, L., and Jentsch, F. (2010), "Neuroethics: Protecting the Private Brain," in *Proceedings of the 3rd Applied Human Factors and Ergonomics (AHFE) International Conference 2010*, Miami, FL, July 17–20.

Keller, T. A., Cowan, N., and Saults, J. S. (1995), "Can Auditory Memory for Tone Pitch Be Rehearsed?" *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 21, pp. 635–645.

King, C. C. (1991), "Fractal and Chaotic Dynamics in Nervous Systems," *Progress in Neurobiology*, Vol. 36, pp. 279–308, available: http://www.math.auckland.ac.nz/~king/Preprints/pdf/BrChaos.pdf, accessed February 11, 2003.

Kingsley, S. A., Sriram, S., Pollick, A., Caldwell, J., Pearce, F., and Sing, H. (2002), "Physiological Monitoring with High-Impedance Optical Electrodes (Photrodes™)," paper presented at the 23rd Annual Army Science Conference, Orlando, FL, December 2–5.

Kirsh, D. (1996), "Adapting the Environment Instead of Oneself," *Adaptive Behavior*, Vol. 4, pp. 415–452.

Klapp, S. T., and Netick, A. (1988), "Multiple Resources for Processing and Storage in Short-Term Working Memory," *Human Factors*, Vol. 30, pp. 617–632.

Kobus, D. A., Brown, C. M., Morrison, J. G., Kollmorgen, G., and Cornwall, R. (2005, October 31), *DARPA Improving Warfighter Information Intake under Stress—Augmented Cognition Phase II: The Concept Validation Experiment (CVE)*, DARPA/IPTO technical report submitted to CDR Dylan Schmorrow.

Kramer, A. F. (1991), "Physiological Measures of Mental Workload: A Review of Recent Progress," in *Multiple Task Performance*, D. Damos, Ed., Taylor & Francis, London, pp. 279–328.

Krizo, E. J., Wilson, G. F. and Russell, C. A. (2005), "NuWAM: The Development Of A Modular, Real-Time Augmented Cognition System," in *Proceedings of the 11th International Conference on Human–Computer Interaction*, Vol. 11, Foundations of Augmented Cognition, D.D. Schmorrow, Ed., Lawrence Erlbaum Associates, Mahwah, NJ.

Lachman, R., Lachman, J. L., and Butterfield, E. C. (1979), *Cognitive Psychology and Information Processing: An Introduction*, Lawrence Erlbaum Associates, Mahwah, NJ.

Laska, M., and Teubner, P. (1999), "Olfactory Discrimination Ability of Human Subjects for Ten Pairs of Enantiomers," *Chemical Senses*, Vol. 24, No. 2, pp. 161–170.

Lepsien, J., and Nobre, A. C. (2007), "Attentional Modulation of Object Representations in Working Memory," *Cerebral Cortex*, Vol. 17, pp. 2072–2083.

Licklider, L. C. R. (1960), "Man-Computer Symbiosis," *IRE Transactions on Human Factors in Electronics*, Vol. 1, pp. 4–11, available: http://www.memex.org/licklider.pdf.

Livermore, A., and Laing, D. G. (1996), "Influence of Training and Experience on the Perception of Multicomponent Odor Mixtures," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 22, No. 2, pp. 267–277.

Logan, G. D. (2003), "Executive Control of Thought and Action: In Search of the Wild Homunculus," *Current Directions in Psychological Science*, Vol. 12, pp. 45–48.

Luria, R., Sessa, P., Gotler, A., Jolicœur, P., and Dell'Acqua, R. (2010), "Visual Short-Term Memory Capacity for Simple and Complex Objects," *Journal of Cognitive Neuroscience*, Vol. 22, No. 3, pp. 496–512.

Lusted, H. S., and Knapp, R. B. (1996), "Controlling Computers with Neural Signals," *Scientific American*, Vol. 275, pp. 82–87.

Mahrer, P., and Miles, C. (2002), "Recognition Memory for Tactile Sequences," *Memory*, Vol. 10, No. 1, pp. 7–20.

Makovski, T., and Jiang, Y. V. (2007), "Distributing versus Focusing Attention in Visual Short-Term Memory," *Psychonomic Bulletin and Review*, Vol. 14, pp. 1072–1078.

Mathan, S., et al. (2006), "Neurotechnology for Image Analysis: Searching for Needles in Haystacks Efficiently," in *Foundations of Augmented Cognition*, 2nd ed., D. D. Schmorrow, K. M. Stanney, and L. M. Reeves, Eds., Strategic Analysis, Arlington, VA, pp. 3–11.

Mathan, S., et al. (2008), "Rapid Image Analysis Using Neural Signals," in *Conference on Human Factors in Computing Systems*, M. Czerwinski, A. Lund, and D. Tan, Eds., Association for Computing Machinery Press, New York, pp. 3309–3314.

Mathieu, J. E., Goodwin, G. F., Heffner, T. S., Salas, E., and Cannon-Bowers, J. A. (2000), "The Influence of Shared Mental Models on Team Process and Performance," *Journal of Applied Psychology*, Vol. 85, No. 2, pp. 273–283.

Matlin, M. W. (1998), *Cognition*, 4th ed., Hartcourt Brace, Fort Worth, TX.

Matsuda, M. (1998), "Visual Span of Detection and Recognition of a Kanji Character Embedded in a Horizontal Row of Random Hiragana Characters," *Japanese Psychological Research*, Vol. 40, pp. 125–133.

McCabe, D. P., Roediger, III, H. L., McDaniel, M. A., Balota, D. A., and Hambrick, D. Z. (2010), "The Relationship between Working Memory Capacity and Executive Functioning: Evidence for a Common Executive Attention Construct," *Neuropsychology*., Vol. 24, No. 2, pp. 222–243.

Mendez, M. F. (2001), "Visuospatial Deficits with Preserved Reading Ability in a Patient with Posterior Cortical Atrophy," *Cortex*, Vol. 37, pp. 539–547.

Miller, G. A. (1956), "The Magical Number Seven Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psychological Review*, Vol. 63, pp. 81–97.

Moller, H. J., and Mikulis, D. J. (2007), "Neurocognitive Disability, Stroke, and Surgery: A Role for Neuroergonomics?" *Journal of Psychosomatic Research*, Vol. 63, pp. 613–615.

Morrison, J. G., Kobus, D., and St. John, M. (2003), "DARPA Augmented Cognition Technology Integration Experiment (TIE)," available: http://www.tadmus.spawar. navy.mil/AugCog_Brief.pdf, accessed September 3, 2004,

Mussa-Ivaldi, F. A., Miller, L., Rymer, Z., and Weir, R. (2007), "Neural Engineering," in *Neuroergonomics: The Brain at Work*, R. Parasuraman and M. Rizzo, Eds., Oxford University Press, New York, pp. 293–314.

Neisser, U. (1967), *Cognitive Psychology*, Appleton-Century-Crofts, New York.

Norman, D. A., and Shallice, T. (1980), *"Attention to Action: Willed and Automatic Control of Behaviour,"* University of California San Diego CHIP Report 99; reprinted in *Cognitive Neuroscience: A Reader*, M. Gazzaniga, Ed., Blackwell, Oxford, 2000.

Oden, K. B., Tourville, S., and Hale, K. (2010), "Matching the Training to the Trainee: Applying Behavioral Neurophysiology for Adaptive Training," in *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2010*, National Training Systems Association, Arlington, VA.

Ogata, K. (2002), *Modern Control Engineering*, 4th ed., Prentice Hall, Upper Saddle River, NJ.

Packard, A. K., and Doyle, J. C. (1993), "The Complex Structured Singular Value," *Automatica*, Vol. 29, pp. 71–109.

Palmer, E., and Kobus, E. (2010), "Implications of Physiological Measures of Stress for Training Cross-Cultural Decision Making Skills," in *Proceedings of the 3rd Applied Human Factors and Ergonomics (AHFE) International Conference 2010*, Miami, FL, July 17–20.

Parasuraman, R. (2003), "Neuroergonomics: Research and Practice," *Theoretical Issues in Ergonomic Science*, Vol. 4, pp. 5–20

Parasuraman, R. (2009), "Assaying Individual Differences in Cognition with Molecular Genetics: Theory, Empirical Findings, and Applications," *Theoretical Issues in Ergonomics Science*, Vol, 10, pp. 399–416.

Parasuraman, R. (2011), "Neuroergonomics: Brain, Cognition, and Performance at Work," *Current Directions in Psychological Science*, Vol. 20, pp. 181–186.

Parasuraman, R., and Hancock, P.A. (2004), "Neuroergonomics: Harnessing the Power of Brain Science for HF/E," *HFES Bulletin*, Vol. 47, No. 12, pp. 1, 4.

Parasuraman, R., and Jiang, Y. (2011), "Individual Differences in Cognition, Affect, and Performance: Behavioral, Neuroimaging, and Molecular Genetic Approaches," *NeuroImage*, Vol. 59, No. 1, pp. 70–82.

Parasuraman, R., and Wilson, G. F. (2008), "Putting the Brain to Work: Neuroergonomics Past, Present, and Future," *Human Factors*, Vol. 50, No. 3, pp. 468–474.

Penney, C. G. (1989), "Modality Effects and the Structure of Short-Term Verbal Memory," *Memory and Cognition*, Vol. 17, pp. 398–422.

Persson J., and Reuter-Lorenz, P. A. (2008), "Gaining Control: Training Executive Function and Far Transfer of the Ability to Resolve Interference," *Psychological Science*, Vol. 19, No. 9, pp. 881–888.

Pfurtscheller, G., Guger, C., Muller, G., Krausz, G., and Neurper, C. (2000), "Brain Oscillations Control Hand Orthosis in a Tetraplegic," *Neuroscience Letters*, Vol. 292, pp. 211–214.

Phillips, C. L., and Parr, J. M. (1999), *Signals, Systems, and Transforms*, 2nd ed., Prentice Hall, Upper Saddle River, NJ.

Pickering, S. J. (2001), "Cognitive Approaches to the Fractionation of Visuospatial Working Memory," *Cortex*, Vol. 37, pp. 457–473.

Pope, A. T., Bogart, E. H., and Bartolome, D. S. (1995), "Biocybernetic System Validates Index of Operator Engagement in Automated Task," *Biological Psychology,* Vol. 40, pp. 187–195.

Posner, M. I., and DiGirolamo, G. J. (2000), "Cognitive Neuroscience: Origins and Promise," *Psychological Bulletin*, Vol. 126, pp. 873–889.

Posner, M. I., and Konick, A. F. (1966), "Short-Term Retention of Visual and Kinesthetic Information," *Organizational Behavior and Human Performance*, Vol. 1, pp. 71–88.

Pravitha, R., Sreenivasan, R., and Nampoori, V. P. N. (2003), "Complexity of Brain Dynamics Inferred from the Sample Entropy Analysis of Electroencephalogram," in *Proceedings of the National Conference on Nonlinear Systems and Dynamics*, Karagpur, West Bengal, India, December 28–30.

Prinzel, L. J., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., and Pope, A. T. (2000), "A Closed-Loop System for Examining Psychophysiological Measures for Adaptive Task Allocation," *International Journal of Aviation Psychology*, Vol. 10, pp. 393–410.

Prinzel, L. J., Parasuraman, R., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., and Pope, A. T. (2003), *Three Experiments Examining the Use of Electroencephalogram, Event-Related Potentials, and Heart Rate Variability for Real-Time Human-Centered Adaptive Automation Design*, NASA TP-2003-212442, NASA STI Program Office, Hanover, MD, available: http://www.techreports.larc.nasa.gov/ltrs/PDF/2003/tp/NASA-2003-tp212442.pdf, accessed February 9, 2004.

Raj, A. K., Perry, J. F., Abraham, L. J., and Rupert A. H. (2003), "Tactile Interfaces for Decision Making Support under High Workload Conditions," in *Proceedings of Aerospace Medical Association 74th Annual Scientific Meeting*, San Antonio, TX.

Rebolledo-Mendez, G.,  Dunwell, I., Martínez-Mirón, E. A., Vargas-Cerdán, M.D., de Freitas, S., Liarokapis, F., and

García-Gaona, A. R. (2009), "Assessing NeuroSky's Usability to Detect Attention Levels in an Assessment Exercise," in *Human-Computer Interaction, Part I, HCII 2009*, J. A. Jacko, Ed. (LNCS 5610, pp. 149–158), available: http://company.neurosky.com/files/rebolledo-mendez_2009.pdf, accessed November 9, 2010.

Rekha, M., Pravitha, R., Nampoori, V. P. N., and Sreenivasan, R. (2003), "Effect of Fatigue on Mental Performance: A Nonlinear Analysis," in *Proceedings of the National Conference on Nonlinear Systems and Dynamics*, Karagpur, West Bengal, India, December 28–30.

Riener, R. (2007), "Neurorehabilitation Robotics and Neuroprosthetics," in *Neuroergonomics: The Brain at Work*, R. Parasuraman and M. Rizzo, Eds., Oxford University Press, New York, pp. 346–359.

Rizzo, M., and Parasuraman, R. (2007), "Future Prospects for Neuroergonomics," in *Neuroergonomics: The Brain at Work*, R. Parasuraman and M. Rizzo, Eds., Oxford University Press, New York, pp. 381–388.

Roco, M. C., and Bainbridge, W. S., Eds. (2002), *Converging Technologies for Improving Human Performance*, NSF/DOC-sponsored report, available: http://www.wtec.org/ConvergingTechnologies/Report/NBIC_report.pdf, accessed September 28, 2010.

Samman, S. N., Stanney, K. M., Dalton, J., Ahmad, A., Bowers, C., and Sims, V. (2004), "Multimodal Interaction: Multi-Capacity Processing Beyond 7 +/- 2," in *Proceedings of the 48th Annual Human Factors and Ergonomics Society Meeting*, New Orleans, LA, September 20–24.

Sarno, K. J., and Wickens, C. D. (1995), "The Role of Multiple Resources in Predicting Time-Sharing Efficiency," *International Journal of Aviation Psychology*, Vol. 5, pp. 107–130.

Scerbo, M. W., Freeman, F. G., Mikulka, P. J., Parasuraman, R., Di Nocero, F., and Prinzel, L. J. (2001), *The Efficacy of Psychophysiological Measures for Implementing Adaptive Technology*, NASA TP-2001-211018, NASA Langley Research Center, Hampton, VA.

Schmorrow, D., and McBride, D. (2004), "Introduction," Special Issue on Augmented Cognition, *International Journal of Human-Computer Interaction*, Vol. 17, No. 2, pp. 127–130.

Schnell, T., et al. (2006), "Toward the Cognitive Cockpit: Flight Test Platforms and Methods for Monitoring Pilot Mental State," in *Foundations of Augmented Cognition*, 3rd ed., D. D. Schmorrow, K. M. Stanney, and L. M. Reeves, Eds., Strategic Analysis, Arlington VA, pp. 268–278.

Schnell, T., Keller, M., and Macuda, T. (2008a), "Sensor Integration to Characterize Operator State," in *Augmented Cognition: A Practitioner's Guide*, D. D. Schmorrow and K. M. Stanney, Eds., Human Factors and Ergonomics Society, Santa Monica, CA, pp. 41–74.

Schnell, T., Keller, M., and Poolman, P. (2008b), "Quality of Training Effectiveness Assessment (QTEA): A Neurophysiologically Based Method to Enhance Flight Training," paper presented at the Digital Avionics Systems Conference, IEEE/AIAA 27th, St. Paul, MN, October 26–30, 2008.

Shallice, T. (1982), "Specific Impairments of Planning," *Philosophical Transactions of the Royal Society London, Series B*, Vol. 298, pp. 199–209.

Shaw, T., Parasuraman, R., Guagliardo, L., and de Visser, E. (2010), "Towards Adaptive Automation: A Neuroergonomic Approach to Measuring Workload during a Command A Control Task," in *Proceedings of the 3rd Applied Human Factors and Ergonomics (AHFE) International Conference 2010*, Miami, FL, July 17–20.

Sibley, C., Cole, A., Gibson, G., Roberts, D., Barrow, J., Baldwin, C., and Coyne, J. (2010), "Adaptive Training in an Unmanned Aerial Vehicle: Examination of Several Candidate Real-Time Metrics," in *Proceedings of the 3rd Applied Human Factors and Ergonomics (AHFE) International Conference 2010*, Miami, FL, July 17–20.

Sligte, I. G., Vandenbroucke, A. R., Scholte, H., and Lamme, V. (2010), "Detailed Sensory Memory, Sloppy Working Memory," *Frontiers in Psychology*, Vol. 1, Article 175, pp. 1–10.

Smyth, M. M., and Pendleton, L. R. (1990), "Space and Movement in Working Memory," *Quarterly Journal of Experimental Psychology*, Vol. 42, pp. 291–304.

Sperling, G. (1960), "The Information Available in Brief Visual Presentations," *Psychological Monographs*, Vol. 74, No. 11, Whole No. 498.

Sperling, G. (1963), "A Model of Visual Memory Tasks," *Human Factors*, Vol. 5, No. 1, pp. 19–31.

Stanney, K. M. (in press), "Augmenting Human Performance," in *Neuroadaptive Systems: Research, Theory, and Applications*, M. Fafrowicz, T. Marek, W. Karwowski, and D. Schmorrow, Eds., Taylor & Francis/CRC Press, Boca Raton, Fl.

Stanney, K.M., Hale, K.S. and Jones, D. (2009), "Augmented Cognition Design Approaches For Treating Mild Traumatic Brain Injuries," in *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience*, D. Schmorrow, I.V. Estabrooke, and M. Grootjen, Eds., 5th International Conference, Held as Part of HCI International 2009, San Diego, CA, July 19–24, (16), pp. 800–809.

Stanney, K., Kokini, C., Fuchs, S., Axelsson, P., and Phillips, C. (2010), "Auto-Diagnostic Adaptive Precision Training—Human Terrain (ADAPT-HT): A Conceptual Framework for Cross-Cultural Skills Training," in *Proceedings of the 3rd Applied Human Factors and Ergonomics (AHFE) International Conference 2010*, Miami, FL, July 17–20.

Stanney, K., Samman, S., Reeves, L., Hale, K., Buff, W., Bowers, C., Goldiez, B., Nicholson, D., and Lackey, S. (2004), "A Paradigm Shift in Interactive Computing: Deriving Multimodal Design Principles from Behavioral and Neurological Foundations," *International Journal of Human-Computer Interaction*, Vol. 17, No. 2, pp. 229–257.

Sterman, M. B., and Mann, C. A. (1995), "Concepts and Applications of EEG Analysis in Aviation Performance Evaluation," *Biological Psychology*, Vol. 40, pp. 115–130.

Stevens, R. H., Galloway, T., and Berka, C. (2007), "EEG-Related Changes in Cognitive Workload, Engagement and Distraction as Students Acquire Problem Solving Skills," in *Proceedings of the 11th International Conference on User Modeling*.

Sullivan, E. V., and Turvey, M. T. (1974), "On the Short-Term Retention of Serial, Tactile Stimuli," *Memory and Cognition*, Vol. 2, pp. 600–606.

Sulzen, J. (2001), "Modality Based Working Memory," School of Education, Stanford University, available: http://ldt.stanford.edu/~jsulzen/james-sulzen-portfolio/classes/PSY205/modality-project/paper/modality-expt-paper.PDF, accessed February 5, 2003.

Takeuchi, Y. (1999), "Change in Blood Volume in Brain during a Simulated Aircraft Landing Task," *Journal of Occupational Health*, Vol. 42, pp. 60–65.

Tolja, J. (2010), *Neuro Ergonomics of Design*, available: http://bodyconsciousdesign.com/uploads/neuro_ergonomics_of_design_workshop.pdf, accessed October 15, 2010.

Touryan, J., Gibson, L., Horne, J. H., and Weber, P. (2010), "Real-Time Classification of Neural Signals Corresponding to the Detection of Targets In Video Imagery," in *Proceedings of the 3rd Applied Human Factors and Ergonomics (AHFE) International Conference 2010*, Miami, FL, July 17–20.

Trejo, L. J., Wheeler, K. R., Jorgensen, C. C., Rosipal, R., Clanton, S. T., Matthews, B., Hibbs, A. D., Matthews, R., and Krupka, M. (2003), "Multimodal Neuroelectric Interface Development," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 11, No. 2, pp. 199–204.

Tucker, D. M., and Luu, P. (2009), "Operational Brain Dynamics: Data Fusion Technology for Neurophysiological, Behavioral, and Scenario Context Information in Operational Environments," in *Augmented Cognition, HCII 2009*, D. D. Schmorrow et al., Eds., Lecture Notes in Artificial Intelligence, Vol. 5638, pp. 98–104.

Villringer, A., and Chance, B. (1997), "Non-Invasive Optical Spectroscopy and Imaging of Human Brain Function," *Trends in Neuroscience*, Vol. 20, pp. 435–442.

Warm, J. S., Parasuraman, R., and Matthews, G. (2008), "Vigilance Requires Hard Mental Work and Is Stressful," *Human Factors*, Vol. 50, No. 3, pp. 433–441.

Watkins, D. H., and Watkins, O. C. (1974), "A Tactile Suffix Effect," *Memory and Cognition*, Vol. 2, pp. 176–180.

Wickens, C. D. (1984), *Engineering Psychology and Human Performance*, HarperCollins, New York.

Wickens, C. D. (1992), *Engineering Psychology and Human Performance*, 2nd ed., HarperCollins, New York.

Wickens, C. D. (2002), "Multiple Resources and Performance Prediction," *Theoretical Issues in Ergonomics Scientific*, Vol. 3, No. 2, pp. 159–177.

Wickens, C. D., and Hollands, J. G. (2000), *Engineering Psychology and Human Performance*, 3rd ed., Prentice Hall, Upper Saddle River, NJ.

Wickens, C. D., Mavon, A. S., Parasuraman, R., and McGee, A. P. (1998), *The Future of Air Traffic Control: Human Operators and Automation*, National Academy Press, Washington, DC.

Wickens, C. D., and Hollands, J. G. (2000), *Engineering Psychology and Human Performance*, 3rd ed., Prentice Hall, Upper Saddle River, NJ.

Wiener, N. (1948), *Cybernetics, or Control and Communication in the Animal and the Machine*, MIT Press, Cambridge, MA.

Wildey, C., MacFarlane, D. L., Khan, B., Tian, F., Liu, H., and Alexandrakis, G. (2010), "Improved fNIRS Using a Novel Brush Optrode," poster presented at Frontiers in Optics 2010/Laser Science XXVI as part of the Joint FiO/LS Poster Session I (JTuA23), Rochester, NY, October 24–28.

Wilson, G. F. (2001), "In-Flight Psychophysiological Monitoring," in *Progress in Ambulatory Monitoring*, F. Fahrenberg and M. Myrtek, Eds., Hogrefe and Huber Publishers, Seattle, WA, pp. 435–454.

Wilson, G. F. (2002a), "Psychophysiological Test Methods and Procedures," in *Handbook of Human Factors Testing and Evaluation*, 2nd ed., S. G. Charlton and T. G. O'Brien, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 127–156.

Wilson, G. F. (2002b), "An Analysis of Mental Workload in Pilots during Flight Using Multiple Psychophysiological Measures," *International Journal of Aviation Psychology*, Vol. 12, pp. 3–18.

Wilson, G. F., and Eggemeier, F. T. (1991), "Physiological Measures of Workload in Multitask Environments," in *Multiple-Task Performance*, D. Damos, Ed., Taylor & Francis, London, pp. 329–360.

Wilson, G. F., and Russell, C. A. (2004), "Psychophysiologically Determined Adaptive Aiding in a Simulated UCAV Task", in *Human Performance, Situation Awareness, and Automation: Current Research and Trends*, D. A. Vincenzi, M. Mouloua, and P. A. Hancock, Eds., Erlbaum, Mahwah, NJ, pp. 200–204.

Wilson, G. F., and Russell, C. A. (2006), "Psychophysiologically versus Task Determined Adaptive Aiding Accomplishment," in *Foundations of Augmented Cognition*, 2nd ed., D. D. Schmorrow, K. M. Stanney, and L.M. Reeves, Eds., Strategic Analysis, Arlington, VA, pp. 201–207.

Wilson, G. F., and Russell, C. A. (2007), "Performance Enhancement in an Uninhabited Air Vehicle Task Using Psychophysiologically Determined Adaptive Aiding," *Human Factors*, Vol. 49, No. 6, pp. 1005–1018.

Wilson, G. F., and Schlegel, R. E., Eds. (2003), *Operator Functional State Assessment*, Final Report, RTO-TR-HFM-104, NATO, Paris.

Woodin, M. E., and Heil, J. (1996), "Skilled Motor Performance and Working Memory in Rowers: Body Patterns and Spatial Positions," *Quarterly Journal of Experimental Psychology*, Vol. 49, pp. 357–378.

Woolf, B. P., and Regian, J. W. (2000), "Knowledge-Based Training Systems and the Engineering of Instruction," in *Training and Retraining: A Handbook for Business, Industry, Government, and the Military*, S. Tobias and J. D. Fletcher, Eds., MacMillan, New York, pp. 339–356.

Young, P. M. (2001), "Structured Singular Value Approach for Systems with Parametric Uncertainty," *International Journal of Robust and Nonlinear Control*, Vol. 11, pp. 653–680.

Young, P. M., Clegg, B., and Smith, C. A. P. (2004), "Dynamic Models of Augmented Cognition," *International Journal of Human–Computer Interaction*, Vol. 17, No. 2, pp. 259–273.

Zakay, D., and Block, R. A. (2004), "Prospective and Retrospective Duration Judgments: An Executive Control Perspective," *Acta Neurobiolagiae Experimentalis*, Vol. 64, pp. 319–328.

Zhou, K., Doyle, J. C., and Glover, K. (1996), *Robust and Optimal Control*, Prentice Hall, Upper Saddle River, NJ.

# PART 7
# EVALUATION

# CHAPTER 38

# ACCIDENT AND INCIDENT INVESTIGATION

**Patrick G. Dempsey**
**National Institute for Occupational Safety & Health**
**Pittsburgh, Pennsylvania**

## 1  INTRODUCTION

Of all the applications and benefits of human factors found in a comprehensive handbook, determining how a system led to an accident, minor or catastrophic, is among the least proactive. The systems philosophy of human factors is particularly well suited to investigating accidents and incidents to determine the causative influences so that corrective actions can be taken and future systems designed to be safer. An example of this is commercial air travel, where the accident rate per miles flown has decreased substantially over time, in part due to the corrective actions implemented after accidents. The same can be said of highway safety, as fatality rates per mile driven continue to decrease.

The investigation of accidents can take on a variety of forms, and investigations vary in depth from little or no investigation to multiyear investigations costing millions of dollars. Perhaps the most rudimentary of these approaches are summaries of administrative databases used for insurance or regulatory purposes. Beyer (1928) used accident statistics to provide a sense of the seriousness of the burden of occupational injuries and illnesses, in some cases categorizing data due to the loss sources. Although Beyer's (1928) text (and the preceding first and second editions) may have been one of the first to address industrial safety, the topic of accident investigation was not covered explicitly. However, numerous safety innovations contained within clearly resulted from lessons learned from previous accidents.

At the Silver Jubilee Congress of the National Safety Council in October 1938, S. E. Whiting used a simulated confined-entry scenario of workers entering tanks to demonstrate how carbon dioxide in confined spaces can lead to death (Whiting, 1939). Although these were conceptually the opposite of an accident investigation, a key characteristic of the scenarios is that they went beyond descriptions of hardware, including safety equipment. The scenarios alluded to organizational factors such as inaction by bystanders because of concerns over who had decision authority and cognitive considerations such as flawed mental models of what caused workers to become unconscious. The complexities and interactions between system components were apparent in the descriptions. The realistic nature of the simulations suggests that actual cases were used to develop the scenarios.

Within a few decades, accident prevention and investigation began to take on a more sophisticated and systemic view of causation. Heinrich's (1959) accident sequence depicted by dominos acknowledged that the cause of accidents was multifactorial and involved a sequence of factors. The three factors requisite for an accident were ancestry and social environment, fault of the person, and unsafe act and/or mechanical or physical hazard. Although these components may seem dated relative to contemporary theories and philosophies of accident causation, a broader interpretation of ancestry and social environment to include organizational culture and personal factors would be consistent with more current thought. Unsafe acts continue to occur, although the other components would be replaced by more systemic views, such as how a particular system (humans, environment, and hardware) can lead to errors that ultimately culminate in an incident or accident. Similarly, a more contemporary approach would be to determine which organizational factors increase the propensity for unsafe acts.

On the morning of February 1, 2003, the Space Shuttle *Columbia*, launched by the U.S. National Aeronautics and Space Administration (NASA) a few weeks prior on January 16, broke apart during reentry to Earth's atmosphere, leading to the loss of the lives of the

seven astronauts aboard. A piece of insulating foam that was part of the thermal protection system had separated 81.9 s after launch on January 16. When superheated air penetrated the leading-edge insulation, which eventually melted the aluminum structure of the left wing, the weakening of the structure and increasing aerodynamic forces led to loss of control, wing failure, and eventually breakup of the *Columbia* [Columbia Accident Investigation Board (CAIB), 2003].

The CAIB oversaw an accident investigation process that involved a staff over 120 persons in conjunction with 400 NASA engineers lasting nearly seven months. Over 25,000 searchers worked on the ground to collect debris from the spacecraft (CAIB, 2003). Although the physical cause of the *Columbia* accident was attributed to the piece of insulating foam that separated shortly after launch, the CAIB's conclusions regarding the chain of events that led to the disaster reached much further back in time than the foam separation. Examples of the findings included the conclusion that NASA's safety culture had become reactive, complacent, and too optimistic. During the mission and after the foam strike was known, managers resisted new information; thus, communication within the organization was inhibited. Because there were numerous foam incidents during previous missions that did not result in problems, managers were conditioned to believe that foam strikes were maintenance issues to be solved after landing (CAIB, 2003). These and the other extensive findings illustrate just how intricately accident investigations sometimes need to be conducted and how much accident investigation has matured from earlier forms. This particular case also illustrates many of the concepts to be discussed, ranging from the more traditional fault tree analysis, which is more oriented toward hardware, to investigating organizational influences.

The goal of this chapter is to provide an overview of the numerous approaches to accident and incident investigation in occupational settings, covering a range of approaches that range from using administrative databases to more complex systems approaches. More attention will be given to human-centered approaches, as mismatchs between human capabilities and the demands posed by systems are often a key component of accidents. The type of investigation selected will vary depending on the frequency and severity of the incident(s) in question.

## 2  BASIC PRINCIPLES OF ACCIDENT INVESTIGATION

Although organizations should always strive to prevent accidents, a key component of a safety program is having the resources and procedures ready to respond to an accident if an accident should occur. Prior to an accident, it is critical that factors such as authority for the investigation be established. All of the accident investigation methodologies discussed later rely on details of the accident; thus, collecting the information in a timely and thorough manner is critical. Many organizations have accident report and investigation forms specifically developed to collect information about factors such as the demographics of the injured person(s), location,

and ambient conditions at the time of the accident. All employees should know to whom accidents should be reported, in addition to procedures for calling for emergency response from fire departments or emergency medical technicians when necessary.

An important question that often arises relates to who should conduct the investigation. Vincoli (1994) argues that the manager responsible for the employees involved should lead the investigation, as it is his or her responsibility to ensure the safety of the employees. The reasons for this suggestion is because management can, among other items, marshal resources for the investigation and obtain organizational support, define and implement corrective actions, resolve conflicts, and make employees aware of the outcome as well as conduct follow-up procedures to ensure that corrective actions have been taken (Vincoli, 1994). This does not exclude others from contributing, however, and site safety professionals or safety consultants can and should contribute to the process.

Depending on the severity of an accident, others such as insurance company representatives or government investigators may become involved. For example, the Occupational Safety and Health Administration (OSHA) in the United States must be notified whenever a fatality occurs. OSHA collects information on the individual (demographics, experience, nature of injury, etc.), accident data (measurements, photos, workplace layout, etc.), information on the equipment or process being used (machine type, manufacturer, model, warning devices, etc.), information on potential witnesses, and information on whether the employer has an active safety program and whether or not the program or any of its components address the type of accident that has occurred. Similarly, the Mine Safety and Health Administration (MSHA) in the United States requires that mines inform MSHA within 15 min when what are termed "immediately reportable accidents and injuries" such as fatalities, entrapment of an individual for more than 30 min, an unplanned ignition or explosion of dust or gas, or other types of accidents described in Title 30 § 50.2(h) of the U.S. Code of Federal Regulations.

The information collected should provide a sufficient description of the system, including the personnel, the machine/equipment being used, and the task. It is important to try to ascertain what goal the behavior carried out at the time of the accident was directed at, especially if a nonroutine task was being conducted or the operator was troubleshooting. The importance of the task is addressed in more detail in Section 3.3.

An important component of the investigation is recording the scene through photography or videography. Vincoli (1994) provides a comprehensive overview of what the accident photography kit should include. Aside from the camera, scales, rulers, and a perspective grid are recommended. Ability to preview video and still images at the scene can aid in assessing whether the scene has been captured adequately and geometric issues such as potential parallax problems can be assessed.

## 3  EPIDEMIOLOGICAL APPROACHES

Accident investigation is often thought of only as something conducted for catastrophic events, yet the

combined analysis of similar accidents, even if the accidents are somewhat minor, can be a powerful means of determining underlying causes of a class of accidents. Epidemiology, which is the study of the determinants and distribution of health-related states in populations (Last, 1995), is not typically discussed in the context of accident investigation. However, numerous principles of epidemiology are relevant to the study of common classes of accidents, such as injuries caused by a particular object (e.g., a chainsaw) or associated with a particular task (e.g., meat cutting). Several approaches to the study of injuries resulting from accidents are discussed to provide an overview of how epidemiology can contribute to understanding the causes of accidents. It should be noted that all of the methods discussed are not appropriate for investigating single accidents or incidents but rather are cases where a reasonable sample size is available for study. Although this may seem to be a disadvantage, a clear advantage is that the results of these studies can often be implemented and prevent multiple future incidents or accidents.

## 3.1 Administrative Data

Administrative data collected for other purposes, such as paying Workers' Compensation claims or tracking injuries and illnesses to satisfy regulatory requirements (e.g., OSHA Log of Work-Related Injuries and Illnesses—Form 300 used in the United States) have valuable information on the potential factors associated with the injuries and illnesses recorded. Often, these data have short narratives of the accident scenario and may have administrative codes describing the nature of injury (e.g., contusion, laceration, fracture) and cause (e.g., struck by/against, fall on same level).

As an example of the use of administrative data, Dempsey and Hashemi (1999) analyzed a large sample of Workers' Compensation claims attributed to manual materials handling to determine if the claims suggested specific areas for intervention or future research. Although the majority of the claims were due to musculoskeletal overexertion, a number of high-cost traumatic injuries were uncovered that would be better analyzed through one of the more detailed techniques for rarer events discussed later in the chapter. A large number of acute traumatic injuries to the upper extremity, such as lacerations and contusions, also suggested the need for more widespread use of basic personal protective equipment such as gloves.

Moore et al. (2009) extracted injuries classified as "fall from equipment" from the data reported by the MSHA for all injuries, illnesses, and fatalities during 2006 and 2007. They evaluated the circumstances leading to the fall from equipment injuries to develop research questions for future studies and to identify potential accident scenarios or patterns suggestive of preventive approaches. The majority of injuries identified occurred at surface mining facilities (~60%) with fractures and sprains/strains being the most common injuries occurring to the major joints of the body. Nearly 50% of injuries occurred during ingress/egress, predominately during egress, and approximately 25% of injuries occurred during maintenance tasks. The majority of

injuries occurred in relation to large trucks, wheel loaders, dozers, and conveyors/belts. From the data obtained in this study, several different research areas have been identified for future work, including balance and stability control when descending ladders and equipment design for maintenance tasks.

Another example of a study that utilized administrative data was the analysis of motor vehicle crashes in construction work zones by Sorock et al. (1996). The accident narrative fields from insurance claims were used as the basis of the study. Over 3600 claims were analyzed by categorizing the claims according to precrash activity (stopping, merging, cutting off, reversing, precrash error) and crash types (rear-end impact, hitting large object, hitting small object, side impact, overturning). Of the precrash actions classified, stopping was most common, although the group of claims associated with various precrash errors (e.g., lost control, asleep, failure to yield) had the largest mean and median costs.

One method worth mentioning is the discontinued American National Standards Institute (ANSI) standard, ANSI Z16.2, *Information Management for Occupational Safety and Health* (previously titled *Method of Recording Basic Facts Relating to the Nature and Occurrence of Work Injuries*). This method required collecting items such as nature of injury, part of body, source of injury, agency of accident, and unsafe act, which could then be coded using the coding system provided. Although these are not administrative data per se, older safety texts often mention using the coded data similarly to the suggestions for using Workers' Compensation claims that were discussed earlier.

Although administrative data can provide insights into accidents, the data can be of variable quality, particularly narratives. Since the data are not always for the purposes of gaining an understanding of accident causation, information will be missing. There may also be coding errors, especially for body part and nature-of-injury determinations. Often, the people coding these data do not have medical training; thus, the codes selected may represent the closest to what the coder believes to be correct. Little information is usually available to assess validity. More active approaches discussed throughout the remainder of the chapter can be utilized to overcome these limitations.

## 3.2 Case-Crossover Methodology

The case-crossover method is a study design used to investigate transient risk factors for discrete outcomes such as occupational injury. The method has been used to investigate risk factors for myocardial infarction (Mittleman et al., 1993) and more recently to investigate cell phone use and motor vehicle collisions (Redelmeier and Tibshirani, 1997). The method is based on determining what was different in the time period immediately prior to an accident that is different from "normal" conditions. The hazard period investigated varies with the nature and duration of the exposure being investigated (e.g., an object falling has a very short hazard period, whereas a sedating antihistamine may have effects for several hours) (Sorock et al., 2001a). For some types of accidents there can be multiple risk

factors; thus, the longest period needs to be considered. A key advantage of this approach is that subjects act as their own controls, avoiding the difficulties posed by, for example, finding suitable control subjects for a case–control study. Case–control studies can present difficulties when worker populations are limited, and the cases and controls need to be matched by demographics and exposures.

An example of the case-crossover method applied to occupational problems is the study of acute traumatic hand injury (Sorock et al., 2001b, 2004; Lombardi et al., 2003). Workers with acute traumatic hand injuries seeking treatment at several participating clinics were asked to participate in a telephone interview following treatment, preferably within a day of the accident. Volunteers were called and interviewed about eight transient exposures prior to their accident: using a machine, tool, or work material that performed differently than usual; performing an unusual task; using an unusual work method; being distracted or rushed; feeling ill; working overtime; and glove use at the time. The most important risks suggested were using a machine, tool, or work material that performed differently than usual, unusual work methods, and performing unusual tasks. This methodology is appropriate to use for similar types of accidents, and additional studies are currently underway examining other classes of occupational traumatic injuries, such as eye injuries. Kucera et al. (2008) applied the case-crossover design in a study of hand injuries in commercial fishing. While maintenance work was strongly associated with hand injury risk, gloves did not provide a protective effect, as was suggested by the results of Sorock et al. (2004).

### 3.3 Scenario Analysis

Drury and Brill's (1983) scenario analysis is a task analysis–based approach to accident investigation, developed for investigating consumer product accidents. The intention was to go beyond traditional accident investigation techniques to incorporate consideration of the task in addition to characteristics of the person, equipment, and environment. Since task analysis is the basis, uncovering the mismatch between the task demands and the limitations of the human body subsystem of interest is the goal. Although more descriptive than the analytical case-crossover method, the advantage is that the method allows for more open-ended data collection and is more rooted in ergonomics and human performance. Although the method is not rooted in epidemiological methods, there is a clear link, with the goal being to understand the underlying risk factors for accidents.

The scenario analysis approach is based on classifying accidents into hazard patterns (or scenarios), including a description of the victim, product, environment, and task. This approach was considered useful if no more than six hazard patterns describe more than 90% of the in-depth investigations. Once the generic hazard patterns are developed, a questionnaire to collect information is then developed. In addition to the information on the victim (e.g., age, sex, weight, body part injured), environment (e.g., indoor versus outdoor, lighting, weather conditions), and product (e.g., type, make, shape, weight), detailed information on task performance is collected. This includes information on the action intended prior to the accident; at the moment the task could not be completed as intended; at the moment the victim took a new, perhaps corrective action but before the injury occurred; and at the moment of the injury. The relationship between task demands and operator capacity can be assessed at each stage. The actual interview is somewhat longer than the interview typically used in conjunction with the case-crossover method, but this allows for more in-depth information to be collected.

In summary, Drury and Brill's (1983) approach uses archival data to generate hazard patterns, which then form the basis of data collection for future incidents. This is somewhat analogous to the case-crossover method discussed above in that knowledge of prior accidents or risk factors is necessary to formulate the interview for injured workers. The case-crossover method has the advantage of being more analytic, whereas the scenario analysis technique is more capable of uncovering human factors issues that lead to accidents and injury. Both methods were successful in increasing our understanding of the risk factors for the types of accidents studied.

## 4 SYSTEMS SAFETY TECHNIQUES

There are a number of systems safety techniques that can be utilized for proactive investigations of potential risks in a system to maximize reliability as well as for retrospective accident investigations. These methods sometimes encourage concentrating on hardware failures but are nevertheless useful components of accident investigations. Several of these techniques have been in existence for some time and have been refined considerably, the most common of which are discussed next.

### 4.1 Fault Tree Analysis

Fault tree analysis was developed at Bell Laboratories at the request of the U.S. Air Force due to concern over potential catastrophes associated with the Minuteman missile system being developed by Boeing (Hammer, 1985). Many accident investigators find the method particularly useful because it utilizes deductive logic (Vincoli, 1994), although like the technique discussed in Section 4.2, the method is not widely employed any longer due to the time requirements. The original intention was to develop a method that allowed probabilities of different potential sequences culminating in an accident to be estimated. If an accident probability is available, a risk assessment can be performed by multiplying the probabilities of various undesired events by their predicted costs. The question of the value of human life is often the most difficult question to answer, as this approach requires a common denominator across predictions.

Fault tree analysis is conducted through the use of Boolean logic. The top of the fault tree, in the shape of a

rectangle, represents the end effect under investigation, such as an accident. It should be noted that a safe state can be used as the top event to delineate the factors that need to occur to have a safe system. Symbols are then used to represent the different logic operators, including AND gates and OR gates. All possible sequences of events are then mapped out, and as the procedure becomes more complex, the shape of a tree sometimes becomes apparent. The events or system states that need to occur before failure are mapped out. If probabilities for each logic gate are available or can be estimated, probabilities for different branches can be estimated.

Seven different fault trees were constructed during investigation of the *Columbia* tragedy discussed earlier (CAIB, 2003). The number of elements in the fault trees ranged from 3 to 883, the latter illustrating how complex fault trees are when systems are complex. Although fault trees are not as widely used as they were in the past, they are still a rigorous tool for accident investigation, as evidenced by the *Columbia* investigation.

## 4.2 Management Oversight and Risk Tree

The management oversight and risk tree (MORT) technique has similarities to fault tree analysis in that it also uses Boolean logic in a graphical format. MORT can be used to assess the adequacy of safety program elements, or it can be used retrospectively to investigate accidents, in particular the management components that may have contributed to failure, for example, by creating conditions conducive to being complacent about safety or failure to correct previous safety issues. Rather than simply diagramming a physical system, MORT includes systems issues such as hazard review processes, assumed risks, and safety program review. An extensive discussion of MORT, including examples of completed trees, is provided by Johnson (1973).

## 5 SWISS CHEESE AND HUMAN FACTORS ANALYSIS AND CLASSIFICATION SYSTEM

Reason (1997) used a "Swiss cheese" metaphor to illustrate that the culmination of events in damage to humans or assets depends on failures at different levels. Holes in different layers of defenses that could allow penetration of accident trajectories led to the Swiss cheese metaphor. The holes are the result of what Reason calls either active failures or latent conditions. Reason (1997) defines *active failures* to be unsafe acts by personnel that are likely to have a direct impact on the safety of the system. *Latent conditions* arise from strategic and other top-level decisions (e.g., poor design, undetected manufacturing defects) that spread through organizations and affect the corporate culture. The CAIB (2003) report discussed earlier is an excellent reference for those wishing an in-depth view of latent failures identified during the *Columbia* investigation. Unfortunately, many industrial safety programs often fail to address or investigate these issues, due to an acute focus on procedures, maintenance, and similar factors.

Figure 1 illustrates several of the concepts associated with the Swiss cheese model. The box at the top illustrates the layered defenses concept and the notion that the accident trajectory can be stopped by different defenses or, alternatively, that different components of an organization need to be coordinated to prevent accidents. The triangle is the system that produces the accident, comprised of the personnel (unsafe acts), the workplace, and organizational factors. Causation is bottom up, whereas the accident investigation process is top down.

The Human Factors and Analysis and Classification System (HFACS) (Shappell and Wiegmann, 2001) puts into operational terms the concepts of active and latent failures from Reason's (1997) Swiss cheese model.
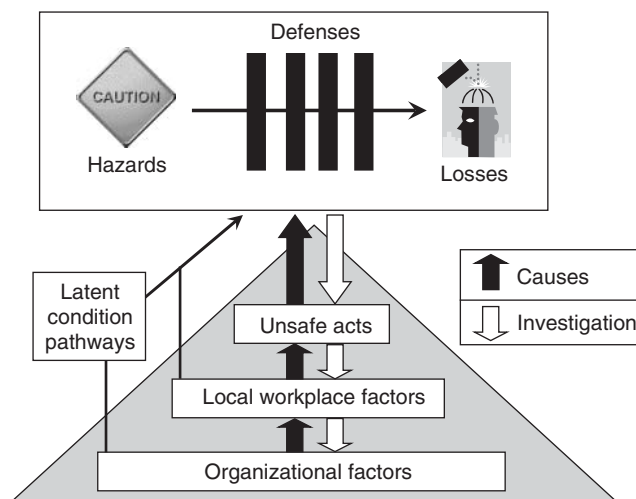


**Figure 1** Stages in the development and investigation of an organizational accident. (Adapted from Reason, 1997.)

Wiegmann and Shappell (2003) provide a comprehensive overview of HFACS, including illustrative case studies of previous accident investigations. HFACS was developed and refined by analyzing accident reports. Although the approach is discussed within the aviation context, some adaptation provides an excellent method that can be applied beyond aviation.

The "unsafe acts" portion of HFACS is broken down into errors and violation. Errors are further broken down into skill-based, decision, and perceptual errors. Regardless of the context in which this taxonomy is used, human factors principles will be critical to understanding the human capabilities or limitations that contributed to the error or errors that led to the accident. Violations are classified as routine or exceptional, with exceptional representing more egregious violations.

The "preconditions for unsafe acts" component of HFACS describes the environmental factors, conditions of operators, and personnel factors that can lead to increased propensity for unsafe acts. Wiegmann and Shappell (2003) provide a broad range of examples for each of these. The "conditions of operators" concept has drawn interest for many occupations over the years, especially in the area of testing operators, particularly in the transportation industry, for what has been termed fitness for duty.

HFACS includes an unsafe supervision component broken down into inadequate supervision, planned inappropriate operations, failure to correct problems, and supervisory violations. Thus, management accountability and culpability is a critical component of HFACS. These factors become particularly important for nonroutine work such as construction, where planning a job and the safety measures taken by supervisors and employees are critical and the lack of such planning has led to accidents.

The last component of HFACS is that of organizational influences, which includes resource management, organizational climate, and organizational process. This is perhaps the most difficult component to describe during an investigation. To truly understand factors such as organizational customs, communication in an organization, and procedural influences, a great deal of data gathering through interviews may be required. The CAIB (2003) report is an excellent example of a comprehensive set of findings regarding organizational influences. The interested reader is strongly encouraged to consult the case studies provided by Wiegmann and Shappell (2003) for in-depth information on application of this taxonomic system for investigating accidents.

Patterson and Shappell (2010) recently used a modified version of HFACS to analyze incident and accident cases from mines in Queensland, Australia. The HFACS nomenclature was modified to better fit language used in the mining industry. The analysis of 508 "high-potential incidents" and lost-time cases revealed that skill-based errors were the most common unsafe act, with no significant differences across mine types. Decision errors, most of which were identified to be procedural errors, did significantly vary across mine types with surface quarries having the highest percentage (48%) and underground coal mines having the lowest percentage (23.1%). Findings for unsafe acts were consistent across the 2004–2008 time period studied. The authors suggested that the results could be used to develop data-driven interventions to address the most significant sources of human factors deficiencies found.

## 6   CONCLUSIONS

Accident investigation can take on many forms, ranging from an analysis of administrative data to investigations of single incidents that last for months. Regardless of the nature or severity of the accident, gaining an understanding of the human capabilities and limitations contributing to the accident is often the key to understanding causation. Even in cases where "hardware faults" were seemingly the cause, there are causative contributions influenced by humans.

One of the best ways to explore different approaches to accident investigation and to gain an understanding of the scope and depth of different investigations is through published case studies. An interesting set of case studies centered around human factors is presented by Casey (1993), with human factors violations highlighted for a range of accident types, from the Bhopal disaster to a medical context. A case study describing how three cosmonauts from the former Soviet Union died in 1971 due to a pressure equalization valve that could not be turned quickly enough following depressurization is presented. Although the valve was intended to be used under such situations, operation had not been tested under the extreme conditions where physical capabilities were greatly reduced. Although the particular physical capabilities required were not difficult to understand or even to define empirically, the underlying issue was why no one had the foresight to test the system under operational conditions (or more realistically, simulated conditions). More in-depth cases studies are also provided, including an analysis of the Bhopal Union Carbide disaster in India in 1984 which killed more than 2500 residents. Perhaps most distressing about many of these case studies is that the actions that would have prevented many of the disasters were not excessively burdensome or technically infeasible.

Kletz (2001) provides an illustrative set of in-depth case studies, many of which are rather famous due to their severity, including the Three-Mile Island and Chernobyl nuclear disasters and the King's Cross railway station fire. The case studies include detailed overviews, in some cases engineering drawings of the relevant physical systems or components. Recommendations for prevention or mitigation at each stage of the accidents are provided, as are comprehensive references for additional sources of information. One advantage of studying such well-known disasters is that there is often detailed information that has been gathered from witnesses, physical evidence, and so on. Although most accident investigations will not reach this level of scope and depth, these case studies provide one of the best means of becoming familiar with accident investigation techniques and the wide array of people, equipment, procedures, and

organizational characteristics that often need to be considered during an accident investigation.

As was mentioned earlier, aviation safety has been increasing, in part due to changes made as a result of accidents. A recent survey of accident investigators from several different industries in Sweden (Rollenhagen et al., 2010) revealed that the phase of developing remedial actions following investigations was comparatively brief. When asked about the existence of particular methods or strategies for developing recommendations, "about 50%" of the respondents answered no. This phase should not be overlooked, particularly since preventing further incidents is the most positive outcome of a previous incident. Many organizations have begun to use commercial products for accident and incident investigation. Examples in the United States include TapRoot® (http://www.taproot.com/) and the "5 Whys" strategy (http://www.mindtools.com/) that has roots in the Toyota production system philosophy.

**Disclaimer:** The findings and conclusions in this chapter are those of the author and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

## REFERENCES

Beyer, D. S. (1928), *Industrial Accident Prevention*, 3rd ed., Houghton Mifflin, Boston.

Casey, S. M. (1993), *Set Phasers on Stun and Other True Tales of Design, Technology, and Human Error*, Aegean Publishing, Santa Barbara, CA.

Columbia Accident Investigation Board (CAIB) (2003), *Columbia Accident Investigation Board Report*, Vol. I, National Aeronautics and Space Administration and U.S. Government Printing Office, Washington, DC.

Dempsey, P. G., and Hashemi, L. (1999), "Analysis of Workers' Compensation Claims Associated with Manual Materials Handling," *Ergonomics*, Vol. 42, No. 1, pp. 183–195.

Drury, C. G., and Brill, M. (1983), "Human Factors in Consumer Product Accident Investigation," *Human Factors*, Vol. 25, pp. 329–342.

Hammer, W. (1985), *Occupational Safety Management and Engineering*, 3rd ed., Prentice-Hall, Englewood Cliffs, NJ.

Heinrich, H. W. (1959), *Industrial Accident Prevention: A Scientific Approach*, 4th ed., McGraw-Hill, New York.

Johnson, W. G. (1973), *MORT: The Management Oversight and Risk Tree*, SAN 821-2, U.S. Atomic Energy Commission and U.S. Government Printing Office, Washington, DC.

Kletz, T. A. (2001), *Learning from Accidents*, 3rd ed., Gulf Professional Publishing, Oxford.

Kucera, K. L., Loomis, D., and Marshall, S. W. (2008), "A Case Crossover Study of Triggers for Hand Injuries in Commercial Fishing," *Occupational and Environmental Medicine*, Vol. 65, pp. 336–341.

Last, J. M. (1995), *A Dictionary of Epidemiology*, 3rd ed., Oxford University Press, New York.

Lombardi, D. A., Sorock, G. S., Hauser, R. B., Nasca, P. C., Eisen, E. A., Herrick, R. F., and Mittleman, M. A. (2003), "Temporal Factors and the Prevalence of Transient Exposures at the Time of an Occupational Traumatic Hand Injury," *Journal of Occupational and Environmental Medicine*, Vol. 45, pp. 832–840.

Mittleman, M. A., Maclure, M., Tofler, G. H., Sherwood, J. B., Goldberg, R. J., and Muller, J. E. (1993), "Triggering of Acute Myocardial Infarction by Heavy Physical Exertion," *New England Journal of Medicine*, Vol. 329, pp. 1677–1683.

Moore, S. M., Porter, W. L., and Dempsey, P. G. (2009), "Fall from Equipment Injuries in U.S. Mining: Identification of Specific Research Areas for Future Investigation," *Journal of Safety Research*," Vol. 40, No. 6, pp. 455–460.

Patterson, J. M., and Shappell, S. A. (2010), "Operator Error and System Deficiencies: Analysis of 508 Mining Incidents and Accidents from Queensland, Australia Using HFACS," *Accident & Analysis Prevention*, Vol. 42, pp. 1379–1385.

Reason, J. (1997), *Managing the Risks of Organizational Accidents*, Ashgate Publishing, Aldershot, Hampshire, England.

Redelmeier, D. A., and Tibshirani, R. J. (1997), "Association between Cellular-Telephone Calls and Motor Vehicle Collisions," *New England Journal of Medicine*, Vol. 336, No. 7, pp. 453–458.

Rollenhagen, C., Westerlund, J., Lundberg, J., and Hollnagel, E. (2010), "The Context and Habits of Investigation Practices: A Study of 108 Swedish Investigators," *Safety Science,* Vol. 48, pp. 859–867.

Shappell, S., and Wiegmann, D. (2001), "Applying Reason: The Human Factors Analysis and Classification System," *Human Factors and Aerospace Safety*, Vol. 1, pp. 59–86.

Sorock, G. S., Ranney, T. A., and Lehto, M. R. (1996), "Motor Vehicle Crashes in Roadway Construction Workzones: An Analysis Using Narrative Text from Insurance Claims," *Accident Analysis and Prevention*, Vol. 28, No. 1, pp. 131–138.

Sorock, G. S., Lombardi, D. A., Gabel, C. L., Smith, G. S., and Mittleman, M. A. (2001a), "Case-Crossover Studies of Occupational Trauma: Methodological Caveats," *Injury Prevention*, Vol. 7, Suppl. I, pp. i38–i42.

Sorock, G. S., Lombardi, D. A., Hauser, R. B., Eisen, E. A., Herrick, R. F., and Mittleman, M. A. (2001b), "A Case-Crossover Study of Occupational Traumatic Hand Injury: Methods and Initial Findings," *American Journal of Industrial Medicine*, Vol. 39, pp. 171–179.

Sorock, G. S., Lombardi, D. A., Peng, D. K., Hauser, R. B., Eisen, E. A., Herrick, R. F., and Mittleman, M. A. (2004), "Glove Use and the Relative Risk of Acute Hand Injury: A Case-Crossover Study," *Journal of Occupational and Environmental Hygiene*, Vol. 1, pp. 182–190.

Vincoli, J. W. (1994), *Basic Guide to Accident Investigation and Loss Control*, Wiley, New York.

Whiting, S. E. (1939), "Men Who Go Down to Their Death in Tanks," demonstration at the Silver Jubilee of the National Safety Council, October 10–14, 1938, Liberty Mutual Insurance Company, Boston.

Wiegmann, D. A., and Shappell, S. A. (2003), *A Human Error Approach to Aviation Accident Analysis: The Human Factors Analysis and Classification System*, Ashgate Publishing, Aldershot, Hampshire, England.

# CHAPTER 39

# HUMAN FACTORS AND ERGONOMICS AUDITS

**Colin G. Drury**
University at Buffalo: State University of New York
Buffalo, New York


**Patrick G. Dempsey**
National Institute for Occupational Safety & Health
Pittsburgh, PA

## 1  PURPOSE

This chapter has two interrelated aims. First, we examine the human–machine tasks of inspecting, checking, and auditing to provide understandings that can guide work design, equipment design, and job aid development. Second, we apply this knowledge to inspecting, checking, and auditing of human factors/ergonomics aspects of human–machine systems. As part of this aim, we provide a detailed review and worked example of recent human factors/ergonomics audit programs. Throughout the chapter examples are given from a wide range of domains, from product usability audits, through aviation preflight checklists, to inspection of products for quality assurance.

## 2  INSPECTING, CHECKING, AND AUDITING

The idea of inspecting is as old as civilization. In the Sumerian epic *Gilgamesh*, almost 5000 years old, the narrator invites the reader to examine the quality of the walls of Uruk, built by Gilgamesh, the king (*Gilgamesh*, Tablet 1):

> Look at its wall which gleams like copper (?),
>
> inspect its inner wall, the likes of which no one can equal!
>
> Take hold of the threshold stone—it dates from ancient times!
>
> Go close to the Eanna Temple, the residence of Ishtar,
>
> such as no later king or man ever equaled!

> Go up on the wall of Uruk and walk around,
>
> examine its foundation, inspect its brickwork thoroughly.
>
> Is not (even the core of) the brick structure made of kiln-fired brick,
>
> and did not the Seven Sages themselves lay out its plans?"

The essence of the examining function is all there already. Bodily senses (look, take hold of) are used to compare the existing item (wall) with some implied or actual standard (e.g., kiln-dried brick). Inspection can have more formal definitions (e.g., in dictionaries and quality control texts), but a simple definition from Drury (2002, p. 27) provides a reasonable modern start: "The [test and] inspection system determines the suitability of a product or process to fulfill its intended function, within given parameters of accuracy, cost and timeliness."

Inspection is thus a decision function, for management, for the general public, and for an individual. Should the process be stopped before it produces more defects? Does this café meet local standards of cleanliness, so that people can eat there safely? Is this aircraft safe for me to fly in? Does this box of strawberries contain any bad fruit? The most basic decision is a go/no-go decision: Does this item or process fulfill its intended function or not? In practice, the amount of inspection, checking, and auditing needed to reach such a simple conclusion may be considerable. For example, a national safety board may need considerable evidence after a major accident to determine that the system (aircraft,

train, ferry, spacecraft) is fit to resume normal operations. As noted above, most inspection decisions are simpler than this.

In principle, inspection can be done by the producer or consumer of the goods or services directly, but this is not always satisfactory. Thus, the person machining a part can determine whether or not it meets standards, or the ultimate end user can inspect the part. Much of the quality revolution from the 1970s onward has been concerned with pushing such decisions back along the production system to ensure that decisions are made at the source (e.g., Evans and Lindsay, 1993; Drury, 1997) to prevent errors from propagating through a production system. Indeed, the preferred solution is to inspect the *process* rather than the *product* to ensure that defects are extremely unlikely to be produced. That is the aim of in-process statistical process control (e.g., Devor et al., 1992). Unlike our example of the box of strawberries, most ultimate consumers are not equipped to be able to inspect or check the complex devices and processes they use in daily living. In an agrarian economy, a farmer could examine a spade produced for him by the local blacksmith (e.g., Jones, 1981) with enough skill and visual/haptic information to reach a reasonable conclusion about whether the design and construction quality met his requirements. In our more specialized economy, consumers cannot make valid quality judgments on automobiles, computers, or even the safety of a local chemical plant because they lack both the specialized knowledge required and access to the inner workings of the product or system.

In such cases, customers rely on the judgment of professional inspectors, checkers, and auditors to make informed decisions. This reliance raises questions of honesty, trust, competence, and human–machine system design. From a sociotechnical systems perspective, inspectors must be seen as independent of producers of goods and services or their findings will not be accepted. For example, in civil aviation the U.S. Federal Aviation Administration (FAA) decrees that airline inspectors charged with checking airworthiness are kept organizationally independent of aviation maintenance technicians, who perform repairs, adjustments, and replacements. This independence can lead to role ambiguity and conflict in their job. For example, McKenzie (1958) noted that inspection is always of people: The inspector is judging the work of others by examining the outputs. Early work by McKenzie (1958), Thomas and Seaborne (1961), and Jamieson (1966) established that social pressures are an important part of the inspector's job and inspectors can at times change their behavior and performance in response to such pressures. The author worked in one factory where the customer returned shipments of product when they could not be used right away (e.g., during a strike) by finding a defect in the shipment. Later, when the customer needed the product, the factory repackaged the same shipments, often receiving complements on their improved quality!

This chapter covers both inspection and auditing, so that we need to establish that auditing and inspection do have commonalities or are indeed both instances of the same systems function and human behavior. Both clearly involve making decisions about the fitness of a product/service (inspection) or a system (auditing) for its intended use, with heavy overtones of protecting the public. However, a more detailed proof of congruence must be postponed until we have considered each process in more analytic detail (Section 4).

## 3  INSPECTION AND HUMAN FACTORS ERGONOMICS

As noted above, the study of inspection and inspectors is at least 50 years old, although most concentrated on inspection of products in a manufacturing setting. There are many other types, as noted by Drury (2002, 2009):

- *Regulatory inspection*—to ensure that regulated industries meet or exceed regulatory norms. Examples are review of restaurants against local service codes, fire safety inspection of buildings, and safety inspection of workplaces.

- *Medical inspection*—to ensure that a patient receives correct diagnosis of medical conditions. An example is inspection of mammograms (Nodine et al., 2002; Chen et al., 2008).

- *Maintenance*—to detect failure arising during the service life of a product. This failure detection function can be seen in inspection of road and rail bridges for structural determination or of civil airliners for stress cracks or corrosion.

- *Security*—to detect items deliberately concealed. These may be firearms or bombs carried onto aircraft, drugs smuggled across borders, or camouflaged targets in aerial photographs. Examples are X-ray inspection of airline baggage (e.g., Gale et al., 2000: Hsiao et al., 2008; Ghylin et al., 2007). They can also be suspicious happenings on a real-time video monitor at a security station. Law enforcement has many examples of searching crime sites for evidence.

- *Design review*—to detect discrepancies or problems with new designs. Examples are the checking of building drawings for building code violations, of chemical plant blueprints for possible safety problems, or of new restaurant designs for health code violations.

- *Functionality testing*—to detect lack of functionality in a completed system. This functional inspection can often include problem diagnosis, as with checks of avionics equipment in aircraft. Often functional inspection is particularly dangerous and costly, as in test flying aircraft or checking out procedures for a chemical process.

All of these inspection applications have much in common. Indeed, Drury (2003) has proposed a unified model of inspection in the security domain and shows how it can apply to all security inspection systems. This model was hardly new; it was based

**Table 1  Generic Function, Outcome, and Error Analysis of Test and Inspection**

| Function | Outcome | Errors | |
|---|---|---|---|
| **Setup** | Inspection system functional, calibrated correctly, and capable | 1.1. | Incorrect equipment |
| | | 1.2. | Nonworking equipment |
| | | 1.3. | Incorrect calibration |
| | | 1.4. | Incorrect or inadequate system knowledge |
| **Present** | Item (or process) presented to inspection system | 2.1. | Wrong item presented |
| | | 2.2. | Item mispresented |
| | | 2.3. | Item damaged by presentation |
| **Search** | Indications of all possible nonconformities detected, located | 3.1. | Indication missed |
| | | 3.2. | False indication detected |
| | | 3.3. | Indication mislocated |
| | | 3.4. | Indication forgotten before decision |
| **Decision** | All indications located by search measured correctly and classified correct outcome decision reached | 4.1. | Indication measured incorrectly |
| | | 4.2. | Indication classified incorrectly |
| | | 4.3. | Wrong outcome decision |
| | | 4.4. | Indication not processed |
| **Respond** | Action specified by outcome taken correctly | 5.1. | Nonconforming action taken on conforming item |
| | | 5.2. | Conforming action taken on nonconforming item |

*Source*: Drury (2002).

on function analytical models developed earlier by Drury (1978), Sinclair (1984), and Wang and Drury (1989). A recent incarnation can be seen in Jiang et al. (2003). Table 1 gives a generic functional breakdown of inspection, showing the major functions of inspection with the correct outcomes and errors arising from each function.

The reader is referred directly to Drury (2002) for a detailed consideration of inspection, including automated inspection and test, and a design methodology for "design for inspectability." In the current chapter, only an overview is provided to help understand the context for both checklists and audits. Each function is considered in the order given in Table 1. A single example of safety inspection of a workplace will be used to illustrate each function:

1. *Setup.* In this function, the inspection system is prepared for use. Needed tools, equipment, and supplies are procured, procedures are available to aid the inspector, and the inspector has been trained to perform the task correctly. For a workplace safety inspection, there will be some equipment needing calibration (e.g., psychrometer, air-sampling systems), a written procedure in the form of a checklist or computer program (e.g., Wilkins et al., 1997), and safety inspectors who have undertaken training and often certification. Certification is one way in which the competence and independence of inspectors are maintained, leading presumably to a higher degree of public trust in the inspection process. The SETUP function places demands on the regulatory system to provide the needed antecedents of effective inspection (i.e., sufficient resources).

2. *Present.* Here the inspector (suitably equipped) and the entity to be inspected come together so

that inspection can take place. The narrator of *Gilgamesh* urges the reader to "...go up to the wall of Uruk and walk around." Often, this function is purely mechanical: The manufacturing inspector has products arrive and depart on a conveyor, the safety inspector accesses all areas of a factory that could contain indications of an unsafe condition, and the FAA inspector goes to the file cabinets where maintenance records are kept to check that maintenance was performed and signed off correctly. The managerial implications for PRESENT are that inspectors must know the places they need to examine, and the organization being inspected must provide open and timely access to such sites. In safety inspection by a regulatory agency, there is often special legal provision for sites to be made available with little or no prior notice to prevent concealment of safety concerns known to management.

3. *Search.* Here we arrive at the first of the two most important and error-prone functions of inspection. Search is typically a sequential serial process during which the entire item to be inspected is brought under scrutiny piece by piece. The most obvious form of search is visual, and this has a long history of study, going back to the earliest days of human factors (Blackwell, 1946; Lamar, 1960). People (i.e., inspectors) search an area by successive visual fixations where the eye remains essentially stationary. What can be detected during a single fixation is a function of target and background characteristics (e.g., Overington, 1973) with considerable research (e.g., Treisman, 1986) on combinations of target and background conditions favoring rapid, parallel ("preattentive") search. Mathematical models (e.g., Morawski

et al., 1980; Wolfe, 1994) of the visual search process emphasize two features:

a. Over what area around the fixation point is the target detectable in a single fixation ("visual lobe") (e.g., Chan and Chiu, 2010)?

b. How are successive fixations sequenced to achieve coverage of the entire area?

Visual lobe models (e.g., Engel, 1971; Eriksen, 1990) determine how much area can be covered in a single fixation and hence the number of fixations required for coverage. The time to search an item completely varies directly with the number of fixations required and hence the reciprocal of lobe area (Morawski et al., 1980). Successive fixations are determined partly by top-down strategy and partly by bottom-up features of the currently fixated area (Wolfe, 1994). Top-down strategy has been modeled (e.g., Hong and Drury, 2002) as sequential, random, or a mixture (e.g., Arani et al., 1984). It is determined in part by the inspector's knowledge and expectations of where targets are likely. In an inspection context, a key derivation from search models is the *stopping time*, when the inspector decides that enough time has been spent on one item and moves to the next item. Stopping time chosen by inspectors accords quite well with the predictions of optimum stopping models (e.g., Chi and Drury, 1995; Baveja et al., 1996). Stopping time is a physical manifestation of how many resources the inspector (or the system giving inspection instructions) is willing to devote to each item inspected.

Visual search is not the only form of search important in inspection; there is also procedural search, where an inspector goes through a list of places that require examination. Procedural search is used extensively in aviation checklists (see later) where, for example, a preflight inspection of a general aviation aircraft follows a written procedure requiring examination of control surfaces, tires, fuel, structural joints, and so on. For each item on the checklist, the inspector is trained on what defects to look for. An example is fuel, where a small sample of fuel is drawn from a low point in the fuel line to check for water contamination, with water drops appearing as spheres in the fuel sample.

In a safety inspection, the inspector will have a list of key items/areas to search for indications of lack of safety. Inspectors will use their senses to examine each item in this procedural search, often requiring visual search within the procedural search. For example, inspectors must check that guarding is present on moving machinery, a largely procedural task. When they check safety records, such as the OSHA log in the United States, a visual search is required to determine whether all fields have been filled in and signed correctly. Note that

in both these examples considerable knowledge and skill are required to understand what would be an indication of a safety violation.

As noted in Table 1, the successful outcome of search is something detected that *could* be a defect. This is known in the nondestructive inspection (NDI) community as an *indication*. Subsequent inspection functions are concerned with how to deal with each indication. Note that if search fails, the indication is missed and subsequent functions cannot proceed. At least in visual inspection there is considerable evidence (Drury and Sinclair, 1983; Drury et al., 1997) that the search function is quite error prone, with only about 50% of defects ever being located as indications.

4. *Decision.* This is the function in which the indication is judged against a standard to determine whether it is a true defect. If inspection is about decision, this function represents the essence of inspection. Decision requires human (or machine-aided) judgment against a standard, so a standard must be prespecified. Standards in manufacturing inspection can come from physical properties (hardness, conductivity, surface finish) that can be measured with appropriate gauging. The decision in such cases of what the statistical quality control (SQC) community calls *variables inspection* can be automated quite simply and is thus rarely an appropriate human task. Not all measurements and standards can be implemented so simply. How do we quantify blemishes in the surface finish of automobile paint work (Lloyd et al., 2000) or corrosion areas on an aircraft fuselage (Wenner and Drury, 1996). These examples of *attributes inspection* require a complex, typically human judgment. Often, signal detection theory (SDT) has been used as a model of this part of the inspection process (e.g., Drury and Sinclair, 1983; Chi and Drury, 1995) although at times it has been misapplied to the overall inspection process, hence lumping search errors with wrong decisions to accept a true defect (e.g., Drury and Addison, 1973). SDT suggests a separation of decision difficulty (discriminability) from bias in reporting/not reporting defects (criteria), thus providing a useful link to different remedial actions, depending on whether the discriminability or criterion needs changing. For example, Drury et al. (1997) found that the decision function of inspection was extremely inconsistent between inspectors, implying the need for better training and job aids in aircraft inspection.

For safety inspection, discriminability represents the decision difficulty. This can be very easy when the implied standard is zero (e.g., *any* missing machine guard is a violation) or more difficult (e.g., how much untidiness represents "poor housekeeping"). The decision criterion reflects the willingness to report. From SDT

this is a function of the a priori probability of a defect and the relative costs of the two errors:

- *Miss*: not reporting a true defect
- *False alarm*: reporting an indication that was not a true defect

In safety inspection, the pressures on the inspector can be high. A false alarm can be used by the factory being inspected to bring disrepute on a regulatory agency and on the actual inspector. A miss can lead to an accident, injury, or even a Bhopal-like disaster. Similar pressures exist for aircraft inspectors, security operators, and even the intelligence community. The decision to report becomes even more difficult (from SDT) when the defect found is extremely rare. A recent example is the failure to find a crack in the titanium hub of a jet engine that caused the accident in 1997 at Pensacola, Florida. The inspector, despite years of experience, had never encountered a crack in a titanium hub before. More details on mathematical approaches to the decision function may be found in Drury (2002). An example of applying a mathematical model of the search plus decision functions to security inspection can be found in Ghylin et al. (2007).

5. *Response.* When the decision has been made, the action chosen must be taken. In a manufacturing context, this can be as simple as removing a defective item from the production process or as complex as stopping the process to diagnose the "root cause" of the defect being produced. In more general contexts, the action is often written (e.g., a repair order for a crack in an aircraft structure, written warning of unsanitary conditions in a restaurant, or a safety citation to company management for a missing machine guard).

Although response is a relatively mechanical function, it can be subject to errors. Aircraft inspectors who intend to write up all defects at the end of inspection can forget some defects, surgeons can mark the wrong leg for amputation, and the safety citation can be incomplete. This is one function where computer-based automation can help by making response simple and immediate. For example, Drury et al. (2000) developed a computer-based task card system for aircraft inspection that allowed easy generation of repair orders. Even more could have been done with drop-down menus for fault type (crack, corrosion, etc.) or a search function for the appropriate reference in the structural repair manual.

Throughout this short treatment of inspection, we have seen where and why good human factors/ergonomics practices can reduce error potential. There are also overarching considerations of job design and automation that can have a great impact on errors [e.g., the hybrid automation studies of Hou et al. (1993) and Jiang et al. (2003)]. Again, details may be found in Drury (2002).

The audit and checking activities associated with nonmanufacturing applications can now be placed in a suitable context.

## 4 CHECKING AND CHECKLISTS

When inspection is too complex to be carried out by an inspector unaided by procedural notes, a job aid is required to lead the inspector through the task. The nonmanufacturing examples given already (e.g., aircraft maintenance, safety inspectors) are typical of those requiring and using job aids. The simplest job aid for any procedural task (e.g., preparation for landing an aircraft) is the checklist. All pilots carry checklists for many complex procedures when a sequence of actions must be performed in a standard order. In addition to the landing preparation noted above, there are checklists for preflight inspection, startup/taxi, pre-takeoff, climb, cruise, postlanding, and engine shutdown. These are typically short laminated paper lists in general aviation or computer-based lists for corporate and passenger jets. Glider pilots use laminated checklists but also use mnemonics, such as "STALLS" for prelanding, where they might not have time to consult a written checklist. Safety inspectors in industry typically use a written checklist of several pages.

If the form of a checklist (paper, computer, mnemonic) can vary, so can the content and structure. Most checklists are used as memory aids for well-practiced tasks, so that they are structured as lists of commands, each of which is relatively tense: "switch both magnetos on: open fuel cock: prime engine for 5 seconds." The user is expected to know which are the magneto switches and which way they move for "on." Users are also expected to understand *why* each action is required so that they have some strategy for recovering from malfunctions (e.g., one magneto not working correctly).

In contrast, more detailed procedures are used for inspection and maintenance of aircraft and spacecraft. These procedures will spell out each step in detail, often with part numbers and numerical settings (e.g., tightening torque), and include warnings and cautions as well as a rationale for the overall procedure. With a computer-based system, detailed procedures can also be viewed as checklists: for example, when the procedure is being repeated after a short interval. Drury et al. (2000) used simple hypertext links to move between the checklist steps and the more detailed procedures in their program for inspection workcards. Much human factors design and evaluation has gone into the physical design of such procedures (e.g., Patel et al., 1994; Chervak et al., 1996; Drury et al., 2000), so in the remainder of this section we concentrate on classic checklists, as they are most often encountered in nonmanufacturing inspection and audit.

Checklists have their limitations, though. The cogent arguments put forward by Easterby (1967) provide a good early summary of these limitations in the context of design checklists, and most are still valid today. Checklists are only of use as an aid to designers of systems at the earliest stages of the process. By concentrating

on simple questions, often requiring yes or no answers, some checklists may reduce human factors to a simple stimulus–response system rather than encouraging conceptual thinking. Easterby quotes Miller (1967): "I still find that many people who should know better seem to expect magic from analytic and descriptive procedures. They expect that formats can be filled in by dunces and lead to inspired insights.... We should find opportunity to exorcise this nonsense" Easterby, 1967, p. 554).

Easterby finds that checklists can have a helpful structure but often have vague questions, make nonspecified assumptions, and lack quantitative detail. Checklists are seen as appropriate for some parts of ergonomics analysis (as opposed to synthesis) and are even more appropriate to aid operators (not ergonomists) in following procedural steps. Clearly, we should be careful, even 30 years on, to heed these warnings. Many checklists are developed, and many of these published, that contain design elements fully justifying such criticisms.

Most formal studies of checklist use have been in an aviation context, both in maintenance (Pearl and Drury, 1995) and in preflight inspection (Ockerman and Pritchett, 1998, 2000, 2004). They have also found widespread use in the flight operations side of aviation, with detailed analysis by Degani and Wiener (1990). The Degani and Wiener (1990) study laid the basis for much subsequent work on checklists. They analyzed incident reports from the National Transportation Safety Board (NTSB) and ASRS (NASA's Aviation Safety Reporting System), finding that the main checklist errors resulted from overlooking items following interruptions or distractions, particularly when working under time pressure or toward the end of the working day. Their recommended countermeasures were use of a "challenge and response" operating philosophy, grouping several items together and using a logical flow pattern. In particular, they advocated a "geographical" sequence of steps, good formatting/typography, and that "operators should keep checklists as short as possible to minimize interruptions." They also reviewed then-current technologies that could assist checklist use. An earlier study of checklists for circuit board inspection (Goldberg and Gibson, 1986) also found that a logically organized checklist outperformed a randomly organized one.

Patel et al. (1993) found that during the initial inspection of an aircraft on arrival at maintenance the sequence of tasks in the checklist or workcard did not match the sequence of tasks that aircraft maintenance technicians typically followed. In a related study by Pearl and Drury (1995), questionnaires and videotapes showed that mechanics tended to sequence their tasks using spatial cues on the airplane rather than the order specified on their workcard. The study also revealed that aviation maintenance technicians who performed low-level inspections used spatial locations of tasks to sequence them. In addition, many aircraft mechanics rarely used the checklist and viewed it as an only guide for inexperienced mechanics. Experienced inspectors felt that they had acquired sufficient skill to perform the inspection task using their memory and referred to the checklist only occasionally.

A more recent series of investigations (Ockerman and Pritchett, 1998, 2000, 2004) have examined the relationship between the medium (paper vs. wearable computers) on which the procedure was displayed, the presentations of procedure context, overreliance, and inspection performance for a preflight inspection task. The studies found that inspection performance could be influenced by the presence of procedure context information presented with procedures. The 1998 study also observed that one-third of the participants used their memory and not the task guidance system to perform the preflight inspection. They observed that in some sessions the subjects performed the task from memory and consulted the checklist only to see if anything was forgotten, echoing the Pearl and Drury (1995) findings from maintenance.

Computer applications for checklists were also advocated by Degani and Wiener (1990) and tested in an aviation maintenance/inspection environment by Drury et al. (2000). The latter study measured the impact of a hypertext-based computer program on the usability of work documentation in maintenance and inspection. Based on data collected in 1992–1993 at an airline partner, they concluded that computer-based inspection job aids were effective, although much of their effectiveness was attributed to good job aid design rather than computerization per se. Their task was one of detailed inspection of aircraft structures and used a checklist only as a top-level job aid, with more detailed instructions and data available via hyperlinks.

Major findings of all of these studies should be applicable to audit checklists, despite their somewhat different domains within aviation. Checklists are good job performance aids for repetitive tasks. They involve little explanation of detail or rationale for the sequence of operations, being mainly reminders of the correct sequence, often with facilities for marking (the "check" in "checklist") each item as it is performed to minimize the effect of interruptions or distractions (Degani and Wiener, 1990). The findings include (1) a geographical sequence is probably best; (2) good design principles should be followed; (3) technology can improve checklists; and (4) checklists are not always used, with reliance often placed on memory. The latter finding was reinforced by Wenner and Drury (2000), who note that some people did not read or follow very explicit instructions for performing a task.

Recently, two closely related studies of checklists were performed using a simulated repetitive aircraft inspection task with engineering student participants. The first (Larock, 2000; Larock and Drury, 2003) measured the effects of checklist layout and the number of sign-offs when a task was repeated eight times on eight days. The second (Pai, 2003) examined the use of computer-based checklists under the best design conditions found by Larock and Drury (2003) using the same task repeated over six days.

The first study compared functionally ordered and spatially ordered checklists and also whether each of the 108 items had to be signed off individually or whether they were signed off in 37 logically related subsets. As expected, spatial ordering was better than

functional ordering for both accuracy and speed. The number of sign-offs and the checklist layout interacted for sequence errors, where the best combination was spatial ordering and signing off in 37 groups. Over the course of eight daily trials, participants became faster at the task but tended to develop a spatial strategy for either checklist. The second study used this combination to test the efficacy of various computer implementations of the original paper checklist. A personal digital assistant (Palm-Pilot) was used with its built-in application of the "to-do list." A more user-friendly program was written specifically for the task studied and was implemented on the PDA and on a laptop computer. The conclusion was that the three computer implementations did not differ from each other, and all gave better speed and accuracy than those for the paper-based checklist.

These studies reinforce conclusions 1–4 noted above and showed that checklist behavior is not merely an artifact of using aviation professionals as subjects. Checklists emerge as a powerful tool, but one that needs careful human factors design to reach its maximum performance. Outside of aviation operations, there is little evidence that checklists are designed with these human factors findings in mind.

To develop checklists for auditing safety, ergonomics, or human factors, the design principles above should be followed. In addition, checklists need to be validated with actual users to ensure that their content, structure, and format do indeed lead to reliable performance.

## 5  AUDITING WITH SPECIFIC APPLICATION TO HUMAN FACTORS

When we audit an entity, we perform an examination of it. Dictionaries typically emphasize official examinations of (financial) accounts, reflecting the accounting origin of the term. Accounting texts go further: for example, "testing and checking the records of an enterprise to be certain that acceptable policies and practices have been consistently followed" (Carson and Carlson, 1977, p. 2). In the human factors field, the term is broadened to include nonfinancial entities but remains faithful to the concepts of checking, acceptable policies/practices, and consistency.

As with inspecting, auditing is mentioned in antiquity, at least in current translations: "[H]e who does not pay the fine annually shall owe ten times the sum, which the treasurer of the goddess shall exact; and if he fails in doing so, let him be answerable and give an account of the money at his audit" (Plato, *Laws*, Book VI).

Human factors audits can be applied, as can human factors itself, to both products and processes. Both applications have much in common, as any *process* can be considered as a *product* of a design procedure, but in this section we emphasize process audits because product evaluation is covered in detail in Chapter 50. Product usability audits have their own history (e.g., Malde, 1992), which is best accessed through the product design and evaluation literature (e.g., McClelland, 1990).

Auditing, like inspection, proceeds through a series of functional steps. For example, an audit by a certified public accountant would comprise the following steps (adapted from Koli, 1994):

1. *Diagnostic Investigation*. Describe the business and highlight areas requiring increased care and high risk.
2. *Test for Transaction*. Trace samples of transactions grouped by major area and evaluate.
3. *Test of Balances*. Analyze content.
4. *Formation of Opinion*. Communicate judgment in an audit report.

There are obvious direct parallels with the functions of inspection (Table 1), as noted by Drury (2009): *Diagnostic Investigation* comprises the Setup task, *Test for Transaction* comprises the Present and Search tasks, *Test of Balances* is the Decision task while *Formation of Opinion* is the Response task.

### 5.1  Need for Auditing Human Factors

Human factors or ergonomics programs have become a permanent feature of many companies, with typical examples shown in Alexander and Pulat (1985). As with any other function, human factors/ergonomics needs tools to measure its effectiveness. Earlier, when human factors operated through individual projects, evaluation could take place on a project-by-project basis. Thus, the interventions to improve apparel sewing workplaces described by Drury and Wick (1984) could be evaluated to show changes in productivity and reductions in cumulative trauma disorder causal factors. Similarly, Hasslequist (1981) showed productivity, quality, safety, and job satisfaction following human factors interventions in a computer component assembly line. In both cases, the objectives of the intervention were used to establish appropriate measures for the evaluation.

Ergonomics/human factors, however, is no longer confined to operating in a project mode. Increasingly, the establishment of a permanent function within an industry has meant that ergonomics is more closely related to the strategic objectives of the company. As Drury et al. (1989) have observed, this development requires measurement methodologies that also operate at the strategic level. For example, as a human factors group becomes more involved in strategic decisions about identifying and choosing the projects it performs, evaluation of the individual projects is less revealing. All projects performed could have a positive impact, but the group could still have achieved more with a more astute choice of projects. It could conceivably have had a more beneficial impact on the company's strategic objectives by stopping all projects for a period to concentrate on training the management, workforce, and engineering staff to make more use of ergonomics.

Such changes in the structure of the ergonomics/human factors profession indeed demand different evaluation methodologies. A powerful network of individuals, for example, who can, and do, call for human factors input in a timely manner can help an enterprise more than a number of individually successful project outcomes. Audit programs are one of the ways in which

such evaluations can be made, allowing a company to focus its human factors resources most effectively. They can also be used in a prospective, rather than retrospective, manner to help quantify the needs of the company for ergonomics/human factors. Finally, they can be used to determine which divisions, plants, departments, or even product lines are in most need of ergonomics input.

## 5.2 Design Requirements for Audit Systems

Returning to the definition of an audit, the emphasis is on checking, acceptable policies, and consistency. The aim is to provide a fair representation of the business for use by third parties. A typical audit by a certified public accountant follows the steps outlined in the previous section (diagnostic investigation, transaction test, balances test, opinion formation).

Such a procedure can also form a logical basis for human factors audits. The first step chooses the areas of study, the second samples the system, the third analyzes these samples, and the final step produces an audit report. These define the broad issues in human factors audit design:

1. *How to sample the system.* How many samples are to be used, and how are they distributed across the system?

2. *What to sample.* What specific factors are to be measured, from biomechanical to organizational?

3. *How to evaluate the sample.* What standards, good practices, or ergonomic principles are to be used for comparison?

4. *How to communicate the results.* What techniques are to be used for summarizing the findings, and how far can separate findings be combined?

A suitable audit system needs to address all of these issues, but some overriding design requirements must first be specified.

### 5.2.1 Breadth, Depth, and Application Time

Ideally, an audit system would be broad enough to cover any task in any industry, would provide highly detailed analysis and recommendations, and would be applied rapidly. Unfortunately, the three variables of breadth, depth, and application time are likely to trade off in a practical system. Thus, a thermal audit (Parsons, 1992) sacrifices breadth to provide considerable depth based on the heat balance equation but requires measurement of seven variables. Some can be obtained rapidly (air temperature, relative humidity), but some take longer (clothing insulation value, metabolic rate). Conversely, structured interviews with participants in an ergonomics program (Drury, 1990a) can be broad and rapid but quite deficient in depth.

At the level of audit instruments such as questionnaires or checklists, there are comprehensive surveys such as the Position Analysis Questionnaire (McCormick, 1979); the Arbeitswissenschaftliche Erhebungsverfahren zur Tätikgkeitsanalyse (AET) (Rohmert

and Landau, 1989), which takes 2–3 h to complete; or the simpler Work Analysis Checklist (Pulat, 1992). Alternatively, there are simple single-page checklists such as the Ergonomics-Working Position-Sitting Checklist (SHARE, 1990), which can be completed in a few minutes. Analysis and reporting can range in depth from merely tabulating the number of ergonomic standards violated to expert systems that provide prescriptive interventions (Ayoub and Mital, 1989).

Most methodologies fall between the various extremes given above, but the goal of an audit system with an optimum trade-off between breadth, depth, and time is probably not realizable. A better practical course would be to select several instruments and use them together to provide the specific breadth and depth required for a particular application.

### 5.2.2 Use of Standards

The human factors/ergonomics profession has many standards and good practice recommendations. These differ by country [American National Standards Institute (ANSI), British Standards Institution (BSI), German Institute for Standardization (DIN)], although commonality is increasing through joint standards such as those of the International Organization for Standardization (ISO). Some standards are quantitative, such as heights for school furniture (BSI, 1980), sizes of characters or a video terminal display (VDT) screen (ANSI/HFES-200), and occupational exposure to noise. Other standards are more general in nature, particularly those which involve management actions to prevent or alleviate problems, such as the Occupational Safety and Health Administration (OSHA, 1990) guidelines for meatpacking plants. Generally, standards are more likely to exist for simple tasks and environmental stressors and are hardly to be expected for the complex cognitive activities with which human factors predictions increasingly deal. Where standards exist, they can represent unequivocal elements of audit procedures as a workplace that does not meet these standards is in a position of legal violation. A human factors program that tolerates such legal exposure should clearly be held accountable in any audit. A comprehensive listing of standards pertaining to human factors and ergonomics can be found in the appropriate handbook (Karwowski, 2005).

Merely meeting legal requirements, however, is an insufficient test of the quality of ergonomics/human factors efforts. Many legal requirements are arbitrary or outdated: for example, weight limits for manual materials handing in some countries. Additionally, other aspects of a job with high ergonomic importance may not be covered by standards: for example, the presence of multiple stressors, work in restricted spaces resulting in awkward postures, or highly repetitive upper extremity motions. Finally, there are many "human factors good practices" that are not the subject of legal standards. Examples are the National Institute for Occupational Safety and Health (NIOSH) lifting equation (Waters et al., 1993), the Illuminating Engineering Society (IES, 1993) codes, or the zones of thermal comfort defined by the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE, 1989)

or Fanger (1970). In some cases, standards are available in a different jurisdiction from that being audited. As an example, the military standard MIL-1472D (U.S. DoD, 1989) provides detailed standards for control and display design that are equally appropriate to process controls in manufacturing industry but have no legal weight there.

Despite the lack of legislation covering many human factors concerns, standards and other instantiations of good practice do have a place in ergonomics audits. Where they exist, they can be incorporated into an audit system without becoming the only criterion. Thus, noise levels in the United States have a legal limit of 90 dBA for hearing protection purposes. But at levels far below this, noise can disrupt communications (Jones and Broadbent, 1987) and distract from task performance. An audit procedure can assess the noise on multiple criteria (i.e., on hearing protection and on communication interruptions), with the former criterion used on all jobs and the latter only where verbal communication is an issue.

If standards and other good practices are used in a human factors audit, they provide a quantitative basis for decision making. Measurement reliability can be high and validity self-evident for legal standards. However, it is good practice in auditing to record only the measurement used, and not its relationship to the standard, which can be established later. This removes any temptation by the analyst to "bend" the measurement to reach a predetermined conclusion or to become complacent when the measurement is somewhat below the standard yet still potentially a detriment to human performance. Illumination measurements, for example, can vary considerably over a workspace, so that the audit question.

Is the work surface illumination > 750 lux?

☐    yes    ☐    no

could be answered legitimately either way for some workspaces by choice of sampling point. Such temptation can be removed, for example, by the following audit question:

What is the illumination at four points on the workstation?

☐    ☐    ☐    ☐    lux

Later analysis can establish whether, for example, the mean exceeds 750 lux or whether any of the four points fall below this level.

It is also possible to provide later analyses that combine the effects of several simple checklist responses, as in Parsons's (1992) thermal audit, where no single measure would exceed good practice even though the overall result would be cumulative heat stress.

### 5.2.3 Evaluation of an Audit System

For a methodology to be of value, it must demonstrate validity, reliability, sensitivity, and usability. Most texts that cover measurement theory treat these aspects in detail (e.g., Kerlinger, 1964). Shorter treatments are found within human factors methodology texts (e.g., Drury, 1990b; Osburn, 1987).

*Validity* is the extent to which a methodology measures the phenomenon of interest. Does our ergonomics audit program indeed measure the quality of ergonomics in the plant? It is possible to measure validity in a number of ways, but ultimately all are open to argument. For example, if we do not know the "true" value of the "quality of ergonomics" in a plant, how can we validate our ergonomics audit program? Broadly, there are three ways in which validation can be tested.

*Content validity* is perhaps the simplest but least convincing measure. If each item of our measurement device displays the correct content, validity is established. Theoretically, if we could list all of the possible measures of a phenomenon, content validity would describe how well our measurement device samples these possible measures. In practice, it is assessed by having experts in the field judge each item for how well its content represents the phenomenon studied. Thus, the heat balance equation would be judged by most thermal physiologists to have a content that well represents the thermal load on an operator. Not all aspects are as easily validated!

A recent investigation of the content validity of five occupational health and safety management audits (Robson et al., 2010) against a published Canadian standard on occupational safety and health management found that 74% of the standard's content was partially (40%) or fully (34%) represented across all audits. In some cases, particular management elements audited were not covered well with considerable variability across program elements. Thus, even seemingly straightforward content validity of following a prescriptive standard may not be realized in practice if careful attention is not paid during the development and testing of the audit prior to implementation.

*Concurrent* (or *prediction*) *validity* has the most immediate practical impact. It measures empirically how well the output of the measurement device correlates with the phenomenon of interest. Of course, we must have an independent measure of the phenomenon of interest, which raises difficulties. To continue our example, if we used the heat balance equation to assess the thermal load on operators, there should be a high correlation between this and other measures of the effects of thermal load. Perhaps measures such as frequency of temperature complaints or of heat disorders: heat stroke, hyperthermia, hypothermia, and so on. In practice, however, measuring such correlations would be contaminated by, for example, the propensity to report temperature problems or individual acclimatization to heat. Overall outputs from a human factors audit (if such overall outputs have any useful meaning) should correlate with other measures of ergonomic inadequacy, such as injuries, turnover, quality measures, or productivity. Alternatively, we can ask how well the audit findings agree with independent assessments of qualified human factors engineers (Keyserling et al., 1992; Koli et al., 1993) and thus validate against one interpretation of current good practice.

Finally, there is *construct validity*. This is concerned with inferences made from scores, evaluated by considering all empirical evidence and models. Thus, a model

may predict that one of the variables being measured should have a particular relationship to another variable not in the measurement device. Confirming this relationship empirically would help validate the particular construct underlying our measured variable. Note that different parts of an overall measurement device can have their construct validity tested in different ways. Thus, in a broad human factors audit, the thermal load could differentiate between groups of operators who do and do not suffer from thermal complaints. In the same audit a measure of difficulty in a target aiming task could be validated against Fitts's law. Other ways to assess construct validity are those that analyze clusters or factors within a group of measures. Different workplaces audited on a variety of measures and the scores, which are then subjected to factor analysis, should show an interpretable, logical structure in the factors derived. This method has been used on large databases for job-evaluation-oriented systems such as McCormick's position analysis questionnaire (PAQ) (McCormick, 1979).

*Reliability* refers to how well a measurement device can repeat a measurement on the same sample unit. Classically, if a measurement $X$ is assumed to be composed of a true value $X_t$ and a random measurement error $X_e$, then

$$X = X_t + X_e$$

For uncorrelated $X_t$ and $X_e$, taking variances gives

$$\text{Variance}(X) = \text{Variance}(X_t) + \text{Variance}(X_e)$$

or

$$V(X) = V(X_t) + V(X_e)$$

We can define the reliability of the measurement as the fraction of measurement variance accounted for by true measurement variance:

$$\text{reliability} = \frac{V(X_t)}{V(X_t) + V(X_e)}$$

Typically, reliability is measured by correlating the scores obtained through repeated measurements. In an audit instrument, this is often done by having two (or more) auditors use the instrument on the same set of workplaces. The square of the correlation coefficient between the scores (either overall scores or separately for each logical construct) is then the reliability. Thus, PAQ was found to have an overall reliability of 0.79, tested using 62 jobs and two trained analysts (McCormick, 1979).

In the absence of direct measurement capabilities, many checklists rely on observation of workers and subsequent categorization (i.e., rating) of parameters of interest. Postures are an excellent example since most checklists utilize posture categories rather than relying on the rater to estimate angles. This introduces the issues of inter- and intrarater reliability. Inter- and intrarater reliability can be used during the development process to optimize reliability by revising categories to achieve higher reliability. Assessments of interrater reliability are more common, exemplified by the recent investigation of Park et al. (2009) examining interrater reliabilities for postures, materials handling, and noise and vibration exposure among hospital employees. Village et al. (2009) reported interrater reliabilities (kappa values) ranging from 0.21 to 1.0 for categorization of gross postures, trunk angles, and several materials-handling parameters.

*Sensitivity* defines how well a measurement device differentiates between entities. Does an audit system for human–computer interaction find a difference between software generally acknowledged to be "good" and "bad"? If not, perhaps the audit system lacks sensitivity, although of course there may truly be no difference between the systems except blind prejudice. Sensitivity can be affected adversely by poor reliability, which increases the variability in a measurement relative to a fixed difference between entities (i.e., gives a poor signal-to-noise ratio). Low sensitivity can also come from a floor or ceiling effect. These arise where almost all of the measurements cluster at a high or low limit. For example, if an audit question on the visual environment was

Does illumination exceed 10 lux?

☐     yes     ☐     no

almost all workplaces could answer "yes" (although the author has found a number that could not meet even this low criterion). Conversely, a floor effect would be a very high threshold for illuminance. Sensitivity can arise too when validity is in question. Thus, heart rate is a valid indicator of heat stress but not of cold stress. Hence, exposure to various degrees of cold stress would be measured only insensitively by heart rate.

*Usability* refers to the auditor's ease of use of the audit system. Good human factors principles should be followed: for example, document design guidelines in constructing checklists (Wright and Barnard, 1975; Patel et al., 1993). If the instrument does not have good usability, it will be used less often and may even show reduced reliability due to auditors' errors.

## 5.3  Audit System Design

As outlined in Section 2, the audit system must choose a sample, measure the sample, evaluate it, and communicate the results. In this section we approach these issues systematically.

An audit system is not just a checklist; it is a methodology that often includes the technique of a checklist. The distinction needs to be made between methodology and techniques. Almost three decades ago, Easterby (1967) used Bainbridge and Beishon's (1964) definitions:

- *Methodology*: a principle for defining the necessary procedures
- *Technique*: a means to execute a procedural step

Easterby notes that a technique may be applicable in more than one methodology.

### 5.3.1  Sampling Scheme

In any sampling, we must define the unit of sampling, the sampling frame, and the sample choice technique.

For a human factors audit the unit of sampling is not as self-evident as it appears. From a job evaluation viewpoint (e.g., McCormick, 1979), the natural unit is the job that is composed of a number of tasks. From a medical viewpoint the unit would be the individual. Human factors studies focus on the task/operator/ machine/environment (TOME) system (Drury, 1992a,b) or, equivalently, the software/hardware/environment/ liveware (SHEL) system [International Civil Aviation Organization (ICAO), 1989]. Thus, from a strictly human factors viewpoint, the specific combination of TOME can become the sampling unit for an audit program.

Unfortunately, this simple view does not cover all the situations for which an audit program may be needed. Although it works well for the rather repetitive tasks performed at a single workplace typical of much manufacturing and service industry, it cannot suffice when these conditions do not hold. One relaxation is to remove the stipulation of a particular incumbent, allowing for jobs that require frequent rotation of tasks. This means that the results for one task will depend on the incumbent chosen or that several tasks will need to be combined if an individual operator is of interest. A second relaxation is that the same operator may move to different workplaces, thus changing the environment as well as the task. This is typical of maintenance activities, where a mechanic may perform any one of a repertory of hundreds of tasks, rarely repeating the same task. Here, the rational sampling unit is the task, which is observed for a particular operator at a particular machine in a particular environment. Examples of audits of repetitive tasks (Mir, 1982; Drury, 1990a) and maintenance tasks (Chervak and Drury, 1995) are given later to illustrate these different approaches.

Definition of the sampling frame, once the sampling unit is settled, is more straightforward. Whether the frame covers a department, a plant, a division, or an entire company, enumeration of all sampling units is possible at least theoretically. All workplaces, or jobs, or individuals can in principle be listed, although in practice the list may never be up to date in an agile industry where change is the normal state of affairs. Individuals can be listed from personnel records, tasks from work orders or planning documents, and workplaces from plant layout plans. A greater challenge, perhaps, is to decide whether indeed the entire plant really is the focus of the audit. Do we include office jobs or just production? What about managers, foremen, part-time janitors, and so on? A good human factors program would see all of these tasks or people as worthy of study, but in practice they may have had different levels of ergonomic effort expended upon them. Should some tasks or groups be excluded from the audit merely because most participants agree that they have few pressing human factors problems? These are issues that need to be decided explicitly before the audit sampling begins.

Choice of the sample from the sampling frame is well covered in sociology texts. Within human factors it typically arises in the context of survey design (Sinclair, 1990). To make statistical inferences from the sample to the population (specifically to the sampling frame), our sampling procedure must allow the laws of probability to be applied. The sampling methods used most often are described here.

***Random Sampling***    Each unit within the sampling frame is equally likely to be chosen for the sample. This is the simplest and most robust method, but it may not be the most efficient. Where subgroups of interest (strata) exist and these subgroups are not equally represented in the sampling frame, one collects unnecessary information on the most populous subgroups and insufficient information on the least populous. This is because our ability to estimate a population statistic from a sample depends on the absolute sample size and not, in most practical cases, on the population size. As a corollary, if subgroups are of no interest, random sampling loses nothing in efficiency.

***Stratified Random Sampling***    Each unit within a particular stratum of the sampling frame is equally likely to be chosen for the sample. With stratified random sampling we can make valid inferences about each of the strata. By weighting the statistics to reflect the size of the strata within the sampling frame, we can also obtain population inferences. This is often the preferred auditing sampling method, as, for example, we would wish to distinguish between different classes of tasks in our audits: production, warehouse, office, management, maintenance, security, and so on. In this way our audit interpretation could give more useful information concerning where ergonomics is being used appropriately.

***Cluster Sampling***    Clusters of units within the sampling frame are selected, followed by random or nonrandom selection within clusters. Examples of clusters would be the selection of particular production lines within a plant (Drury, 1990a) or selection of "representative" plants within a company or division. The difference between cluster and stratified sampling is that in cluster sampling only a subset of possible units within the sampling frame is selected, whereas in stratified sampling all of the sampling frame is used, as each unit must belong to one stratum. Because clusters are not randomly selected, the overall sample results will not reflect population values, so that statistical inference is not possible. If units are chosen randomly within each cluster, statistical inference within each cluster is possible. For example, if three production lines are chosen as clusters and workplaces sampled randomly within each, the clusters can be regarded as fixed levels of a factor and the data subjected to analysis of variance to determine whether there are significant differences between levels of that factor. What is sacrificed in cluster sampling is the ability to make *population* statements. Continuing this example, we could state that the lighting in line *A* is better than in line *B* or *C* but still not be able to make statistically valid statements about the plant as a whole.

### 5.3.2 Data Collection Instrument

So far we have assumed that the instrument used to collect the data from the sample is based on measured

data where appropriate. Although this is true of many audit instruments, this is not the only way to collect audit data. There have been interviews with participants (Drury, 1990a), interviews and group meetings to locate potential errors (Fox, 1992), and use of archival data such as injury of quality records (Mir, 1982). All have potential uses with, as remarked earlier, a judicious range of methods often providing the appropriate composite audit system.

One consideration regarding audit technique design and use is the extent of computer involvement. Computers are now inexpensive, portable, and powerful, so that they can be used to assist data collection, data verification, data reduction, and data analysis (Drury, 1990a). With the advent of more intelligent interfaces, checklist questions can be answered from mouse clicks on buttons, or selection from menus, as well as the more usual keyboard entry. Data verification can take place at entry time by checking for out-of-limits data, or odd data, such as the ratio of luminance to illuminance implying a reflectivity greater than 100%. In addition, branching in checklists can be made easier, with only valid follow-on questions highlighted. The "checklist user's manual" can be built into the checklist software using context-sensitive help facilities, as in the ergonomics evaluation analysis methodology (EEAM) checklist (Chervak and Drury, 1995). Computers can, of course, be used for data reduction (e.g., finding the insulation value of clothing from a clothing inventory), data analysis, and results presentation.

Having made the case for computer use, some precautions are in order. Computers are still bulkier than simple pencil-and-paper checklists. Computer reliability is not perfect, so that inadvertent data loss is still a real possibility. Finally, software and hardware date much more rapidly than hard copy, so that results stored safely on the latest media may be unreadable 10 years later. How many of us can still read punched cards or 8-in. floppy disks? In contrast, hard-copy records are still available from before the start of the common era.

### Checklists and Surveys as Audit Tools

For many practitioners the proof of the effectiveness of an ergonomics effort lies in the ergonomic quality of the work systems it produces. A plant or office with appropriate human–machine function allocation, well-designed workplaces, comfortable environment, adequate placement/training, and inherently satisfying jobs almost by definition has been well served by human factors. Such a facility may not have human factors specialists, just good designers of environment, training, organization, and so on, working independently, but this would generally be a rare occurrence. Thus, a checklist to measure such inherently ergonomic qualities has great appeal as part of an audit system. We have covered the design aspects of checklists in general, so we concentrate here on their use in the context of human factors/ergonomics audits.

Such checklists are almost as old as the discipline. An early paper by Burger and deJong (1964) lists four earlier checklists for ergonomic job analysis before going on to develop their own. Theirs was commissioned by the International Ergonomics Association

(IEA) in 1961 and is usually known as the IEA checklist. It was based in part on one developed at the Philips Health Centre by G. J. Fortuin and provided in detail in Burger and deJong's paper.

Like any other questionnaire, a checklist needs to have both a helpful overall structure and well-constructed questions. It should also be proven reliable, valid, sensitive, and usable, although precious few meet all these criteria. A recent survey of Certified Professional Ergonomists in the United States (Dempsey et al. 2005) revealed that 70.5% of respondents used checklists, with 67 of 301 respondents that reported using checklists indicating the the checklist was developed on their own or by their employer. It is unlikely these were subjected to tests of reliability or validity, although the usability should be high for those developed by the actual users. The most commonly identified single checklist was referred to as the OSHA checklist, with only 27 responses indicating that there are many checklists used by practicing ergonomists with seemingly few highly popular checklists.

In the remainder of this section, a selection of checklists is presented as typical of (reasonably) good practice. Emphasis will be on objective, structure, and question design. Note that checklists are not the only approach possible. Westwater and Johnson (1995) compared them with expert evaluation and empirical user testing in evaluating PDA design. They concluded that user-based evaluations led to more insights for this evaluation.

1. *IEA Checklist*. The IEA checklist (Burger and deJong, 1964) was designed for ergonomic job analysis over a wide range of jobs. It uses the concept of functional load to give a logical framework relating physical load, perceptual load, and mental load to the worker, the environment, and working methods/tools/machines. Within each cell (or subcell, e.g., physical load could be static or dynamic) the load was assessed on different criteria, such as force, time, distance, occupational medical, and psychological criteria. Table 2 shows the structure and typical questions. Dirken (1969) modified the IEA checklist to improve the questions and methods of recording. He found that it could be applied in a median time of 60 min per workstation. No data are given on evaluation of the IEA checklist, but its structure has been so influential that it is included here for more than historical interest.

2. *Position Analysis Questionnaire*. The PAQ is a structured job analysis questionnaire using 187 worker-oriented elements to characterize the human behaviors involved in jobs (McCormick et al., 1969). The PAQ is structured into six divisions, with the first three representing the classic experimental psychology approach (information input, mental process, work output) and the next a broader sociotechnical view (relationships with other persons, job context, other job characteristics). Table 3 shows these major divisions, examples of job elements in each, and the rating scales employed for response.

**Table 2  IEA Checklist Structure and Typical Questions**
**Structure**

|  |  | A | B | C |
|---|---|---|---|---|
| Load: 1.  Mean<br>      2.  Peaks<br>          (intensity,<br>          frequency,<br>          duration) |  | Worker | Environment | Working<br>Method,<br>Tools,<br>Machines |
| I. Physical load | 1.  Dynamic<br>2.  Static |  |  |  |
| II. Perceptual load | 1.  Perception<br>2.  Selection, decision<br>3.  Control of movement |  |  |  |
| III. Mental load | 1.  Individual<br>2.  Group |  |  |  |

**Typical Question**

| I/B.  Physical<br>Load/Environment |  | 2.1.  Physiological Criteria |
|---|---|---|
|  | 1.  Climate: high and low temperatures |  |
|  | 1.  Are these extreme enough to affect comfort or efficiency? |  |
|  | 2.  If so, is there any remedy? |  |
|  | 3.  To what extent is working capacity adversely affected? |  |
|  | 4.  Do personnel have to be specially selected for work in this particular environment? |  |

**Table 3  PAQ Structure and Scales**
**Structure**

| Division | Definition | Examples of Questions |
|---|---|---|
| 1.  Information input | Where and how does the worker get the information he uses in performing his job? | 1.  Use of written materials<br>2.  Near-visual differentiation |
| 2.  Mental processes | What reasoning, decision making, planning, and information-processing activities are involved in performing the job? | 1.  Level of reasoning in problem solving<br>2.  Coding–decoding |
| 3.  Work output | What physical activities does the worker perform, and what tools or devices does he use? | 1.  Use of keyboard devices<br>2.  Assembling–unassembling |
| 4.  Relationships with other persons | What relationships with other people are required in performing the job? | 1.  Instructing<br>2.  Contacts with public or customers |
| 5.  Job context | In what physical or social contexts is the work performed? | 1.  High temperature<br>2.  Interpersonal; conflict situations |
| 6.  Other job characteristics | What activities, conditions, or characteristics other than those described above are relevant to the job? | 1.  Specified work pace<br>2.  Amount of job structure |

**Scales**

| Types of Scales | | Scales Values | |
|---|---|---|---|
| Code | Type of Rating | Rating | Definition |
| U | Extent of use | N | Does not apply |
| I | Importance of the job | 1 | Very minor |
| T | Amount of time | 2 | Low |
| P | Possibility of occurrence | 3 | Average |
| A | Applicability (yes/no only) | 4 | High |
| S | Special code | 5 | Extreme |

*Source*: McCormick (1979).

Construct validity was tested by factor analyses of databases containing 3700 and 2200 jobs, which established 45 factors. Thirty-two of these fit neatly into the original six-division framework, with the remaining 13 being classified as "overall dimensions." Further proof of construct validity was based on 76 human attributes derived from the PAQ, rated by industrial psychologists and the ratings subjected to principal-component analysis to develop dimensions "which had reasonably similar attribute profiles" (McCormick, 1979, p. 204). As noted earlier, interreliability was 0.79 based on another sample of 62 jobs.

The PAQ covers many of the elements of concern to human factors engineers and has indeed much influenced subsequent instruments, such as AET. With good reliability and useful (although perhaps dated) construct validity, it is still a useful instrument if the natural unit of sampling is the job. The exclusive reliance on rating scales applied by the analyst goes rather against current practice of comparison of measurements against standards or good practices.

3. *AET* (*Arbeitswissenschaftliche Erhebungsverfahren zur Tätikgkeitsanalyse*). The AET has been published in German (Landau and Rohmert, 1981) and later in English (Rohmert and Landau, 1983). It is the job analysis subsystem of a comprehensive system of work studies. It covers "the analysis of individual components of man-at-work systems as well as the description and scaling of their interdependencies" (Rohmert and Landau, 1983, pp. 9–10). As with all good techniques, it starts from a model of the system (Verband für Arbeitsgestaltung, Betriebsorganisation und Unternehmensentwicklung, REFA, 1971; referenced in Wagner, 1989), to which is added Rohmert's stress–strain concept. The latter sees strain as being caused by the intensity and duration of stresses impinging on an operator's individual characteristics. It is seen as useful in the analysis of requirements and work design, organization in industry, personnel management, and vocational counseling and research.

AET itself was developed over many years using PAQ as an initial starting point. Table 4 shows the structure of the survey instrument with typical questions and rating scales. Note the similarity between AET's job demands analysis and the first three categories of the PAQ and between the scales used in AET and PAQ (Table 3).

Measurements of validity and reliability of AET are discussed by H. Luczak in an appendix to Landau and Rohmert (1981), although no numerical values are given. Cluster analysis of 99 AET records produced groupings that supported the AET constructs. Seeber et al. (1989) used AET along with two other work analysis methods in 170 workplaces. They found that AET provided the most differentiating aspects (suggesting sensitivity). They also measured postural complaints and showed that only the AET groupings for 152 female workers found significant differences between complaint levels, thus helping establish construct validity.

Like PAQ before it, AET has been used on many thousands of jobs, mainly in Europe. A sizable database is maintained that can be used for both norming of new jobs analyzed and analysis to test research hypotheses. It remains a most useful instrument for work analysis.

4. *Ergonomics Audit Program* (Mir, 1982; Drury, 1990a). This program was developed at the request of a multinational corporation to be able to audit its various divisions and plants as ergonomics programs were being instituted. The system developed was a methodology of which the workplace survey was one technique. Overall, the methodology used archival data or outcome measures (injury reports, personnel records, productivity) and critical incidents to rank order departments within a plant. A cluster sampling of these departments gives either the ones with the highest need (if the aim is to focus ergonomic effort) or a sample representative of the plant (if the objective is an audit). The workplace survey is then performed on the sampled departments.

The workplace survey was designed based on ergonomic aspects derived from a task/operator/machine/environment model of the person at work. Each aspect formed a section of the audit, and sections could be omitted if they were clearly not relevant (e.g., manual materials-handling aspects for data entry clerks). Questions within each section were based on standards, guidelines, and models, such as the NIOSH (1981) lifting equation, ASHRAE's (1990) *Handbook of Fundamentals* for thermal aspects, and Givoni and Goldman's (1972) model for predicting heart rate. Table 5 shows the major sections and typical questions.

Data were entered into the computer program and a rule-based logic evaluated each section to provide messages to the user in the form of either a "section shows no ergonomic problems" message:

MESSAGE

Results from analysis of auditory aspects:
Everything OK in this section.

or discrepancies from a single input:

MESSAGE

Seats should be padded, covered with nonslip materials, and have the front edge rounded.

**Table 4  AET Structure and Scales**

**Structure**

| Part | Major Divisions | Sections |
|------|-----------------|----------|
| A. Work systems analysis | 1. Work objects | 1.1. Material work objects<br>1.2. Energy as work object<br>1.3. Information as work object<br>1.4. Humans, animals; plants as work objects |
| | 2. Equipment | 2.1. Working equipment<br>2.2. Other equipment |
| | 3. Work environment | 3.1. Physical environment<br>3.2. Organizational and social environment<br>3.3. Principles and methods of remuneration |
| B. Task analysis | 1. Tasks relating to material work objects<br>2. Tasks relating to abstract work objects<br>3. Human-related tasks<br>4. Number and repetitiveness of tasks | |
| C. Job demand analysis | 1. Demands on perception | 1.1. Mode of perception<br>1.2. Absolute/relative evaluation of perceived information<br>1.3. Accuracy of perception |
| | 2. Demands for decision | 2.1. Complexity of decisions<br>2.2. Pressure of time<br>2.3. Required knowledge |
| | 3. Demands for response/activity | 3.1. Body postures<br>3.2. Static work<br>3.3. Heavy muscular work<br>3.4. Light muscular work, active light work<br>3.5. Strenuousness and frequency of movements |

**Scales**

| Types of Scales | | Scales Values | |
|-----------------|--|---------------|--|
| Code | Type of Rating | Duration Value | Definition |
| A | Does this apply? | 0 | Very infrequent |
| F | Frequency | 1 | Less than 10% of shift time |
| S | Significance | 2 | Less than 30% of shift time |
| D | Duration | 3 | 30–60% of shift time |
| | | 4 | More than 60% of shift time |
| | | 5 | Almost continuously during whole shift |

or discrepancies based on the integration of several inputs:

MESSAGE
The total metabolic workload is 174 watts.
Intrinsic clothing insulation is 0.56 clo.
Initial rectal temperature is predicted to be 36.0°C.
Final rectal temperature is predicted to be 37.1°C.

Counts of discrepancies were used to evaluate departments by ergonomics aspect, while the messages were used to alert company personnel to potential design changes. The latter use of the output as a training device for nonergonomic personnel was seen as desirable in a multinational company rapidly expanding its ergonomics program.

Reliability and validity have not been assessed, although the checklist has been used in a number of industries (Drury, 1990a). The workplace survey has been included here because, despite its lack of measured reliability and validity, it shows the relationship between audit as methodology and checklist as technique.

5. *ERGO, EEAM, and ERNAP* (Koli et al., 1993; Chervak and Drury, 1995). These checklists are both part of complete audit systems for different aspects of civil aircraft hangar activities. They were developed for the FAA to provide tools for assessing human factors in aircraft inspection (ERGO) and maintenance (EEAM)

**Table 5  Workplace Survey Structure and Typical Questions**

| Section | Major Classification | Examples of Questions |
|---|---|---|
| 1. Visual aspects | | Nature of task? |
| | | Illuminance at task (midfield, outer field)? |
| 2. Auditory aspects | | Noise level (dBA)? |
| | | Main source of noise? |
| 3. Thermal aspects | | Strong radiant sources present? |
| | | Wet bulb temperature? |
| | | (Clothing inventory) |
| 4. Instruments, controls, displays | Standing vs. seated | Are controls mounted between 30 and 70 in.? |
| | Displays | Signals for crucial visual checks? |
| | Labeling | Are trade names deleted? |
| | Coding | Color codes same for control and display? |
| | Scales, dials, counters | All numbers upright on fixed scales? |
| | Control–display relationships | Grouping by sequence or subsystem? |
| | Controls | Emergency button diameter > 0.75 in.? |
| 5. Design of workplaces | Desks | Seat to underside of desk > 6.7 in.? |
| | Chairs | Height easily adjustable to 15–21 in.? |
| | Posture | Upper arms vertical? |
| 6. Manual materials handling | NIOSH (1981) lifting guide | Task, H, V, D, F |
| 7. Energy expenditure | | Cycle time? |
| | | Object weight? |
| | | Type of work? |
| 8. Assembly/ repetitive aspects | | Seated, standing, or both? |
| | | If heavy work, is bench 6–16 in. below elbow height? |
| 9. Inspection aspects | | Number of fault types? |
| | | Training time until unsupervised? |

activities, respectively. Inspection and maintenance activities are nonrepetitive in nature, controlled by task cards issued to technicians at the start of each shift. Thus, the sampling unit is the task card, not the workplace, which is highly variable between task cards. Their structure was based on extensive task analyses of inspection and maintenance tasks, which led to generic function descriptions of both types of work (Drury et al., 1990). Both systems have sampling schemes and checklists. Both are computer based with initial data collection on either hard copy or direct into a portable computer. Recently, both have been combined into a single program (ERNAP) distributed by the FAA's Office of Aviation Medicine. The structure of ERNAP and typical questions are given in Table 6.

As in Mir's ergonomics audit program, the ERNAP, the checklist is again modular, and the software allows formation of data files, selection of required modules, analysis after data entry is completed, and printing of audit reports. Similarly, the ERGO, EEAM, and ERNAP instruments use quantitative or yes/no questions comparing the value entered with standards and good-practice guides. Each takes about 30 min per task. Output is in the form of an audit report for each workplace, similar

to the messages given by Mir's workplace survey, but in narrative form. Output in this form was chosen for compatibility with existing performance and compliance audits used by the aviation maintenance community.

Reliability of a first version of ERGO was measured by comparing the output of two auditors on three tasks. Significant differences were found at $p < 0.05$ on all three tasks, showing a lack of interrater reliability. Analysis of these differences showed them to be due largely to errors on questions requiring auditor judgment. When such questions were replaced with more quantitative questions, the two auditors had no significant disagreements on a later test. Validity was measured using concurrent validation against six Ph.D. human factors engineers who were asked to list all ergonomic issues on a power plant inspection task. The checklist found more ergonomic issues than the human factors engineers. Only a small number of issues were raised by the engineers that were missed by the checklist. For the EEAM checklist, again an initial version was tested for reliability with two auditors and achieved the same outcome for only 85% of the questions. A modified version was tested and the reliability was considered satisfactory with 93% agreement. Validity was again tested against four human factors engineers;

**Table 6  ERNAP Structure and Typical Questions**

| Audit Phase | Major Classification | Examples of Questions |
|---|---|---|
| I.   Premaintenance | Documentation | Is feedforward information on faults given? |
| | Communication | Is shift change documented? |
| | Visual characteristics | If fluorescent bulbs are used, does flicker exist? |
| | Electric/pneumatic equipment | Do pushbuttons prevent slipping of fingers? |
| | Access equipment | Do ladders lave nonskid surfaces on landings? |
| II.  Maintenance | Documentation | Does inspector sign off workcard after each task? |
| | Communication | Explicit verbal instructions from supervisor? |
| | Task lighting | Light levels in four zones during task (fc)? |
| | Thermal issues | Wet-bulb temperature in hanger bay (°C)? |
| | Operator perception | Satisfied with summer thermal environment? |
| | Auditory issues | Noise levels at five times during task (dBA)? |
| | Electrical and pneumatic | Are controls easily differentiated by touch? |
| | Access equipment | Is correct access equipment available? |
| | Hand tools | Does the tool handle end in the palm? |
| | Force measurements | What force is being applied (kg)? |
| | Manual material handling | Does task require pushing or pulling forces? |
| | Vibration | What is total duration of exposure on this shift? |
| | Repetitive motion | Does the task require flexion of the wrist? |
| | Access | How often was access equipment repositioned? |
| | Posture | How often were following postures adopted? |
| | Safety | Is inspection area cleaned adequately for inspection? |
| | Hazardous material | Were hazardous materials signed out and in? |
| III. Postmaintenance | Buyback | Are discrepancy worksheets readable? |

this time the checklist found significantly more ergonomic issues than the engineers without missing any of the issues they raised.

The ERNAP audits have been included here to provide examples of a checklist embedded in an audit system where the workplace is *not* the sampling unit. They show that nonrepetitive tasks can be audited in a valid and reliable manner. In addition, they demonstrate how domain-specific audits can be designed to take advantage of human factors analyses already made in the domain.

6. *Upper Extremity Checklist* (Keyserling et al., 1993). As its name suggests, this checklist is narrowly focused on biomechanical stresses to the upper extremities that could lead to cumulative trauma disorders (CTDs). It does not claim to be a full-spectrum analysis tool but is included here as a good example of a special-purpose checklist that has been carefully constructed and validated. The checklist (Table 7) was designed for use by management and labor to fulfill a requirement in the OSHA guidelines for meatpacking plants. The aim is to screen jobs rapidly for harmful exposures rather than to provide a diagnostic tool. Questions were designed based on the biomechanical literature, structured into six sections. Scoring was based on simple presence or absence of

a condition or on a three-level duration score. As shown in Table 7, the two or three levels were scored as 0, √, or *, depending on the stress rating built into the questionnaire. These symbols represented insignificant, moderate, or substantial exposures. A total score could be obtained by summing moderate and substantial exposures.

The upper extremity checklist was designed to be biased toward false positives (i.e., to be very sensitive). It was validated against detailed analyses of 51 jobs by an ergonomics expert. Each section (except the first, which recorded only the dominant hand) was considered as giving a positive screening if at least one * rating was recorded. Across the various sections, there was reasonable agreement between checklist users and the expert analysis, with the checklist being generally more sensitive, as was its aim. The original reference shows the findings of the checklist applied to 335 manufacturing and warehouse jobs.

As a special-purpose technique in an area of high current visibility for human factors, the upper extremity checklist has proven validity, can be used by those with minimal ergonomics training for screening jobs, and takes only a few minutes per workstation. The same team has also developed and validated a legs, trunk, and neck job screening procedure along similar lines (Keyserling et al., 1992).

**Table 7  Upper Extremity Checklist: Structure and Scoring**

**Structure**

| Major Section | Examples of Questions |
|---|---|
| Worker information | Which hand is dominant? |
| | Repetitive use of the hands and wrists? |
| | If ''yes,'' then: |
| | Is cycle < 30 s? |
| | Repeated for >50% cycle? |
| Mechanical stress | Do hard or sharp objects put localized pressure on: |
| | Back or side of fingers? |
| | Palm or base of hand? |
| Force | Lift, carry, push, or pull objects >4.5 kg? |
| | If gloves worn, do they hinder gripping? |
| Posture | Is pinch grip used? |
| | Is there wrist deviation? |
| Tools, hand-held objects and equipment | Is vibration transmitted to the operator's hand? |
| | Does cold exhaust air blow on the hand or wristz? |

**Scoring Scheme**

| Question | Scoring | | |
|---|---|---|---|
| Is there wrist deviation? | No | Some | > 33% cycle |
| | 0 | √ | * |
| Overall evaluation: | | | |

total score = number of √ + number of *

7. *Ergonomic Checkpoints.* The workplace improvement in small enterprises (WISE) methodology (Kogi, 1994) was developed by the IEA and the International Labour Office (ILO) to provide cost-effective solutions for smaller organizations. It consists of a training program and a checklist of potential low-cost improvements. This checklist, called *ergonomics checkpoints*, can be used both as an aid to discovery of solutions and as an audit tool for workplaces within an enterprise.

The 128-point checklist has now been published (Kogi and Kuorinka, 1995) and as a book by the ILO (1999). It covers the nine areas shown in Table 8. Each item is a statement rather than a question and is called a *checkpoint*. For each checkpoint there are four sections, also shown in Table 8. There is no scoring system as such; rather, each checkpoint becomes a point of evaluation of each workplace for which it is appropriate. Note that each checkpoint also covers why that improvement is important and a description of the core issues underlying it. Both of these help the move from rule-based reasoning to knowledge-based reasoning as nonergonomists continue to use the checklist. A similar idea was embodied in the Mir (1982) ergonomic checklist.

8. *Operational Demand Evaluation Checklist.* Public et al. (2010) developed this checklist to capture the operational demands on rail signalers. Rail signalers are responsible for directing rail traffic to attain timely performance as well as safety of the traffic as well as equipment and people that could be impacted by a mishap. Rail systems are complex dynamic

**Table 8  Ergonomic Checkpoints**

Structure of the Checklist

| Major Section | Typical Checkpoints |
|---|---|
| Materials handling | Clear and mark transport ways. |
| Handtools | Provide handholds, grips, or good holding points for all packages and containers. |
| Productive machine safety | Use jigs and fixtures to make machine operations stable, safe, and efficient. |
| Improving workstation design | Adjust working height for each worker at elbow level or slightly below it. |
| Lighting | Provide local lights for precision or inspection work. |
| Premises | Ensure safe wiring connections for equipment and lights. |
| Control of hazards | Use feeding and ejection devices to keep hands away from dangerous parts of machinery. |
| Welfare facilities | Provide and maintain good changing, washing, and sanitary facilities to keep good hygiene and tidiness. |
| Work organization | Inform workers frequently about the results of their work. |

**Structure of Each Checkpoint**

| | |
|---|---|
| Why? | Reasons why improvements are important |
| How? | Description of several actions each of which can contribute to improvement |
| Some more hints | Additional points which are useful for attaining the improvement |
| Points to remember | Brief description of the core element of the checkpoint |

*Source*: K. Kogi, private communication, November 13, 1995.

entities, and maintaining mental workload below levels that could lead to errors or inappropriate decisions is important.

Public et al. (2010) began the process of creating the checklist with extensive field data collection with signalers using a range of data collection methods, including interviews, observations, and verbal protocol analysis. A structured interview technique (repertory grid technique) was used to elicit knowledge from rail signalers about the most significant elements identified during initial data collection with respect to influencing signaler mental workload. The identified elements were grouped into categories of operational infrastructure, indicators, process, and service pattern. The repertory grid technique was used to understand which constructs were most meaningful to subject matter experts in describing the overall construct of mental workload and how the elements identified earlier were most able to reflect the presence or absence of the constructs.

The results of the repertory grid technique were then used to develop the checklist. The scoring of high, medium, and low for each element was determined from previously collected data or based upon subject matter experts' opinions where data were lacking. Concurrent validity was assessed by comparing the results of the checklist to a "grading system familiar to the rail industry." The comparison tool graded signal boxes on the "responsibilities and decision-making required as a consequence of the infrastructure controlled and the service operated." Interrater reliability was assessed, but no statistics were presented. Training was developed in an attempt to increase interrater reliability.

9. *Line-Oriented Safety Audit* (*LOSA*). Over 15 years the civil aviation industry has developed a safety audit program for cockpit crew. This was initially developed as a way to check whether the discipline of crew resource management (CRM: e.g., Helmreich et al., 2001) had taken hold in cockpit crews during line operations (Klinect et al., 2003). This was an in-cockpit observation method using trained observers (human factors engineers or experienced airline pilots). The methodology was later refined as a result of new research on threat error management (TEM) and is now an FAA Advisory Circular (FAA, 2006). The actual data collection methodology is quite different from the typical checklist, although it does have boxes to fill in. These are narrative boxes for each stage of flight: predeparture/ axi, take-off/climb, cruise, descent/approach/land/taxi. These raw data are verified by a data verification round table, which can take several days. From this verification comes a set of errors in each of several standard categories, finally being presented stripped of all identification material as error rates in each category. A formal reliability test of LOSA was conducted using the LOSA observer feedback form (LOFF) with two coding exercises using 116 trained observers. The Kudler–Richardson KR-20 coefficient was measured as 0.70, while the split-half Spearman–Brown coefficient was 0.88, both indications of high reliability (Klinect, 2005, pp. 70–71). Currently the LOSA is being extended to airline ground, ramp, and maintenance operations (Ma et al., 2011).

10. Smith (2001) described a production line audit designed to establish a baseline of current ergonomics activities and priorities. The approach used observational task analysis in conjunction with questionnaires and interviews. Significant musculoskeletal stressors were selected as the focus of the audit program, and thus this audit was not designed to audit the potential broad spectrum of ergonomics issues that could be present in an assembly environment. The audit relied on observational analyses of risk factors for each body part, supplemented by body part discomfort maps filled out by line employees, although no details of reliability or validity are given.

11. *Other Checklists.* The sample of successful audit checklists above has been presented in some detail to provide the reader with their philosophy, structure, and sample questions. Rather than continue in the same vein, other interesting checklists are outlined in Table 9. Each entry shows the domain, the types of issues addressed, the size or time taken in use, and whether validity and reliability have been measured. Most textbooks now provide checklists, and a few of these are cited. No claim is made that Table 9 is comprehensive; rather, it is a sampling with references so that readers can find a suitable match to their needs. The first nine entries in the table are conveniently colocated in Landau and Rohmert (1989). Many of their reliability and validity studies are reported in this publication. The next entries are results of the Commission of European Communities fifth ECSC program, reported in Berchem-Simon (1993). Others are from texts and original references. The author has not personally used all of these checklists and so cannot endorse them specifically. Also, omission of a checklist from this table implies nothing about its usefulness.

**Other Data Collection Methods.** Not all data come from checklists and questionnaires. We can audit a human factors program using outcome measures alone. However, outcome measures such as injuries, quality, and productivity are nonspecific to human factors: Many other external variables can affect them. An obvious example is changes in the reporting threshold for injuries, which can lead to sudden apparent increases and decreases in the safety of a department or plant.

**Table 9  Selection of Published Checklists**

| Name | Reference | Coverage | Reliability | Validity |
|---|---|---|---|---|
| TBS | Hacker et al. (1983) | Mainly mental work | | Vs. AET |
| VERA | Volpert et al. (1983) | Mainly mental work | | Vs. AET |
| RNUR | RNUR (1976) | Mainly physical work | | |
| LEST | Guélaud (1975) | Mainly physical work | | |
| AVISEM | AVISEM (1977) | Mainly physical work | | |
| GESIM | GESIM (1988) | Mainly physical work | | |
| RHIA | Leitner and Greiner (1989) | Task hindrances, stress | 0.53–0.79 | Vs. many |
| MAS | Groth (1989) | Open structure, derived from AET | | Vs. AET |
| JL and HA | Mattila and Kivi (1989) | Mental, physical work, hazards | 0.87–0.95 | |
| | Bolijn (1993) | Physical work for women | Tested | |
| | Panter (1993) | Load handling | | |
| | Portillo Sosa (1993) | VDT standards | | |
| Work analysis | Pulat (1992) | Mental and physical work | | |
| Thermal audit | Parsons (1992) | Thermal audit from heat balance | | Content |
| WAS | Yoshida and Ogawa (1991) | Workplace and environment | Tested | Vs. expert |
| Ergonomics | SHARE (1990) | Short workplace checklists | | |
| | Cakir et al. (1980) | VDT checklist | | |
| | Nery (1999) | Meat processing | | |
| | Robson and Wolstenhulme (2010) | Ultrasound scan room | | |

*Source*: First nine from Landau and Rohmert (1989), next three from Berchem-Simon (1993)

Additionally, injuries are (or should be) extremely rare events. Thus, to obtain enough data to perform meaningful statistical analysis may require aggregation over many disparate locations andz/or time periods. In ergonomics audits, such outcome measures are perhaps best left for long-term validation or for use in selecting cluster samples. An example is the "validation" of LOSA by using it to measure the improvements in error reduction after a CRM program was instituted. Ma et al. (2010) quote Croft (2001) showing a 59% decline in unstabilized approaches as measured by LOSA after a CRM program at Continental Airlines.

Besides outcome measures, interviews represent a possible data collection method. Whether directed or not (e.g., Sinclair, 1990), they can produce critical incidents, human factors examples, or networks of communication (e.g., Drury, 1990a) that have value as part of an audit procedure. Interviews are used routinely as part of design audit procedures in large-scale operations such as nuclear power plants (Kirwan, 1989) or naval systems (Malone et al., 1988).

A novel interview-based audit system was proposed by Fox (1992) based on methods developed by British Coal (reported by Simpson, 1994). Here an error-based approach was taken using interviews and archival records to obtain a sampling of actual and possible errors. These were then classified using Reason's (1990) active/latent failure scheme and orthogonally by Rasmussen's (1987) skill-, rule-, and knowledge-based framework. Each active error is thus a conjunction of skill/mistake/violation with skill/rule/knowledge. Within each conjunction, performance-shaping factors can be deduced and sources of management intervention listed. This methodology has been used in a number of mining-related studies (see Section 5.4.2).

It is worth mentioning that the term ergonomics audit is occasionally used by consultants in reference to assessing elements of ergonomics programs either corporatewide or at individual sites. The audits do not necessarily assess the ergonomics of tasks, machines/equipment, or environment but rather assess whether ergonomics processes such as individual workstations are carried out, whether surveillance for injuries is carried out, the nature of policies and procedures, and so on. These are typically carried out through interviews. Although these may be effective for increasing the quality of an ergonomics program, they do not necessarily measure effectiveness of the program.

### 5.3.3  Data Analysis and Presentation

Human factors as a discipline covers a wide range of topics from workbench height to function allocation in automated systems. An audit program can only hope to abstract and present a part of this range. With our consideration of sampling systems and data collection devices we have seen different ways in which an unbiased abstraction can be aided. At this stage the data consist of large numbers of responses to large numbers of checklist items or detailed interview findings. How can, or should, these data be treated for best interpretation?

Here there are two opposing viewpoints: One is that the data are best summarized across sample units but not across topics. This is typically the way the human factors professional community treats the data, giving summaries in published papers of the distribution of responses to individual items on the checklist. In this way, findings can be more explicit: for example, that the lighting is an area that needs ergonomics effort or that

the seating is generally poor. Adding together lighting and seating discrepancies is seen as perhaps obscuring the findings rather than assisting in their interpretation.

The opposite viewpoint, in many ways, is taken by the business community. For some, an overall figure of merit is a natural outcome of a human factors audit. With such a figure in hand, the relative needs of different divisions, plants, or departments can be assessed in terms of ergonomic and engineering effort required. Thus, resources can be distributed rationally from a management level. This view is heard by those in the manufacturing and service industries who after an audit ask "How did we do?" and expect a very brief answer. The proliferation of the spreadsheet, with its ability to sum and average rows and columns of data, has encouraged people to do just that with audit results. Repeated audits fit naturally into this view, as they can become the basis for monthly, quarterly, or annual graphs of ergonomic performance.

Neither view alone is entirely defensible. Of course, summing lighting and seating needs produces a result that is logically indefensible and that does not help diagnosis. But equally, decisions must be made concerning optimum use of limited resources. The human factors auditor, having chosen an unbiased sampling scheme and collected data on (presumably) the correct issues, is perhaps in an excellent position to assist in such management decisions. But so, too, are other stakeholders, primarily the workforce.

Audits are not, however, the only use of some of the data collection tools. For example, the Keyserling et al. (1993) upper extremity checklist was developed specifically as a screening tool. Its objective was to find which jobs/workplaces are in need of detailed ergonomic study. In such cases, summing across issues for a total score has an operational meaning (i.e., that a particular workplace needs ergonomic help).

Where interpretation is made at a deeper level than just a single number, a variety of presentation devices have been used. These must show scores (percent of workplaces, distribution of sound pressure levels, etc.) separately but so as to highlight broader patterns. Much is now known about separate versus integrated displays and emergent features (e.g., Wickens, 1992, pp. 121–122), but the traditional profiles and spider's web charts are still the most usual presentation forms. Thus, Wagner (1989) shows the AVISEM profile for a steel industry job before and after automation. The nine different issues ("rating factors") are connected by lines to show emergent shapes for the old and the new jobs. Landau and Rohmert's (1981) original book on AET shows many other examples of profiles. Klimer et al. (1989) present a spider web diagram to show how three work structures influenced 10 issues from the AET analysis. Mattila and Kivi (1989) present their data on the job load and hazard analysis system applied to the building industry in the form of a table. For six occupations, the rating on five different loads/hazards is presented as symbols of different sizes within the cells of the table.

There is little that is novel in the presentation of audit results: Practitioners tend to use the standard tabular or graphical tools. But audit results are inherently multidimensional, so that some thought is needed if the reader is to be helped toward an informed comprehension of the audit's outcome.

## 5.4   Audit Systems in Practice

Almost any of the audit programs and checklists referred to in previous sections give examples of their use in practice. Only two examples will be given here, as others are readily accessible. These examples were chosen as they represent quite different approaches to auditing.

### 5.4.1   Auditing a Decentralized Business

From 1992 to 1996, a major U.S.-based apparel manufacturer had run an ergonomics program aimed primarily at the reduction of workforce injuries in backs and upper extremities. As detailed in Drury et al. (1999), the company during that time comprised nine divisions and employed about 45,000 workers. Of particular interest was the fact that the divisions enjoyed great autonomy, with only a small corporate headquarters with a single executive responsible for all risk management activities. The company had grown through mergers and acquisitions, meaning that different divisions had different degrees of vertical integration. Hence, core functions such as sewing, pressing, and distribution were common to most divisions, while some also included weaving, dyeing, and embroidery. In addition, the products and fabrics presented quite different ergonomic challenges, from delicate undergarments, through heavy jeans, to knitted garments and even luggage.

The ergonomics program was similarly diverse. It started with a corporate launch by the highest level executives, then was rolled out to the divisions and to individual plants. The pace of change was widely variable. All divisions were given a standard set of workplace analysis and modification tools (based on Drury and Wick, 1984) but were encouraged to develop their own solutions to problems in a way appropriate to their specific needs.

Evaluation took place continuously, with regular meetings between representatives of plants and divisions to present results of before-and-after workplace studies. However, there was a need for a broader audit of the entire corporation aimed at understanding how much had been achieved for the multimillion-dollar investment, where the program was strong or weak, and what program needs were emerging for the future. During 1995, a team of auditors visited all nine divisions and a total of 12 plants spread across eight divisions. This was three years after the initial corporate launch and about two years after the start of shop-floor implementation.

A three-part audit methodology was used. First, a workplace survey was developed based on elements of the program itself, supplemented by direct comparisons to ergonomics standards and good practices. Table 10 shows this 50-item survey form, with data added for the percentage of "yes" answers where the responses were not measures or scale values. The workplace survey was given at a total of 157 workplaces across the 12 plants. Second, a user survey (Table 11) was used in an interview format with 66 consumers of ergonomics,

**Table 10 Ergonomics Audit: Workplace Survey with Overall Data**

| Number | | | Division | Plant | Job Type |
|--------|-----|-----|----------|-------|----------|
| | Yes | No | | Factor | |
| **1. Postural aspects** | | | | | |
| W1 | 68% | | Frequent extreme motions of back, neck, shoulders, wrists | | |
| W2 | 66% | | Elbows raised or unsupported more than 50% of time | | |
| W3 | 22% | | Upper limbs contact nonrounded edges | | |
| W4 | 73% | | Gripping with fingers | | |
| W5 | 36% | | Knee/foot controls | | |
| **1.1. Seated** | | | | | |
| W6 | 12% | | Leg clearance restricted | | |
| W7 | 21% | | Feet unsupported/legs slope down | | |
| W8 | 17% | | Chair/table restricts thighs | | |
| W9 | 22% | | Back unsupported | | |
| W10 | 37% | | Chair height not adjustable easily | | |
| **1.2. Standing** | | | | | |
| W11 | 3% | | Control requires weight on one foot more than 50% time | | |
| W12 | 37% | | Standing surface hard | | |
| W13 | 92% | | Work surface height not adjustable easily | | |
| **1.3. Hand tools** | | | | | |
| W14 | 77% | | Tools require hand/wrist bending | | |
| W15 | 9% | | Tools vibrate | | |
| W16 | 63% | | Restricted to one handed use | | |
| W17 | 39% | | Tool handle ends in palm | | |
| W18 | 20% | | Tool handle has nonrounded edges | | |
| W19 | 56% | | Tool uses only two or three fingers | | |
| W20 | 9% | | Requires continuous or high force | | |
| W21 | 41% | | Tool held continuously in one hand | | |
| **2. Vibration** | | | | | |
| W22 | 14% | | Vibration reaches body from any source | | |
| **3. Manual materials handling** | | | | | |
| W23 | 40% | | More than five moves per minute | | |
| W24 | 36% | | Loads unbalanced | | |
| W25 | 14% | | Lift above head | | |
| W26 | 28% | | Lift off floor | | |
| W27 | 83% | | Reach with arms | | |
| W28 | 78% | | Twisting | | |
| W29 | 60% | | Bending trunk | | |
| W30 | 3% | | Floor wet or slippery | | |
| W31 | 0% | | Floor in poor condition | | |
| W32 | 17% | | Area obstructs task | | |
| W33 | 4% | | Protective clothing unavailable | | |
| W34 | 2% | | Handles used | | |
| **4. Visual aspects** | | | | | |
| W35 | | | Task nature: 1, rough; 2, moderate; 3, fine; 4, very fine | | |
| W36 | | | Glare/reflection: 0, none; 1, noticeable; 2, severe | | |
| W37 | | | Color contrast: 0, none; 1, noticeable; 2, severe | | |
| W38 | | | Luminance contrast: 0, none; 1, noticeable; 2, severe | | |
| W39 | | | Task illuminance (foot candles) | | |
| W40 | 69% | | Luminance: task > midfield > outerfield = yes | | |
| **5. Thermal aspects** | | | | | |
| W41 | | | Dry-bulb temperature (°F) | | |
| W42 | | | Relative humidity (%) | | |

*(continues)*

**Table 10 (*Continued*)**

| Number | | | Division | Plant | Job Type |
|---|---|---|---|---|---|
| | Yes | No | | | Factor |
| W43 | | | | | Airspeed: 1, just perceptible; 2, noticeable; 3, severe |
| W44 | | | | | Metabolic cost |
| W45 | | | | | Clothing (clo value) |
| 6. Auditory aspects | | | | | |
| W46 | | | | | Maximum sound pressure level (dBA) |
| W47 | | | | | Noise sources: 1, m/c; 2, other m/c; 3, general; 4, other |
| 7. General factors | | | | | |
| W48 | | | | | Primary cycle time (seconds) |
| W49 | 62% | | | | Seen ergonomics video |
| W50 | 38% | | | | Any ergonomics changes to workplace or methods |

**Table 11  Ergonomics Audit: User Survey**

| Number | Division | Plant | Job Type |
|---|---|---|---|
| U1. What is ergonomics? | | | |
| U2. Who do you call to do ergonomics? | | | |
| U3. When did you last ask them to do ergonomics? | | | |
| U4. Describe what they did. | | | |
| U5. Who else should we talk to about ergonomics? | | | |
| U6. General comments on ergonomics. | | | |

**Table 12  Ergonomics Audit: Provider Survey**

| Number | Division | Plant | Job Type |
|---|---|---|---|
| P1. What do you do? | | | |
| P2. How do you get contacted to do ergonomics? | | | |
| P3. When were you last asked to do ergonomics? | | | |
| P4. Describe what you did. | | | |
| P5. How long have you been doing ergonomics? | | | |
| P6. How were you trained in ergonomics? | | | |
| P7. What percent of your time is spent on ergonomics? | | | |
| P8. Where do you go for more detailed ergonomics help? | | | |
| P9. What ergonomics implementation problems have you had? | | | |
| P10. How well are you regarded by management? | | | |
| P11. How well are you regarded by the workforce? | | | |
| P12. General comments on ergonomics. | | | |

typically plant managers, production managers, human resource managers, or their equivalent at the division level, usually vice presidents. Finally, a total of 27 providers of ergonomics services were given a similar provider survey (Table 12) interview. Providers were mainly engineers, with three human resources specialists and one line supervisor. From these three audit methods the corporation wished to provide a time snapshot of how effectively the current ergonomics program was meeting their needs for reduction of injury costs. While the workplace survey measured how well ergonomics was being implemented at the workplace, the user and provider surveys provided data on the roles of the decision makers beyond the workplace.

Detailed audit results are provided in Drury et al. (1999), so only examples and overall conclusions are covered in this chapter. Workplaces showed some evidence of good ergonomic practice, with generally satisfactory thermal, visual, and auditory environments. There were some significant differences ($p < 0.05$) between workplace types rather than between divisions or plants; for example, better lighting ($>700$ lux) was associated with inspection and sewing. Also, higher thermal load was associated with laundries and machine load/unload. Overall, 83% of workplaces met the ASHRAE (1990) summer comfort zone criteria. As shown in Table 13, the main ergonomics problem areas were in poor posture and manual materials handling. Where operators were seated (only 33% of all workplaces), seating was relatively good. In fact, many

of the workforce had been supplied with well-designed chairs as part of the ergonomics program.

To obtain a broad perspective, the three general factors at the end of Table 10 were analyzed. Apart from cycle time (W48), the questions related to workers having seen the corporate ergonomics video (W49) and having experienced a workplace or methods change (W50). Both should have received a "yes" response if the ergonomics program were reaching the entire workforce. In fact, both showed highly significant differences between plants, $X_8^2 = 92.0$, $p < 0.001$ and $X_8^2 = 22.2$, $p < 0.02$, respectively). Some of these differences were due to two divisions lagging in ergonomics implementation, but even beyond this there were large between-plant differences. Overall, 62% of the workforce had seen the ergonomics video, a reasonable value but one with wide variance between plants and divisions. Also, 38% of workplaces had experienced some change, usually ergonomics related,

**Table 13 Responses to Ergonomics User**

| Question and Concern | Corporate | | Plant | |
|---|---|---|---|---|
| | Mgt. | Staff | Mgt. | Staff |
| 1.   What is ergonomics? | | | | |
| 1.1.   Fitting job to operator | 1 | 6 | 10 | 5 |
| 1.2.   Fitting operator to job | 0 | 6 | 0 | 0 |
| 2.   Who do you call on to get ergonomics work done? | | | | |
| 2.1.   Plant ergonomics people | 0 | 3 | 3 | 2 |
| 2.2.   Division ergonomics people | 0 | 4 | 5 | 2 |
| 2.3.   Personnel department | 3 | 0 | 0 | 0 |
| 2.4.   Engineering department | 1 | 8 | 6 | 11 |
| 2.5.   We do it ourselves | 0 | 2 | 1 | 0 |
| 2.6.   College interns | 0 | 0 | 4 | 2 |
| 2.7.   Vendors | 0 | 0 | 0 | 1 |
| 2.8.   Everyone | 0 | 1 | 0 | 0 |
| 2.9.   Operators | 0 | 1 | 0 | 0 |
| 2.10.   University faculty | 0 | 0 | 1 | 0 |
| 2.11.   Safety | 0 | 1 | 0 | 0 |
| 3.   When did you last ask them for help? | | | | |
| 3.1.   Never | 0 | 4 | 2 | 0 |
| 3.2.   Sometimes/infrequently | 2 | 0 | 1 | 0 |
| 3.3.   1 year or more ago | 0 | 1 | 4 | 0 |
| 3.4.   1 month or so ago | 0 | 0 | 2 | 0 |
| 3.5.   Less than 1 month ago | 1 | 0 | 3 | 4 |
| 5.   Who else should we talk to about ergonomics? | | | | |
| 5.1.   Engineers | 0 | 0 | 3 | 2 |
| 5.2.   Operators | 1 | 1 | 2 | 0 |
| 5.3.   Everyone | 0 | 0 | 2 | 0 |
| 6.   General ergonomics comments | | | | |
| 6.1.   Ergonomics concerns | | | | |
| 6.1.1.   Workplace design for safety/ease/stress/fatigue | 2 | 5 | 13 | 5 |
| 6.1.2.   Workplace design for cost savings/productivity | 1 | 0 | 2 | 1 |
| 6.1.3.   Workplace design for worker satisfaction | 1 | 1 | 0 | 1 |
| 6.1.4.   Environment design | 2 | 1 | 3 | 0 |
| 6.1.5.   The problem of finishing early | 0 | 0 | 1 | 1 |
| 6.1.6.   The seniority/bumping problem | 0 | 3 | 1 | 0 |
| 6.2.   Ergonomics program concerns | | | | |
| 6.2.1.   Level of reporting of ergonomics | 0 | 1 | 7 | 0 |
| 6.2.2.   Communication/who does ergonomics | 7 | 1 | 4 | 0 |
| 6.2.3.   Stability/staffing of ergonomics | 0 | 0 | 10 | 4 |
| 6.2.4.   General evaluation of ergonomics | | | | |
| Positive | 1 | 3 | 3 | 4 |
| Negative | 4 | 10 | 10 | 3 |
| 6.2.5.   Lack of financial support for ergonomics | 0 | 0 | 1 | 0 |
| 6.2.6.   Lack of priority for ergonomics | 2 | 2 | 1 | 4 |
| 6.2.7.   Lack of awareness of ergonomics | 2 | 1 | 6 | 1 |

a respectable figure after only two to three years of the program.

From the user and provider surveys, an enhanced picture emerged. Again, there was variability between divisions and plants, but 94% of the users defined ergonomics as fitting the job to the operator rather than training or medical management of injuries. Most users had requested an ergonomic intervention within the past two months, but other "users" had never in fact used ergonomics.

The solutions employed ranged widely, with a predominance of job aids such as chairs or standing

pads. Other frequent categories were policy changes (e.g., rest breaks, rotation, box weight reduction) and workplace adjustment to the individual operator. There were few uses of personal aids (e.g., splints) or referrals to physicians as ergonomic solutions. Changes to the workplace clearly predominated over changes to the individual, although a strong medical management program was in place when required. When questioned about ergonomics results, all mentioned safety (or workplace comfort or ease of use), but some also mentioned others. Cost or productivity benefits were the next most common response, with a few additional ones relating to employee relations, absence/turnover, or job satisfaction. Significantly, only one respondent mentioned quality.

The major user concern at the plant level was time devoted to ergonomics by providers. At the corporate level, the need was seen for more rapid job analysis methods and corporate policies (e.g., on back belts or "good" chairs). Overall, 94% of users made positive comments about the ergonomics program.

Ergonomics providers were almost always trained in the corporate or division training seminars, usually near the start of the program. Providers' chief concern was for the amount of time and resources they could spend on ergonomics activities. Typically, ergonomics was only one job responsibility among many. Hence, broad programs, such as new chairs or back belts, were supported enthusiastically, as they gave the maximum perceived impact for the time devoted. Other solutions presented included job aids, workplace redesign (e.g., moving from seated to standing jobs for long-seam sewing), automation, rest breaks, job rotation, packaging changes, and medical management. Specific needs were seen in the area of corporate or supplier help in obtaining standard equipment solutions and of more division-specific training. As with users, the practitioners enjoyed their ergonomics activity and thought it worthwhile.

Recommendations arising from this audit were that the program was reasonably effective at present but had some long-term needs. The corporation sees itself as an industry leader and wants to move beyond a relatively superficial level of ergonomics application. To do this will require more time resources for job analysis and change implementation. Corporate help could also be provided in developing more rapid analysis methods, standardized video-based training programs, and more standardized solutions to recurring ergonomics problems. Many of these changes have since been implemented.

On another level, the audit was a useful reminder to the company of the fact that it had incurred most of the up-front costs of a corporate ergonomics program and was now beginning to reap the benefits. Indeed, by 1996, corporate injury costs and rates had decreased by about 20% per year after peaking in 1993. Clearly, the ergonomics program was not the only intervention during this period, but it was seen by management as the major contributor to improvement. Even on the narrow basis of cost savings, the ergonomics program was a success for the corporation.

## 5.4.2  Error Reduction at a Colliery

In a two-year project reported by Fox (1992) and Simpson (1994), the human error audit described in Section 5.3.2 was applied to two colliery haulage systems. The results of the first study are presented here. In both systems, data collection focused on potential errors and the performance-shaping factors (PSFs) that can influence these errors. Data were collected by "observation, discussion and measurement within the framework of the broader man–machine systems and checklist of PSFs," taking some 30–40 shifts at each site. The entire haulage system from surface operations to delivery at the coal face was covered.

The first study found 40 active failures (i.e., direct error precursors) and 9 latent failures (i.e., dormant states predisposing the system to later errors). Four broad classes of active failures were (1) errors associated with locomotive maintenance (7 errors) (e.g., fitting incorrect thermal cutoffs), (2) errors associated with locomotive operation (10 errors) (e.g., locomotives not returned to the service bay for a 24-h check), (3) errors associated with loads and load security (7 errors), (e.g., failure to use spacer wagons between overhanging loads), and (4) errors associated with the design/operation of the haulage route (10 errors), (e.g., continued use despite potentially unsafe track) plus a small miscellaneous category.

The latent failures were (Fox, 1992) (1) quality assurance in supplying companies, (2) supply-ordering procedures within the colliery, (3) locomotive design, (4) surface "makeup" of supplies, (5) lack of equipment at specific points, (6) training, (7) attitudes to safety, and (8) the safety inspection/reporting/action procedures. As an example of item 3, locomotive design, the control positions were not consistent across the locomotive fleet, despite all originating from the same manufacturer. Using the slip/mistake/violation categorization, each potential error could be classified so that the preferred source of action (intervention) could be specified.

This audit led to the formation of two teams, one to tackle locomotive design issues and the other for safety reporting and action. As a result of team activities, many ergonomic actions were implemented. These included management actions to ensure a uniform wagon fleet, autonomous inspection/repair teams for tracks, and multifunctional teams for safety initiatives.

The outcome was that the accident rate dropped from 35.40 per 100,000 person-shifts to 8.03 in one year. This brought the colliery from worst in the regional group of 15 collieries to best in the group and indeed in the United Kingdom. In addition, personnel indicators, such as industrial relations climate and absence rates, improved.

## 6  CONCLUSIONS

In this chapter we have arrived at human factors audits through a context of inspection and checklist design. It should be obvious by now that checklists are a subset of audits, which are in turn a subset of inspection. Within the context of inspection, we have seen that all inspections follow a short logical sequence of functions

and that each function has considerable scope for model-based and empirical design to improve the human factors and system performance. Nonmanufacturing applications have been emphasized, with the focus on processes and broader systems rather than on repetitively produced products. Audits have been shown to be functionally similar to inspections.

Inspecting, checking, and auditing are interesting, as they all have human factors design aspects but can all be applied to both the processes being audited and the auditing process itself. Whether inspecting nonmanufacturing items or checking items on a checklist or performing an audit, there is prescriptive advice on how to develop or choose a system that accords with human factors good practices.

**Disclaimer**: The findings and conclusions in this chapter are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

## REFERENCES

Alexander, D. C., and Pulat, B. M. (1985), *Industrial Ergonomics: A Practitioner's Guide*, Industrial Engineering and Management Press, Atlanta, GA.

American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) (1989), "Physiological Principles, Comfort and Health," in *Fundamentals Handbook*, ASHRAE, Atlanta, GA, Chapter 8.

American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) (1990), *Handbook of Fundamentals*, ASHRAE, Atlanta, GA.

Arani, T., Karwan, M. H., and Drury, C. G. (1984), "A Variable-Memory Model of Visual Search," *Human Factors*, Vol. 26, No. 6, pp. 631–639.

AVISEM (1977), *Techniques d'amélioration des conditions de travail dans l'industrie*, Editions Hommes et Techniques, Suresnes, France.

Ayoub, M. M., and Mital, A. (1989), *Manual Materials Handling*, Taylor & Francis, London.

Bainbridge, L., and Beishon, R. J. (1964), "The Place of Checklists in Ergonomic Job Analysis," in *Proceedings of the 2nd I.E.A. Congress*, Dortmund, Ergonomics Congress Proceedings Supplement.

Baveja, A., Drury, C. G., Karwan, M., and Malone, D. M. (1996), "Derivation and Test of an Optimum Overlapping-Lobe Model of Visual Search," *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. 28, pp.161–168.

Berchem-Simon, O., Ed. (1993), *Ergonomics Action in the Steel Industry*, EUR 14832 EN, Commission of the European Communities, Luxembourg.

Blackwell, H. R. (1946), "Contrast Thresholds of the Human Eye," *Journal of the Optical Society of America*, Vol. 36, pp. 624–643.

Bolijn, A. J. (1993), *Research into the Employability of Women in Production and Maintenance Jobs in Steelworks*, Commission of the European Communities, Luxembourg.

British Standards Institution (BSI) (1980), "Educational furniture," *Specification for Functional Dimensions, Identification and Finish of Chairs and Tables for Educational Institutions*, BS5873-1: 1980, BSI, London.

Burger, G. C. E., and de Jong, J. R. (1964), "Evaluation of Work and Working Environment in Ergonomic Terms," *Aspects of Ergonomic Job Analysis*, Vol. 7, pp. 185–201.

Cakir, A., Hart, D. M., and Stewart, T. F. M. (1980), *Visual Display Terminals*, Wiley, New York.

Carson, A. B., and Carlson, A. E. (1977), *Secretarial Accounting*, 10th ed., South-Western, Cincinnati, OH.

Chan, A. H. S., and Chiu, C. H. Y., (2010), "Visual Lobe Shape Characteristics of Experienced Inspectors and Inexperienced Subjects," *Human Factors and Ergonomics in Manufacturing and Service Industries*, Vol. 20.5, pp. 367–377.

Chen, Y., Gale, A., and Scott, H. (2008), "Anytime, Anywhere Mammographic Interpretation Training," in *Contemporary Ergonomics 2008*, P. D. Bust, Ed., Taylor & Francis, London, pp. 375–380.

Chervak, S., and Drury, C. G. (1995), "Simplified English Validation," in *Human Factors in Aviation Maintenance: Phase 6 Progress Report*, DOT/FAA/AM-95/xx, Federal Aviation Administration, Office of Aviation Medicine, National Technical Information Service, Springfield, VA.

Chervak, S., Drury, C. G., and Ouellette, J. L. (1996), "Simplified English for Aircraft Workcards," in *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*, pp. 303–307.

Chi, C.-F., and Drury, C. G. (1995), "A Test of Economic Models of Stopping Policy in Visual Search," *IIE Transactions*, Vol. 27, pp. 392–393.

Croft, J. (2001). Research perfect new ways to monitor pilot performance. *Aviation Week & Space Technology*, 155, 76.

Degani, A., and Wiener, E. L. (1990), *Human Factors of Flight-Deck Checklists: The Normal Checklist*, Contractor Report 177548, NASA Ames Research Center, Moffett Field, CA.

Dempsey, P. G., McGorry, R. W., and Maynard, W. S. (2005), "A Survey of Tools and Methods Used by Certified Professional Ergonomists," *Applied Ergonomics*, Vol. 36, No. 4, pp. 489–503.

Devor, R. E., Change, T.-H., and Sutherland, J. W. (1992), *Statistical Quality Design and Control*, Prentice-Hall, Englewood Cliffs, NJ.

Dirken, J. M. (1969), "An Ergonomics Checklist Analysis of Printing Machines, *ILO (Geneva)*, Vol. 2, pp. 903–913.

Drury, C. G. (1978), "Integration of Human Factors Models into Statistical Quality Control," *Human Factors*, Vol. 20, No. 5, pp. 561–572.

Drury, C. G. (1990a), "The Ergonomics Audit," in *Contemporary Ergonomics*, E. J. Lovesey, Ed., Taylor & Francis, London, pp. 400–405.

Drury, C. G. (1990b), "Computerized Data Collection in Ergonomics," in *Evaluation of Human Work*, J. R. Wilson and E. N. Corlett, Eds., Taylor & Francis, London, pp. 200–214.

Drury, C. G. (1992a), "Inspection Performance," in *Handbook of Industrial Engineering*, G. Salvendy, Ed., Wiley, New York, pp. 2282–2314.

Drury, C. G. (1992b), "Design for Inspectability," in *Design for Manufacturability: A Systems Approach to Concurrent Engineering and Ergonomics*, M. H. Helander and M. Nagamachi, Eds., Taylor & Francis, London, pp. 204–216.

Drury, C. G. (1997), "Ergonomics Audits: Why and How," in *Proceedings of the 13th Triennial Congress of the International Ergonomics Association'97*, Vol. 2, Tampere, Finland, pp. 284–286.

Drury, C. G. (2002), "A Unified Model of Security Inspection," in *Proceedings of the Aviation Security Technical Symposium*, Federal Aviation Administration, Atlantic City, NJ, pp. 27–30.

Drury, C. G. (2003, April–May), "Service Quality and Human Factors," *AI and Society*, Vol. 17, No. 2, pp. 78–96.

Drury, C. G. (2009), "Inspection: From Goods through Passengers to Systems," in P. D. Bust (Ed) *Contemporary Ergonomics 2009*, Taylor and Francis, London, pp. 3–15.

Drury, C. G., and Addison, J. L. (1973), "An Industrial Study of the Effects of Feedback and Fault Density on Inspection Performance," *Ergonomics*, Vol. 16, pp. 159–169.

Drury, C. G., and Sinclair, M. A. (1983), "Human and Machine Performance in an Inspection Task," *Human Factors,* Vol. 25, No. 4, pp. 391–400.

Drury, C. G., and Wick, J. (1984), "Ergonomic Applications in the Shoe Industry," in *Proceedings of the International Conference on Occupational Ergonomics*, Vol. 1, pp. 489–483.

Drury, C. G., Kleiner, B. M., and Zahorjan, J. (1989), "How Can Manufacturing Human Factors Help Save a Company: Intervention at High and Low Levels," in *Proceedings of the Human Factors Society 33rd Annual Meeting*, Denver, CO, pp. 687–689.

Drury, C. G., Prabhu, P., and Gramopadhye, A. (1990), "Task Analysis of Aircraft Inspection Activities: Methods and Findings," in *Proceedings of the Human Factors Society 34th Annual Meeting*, Santa Monica, CA, pp. 1181–1185.

Drury, C. G., Spencer, F. W., and Schurman, D. (1997), "Measuring Human Detection Performance in Aircraft Visual Inspection," in *Proceedings of the 41st Annual Human Factors and Ergonomics Society Meeting*, Albuquerque, NM.

Drury, C. G., Broderick, R. L., Weidman, C. H., and Reynolds Mozrall, J. L. (1999), "A Corporate-Wide Ergonomics Programme: Implementation and Evaluation," *Ergonomics*, Vol. 42, No. 1, pp. 208–228.

Drury, C. G., Patel, S. C., and Prabhu, P. V. (2000), "Relative Advantage of Portable Computer-Based Workcards for Aircraft Inspection," *International Journal of Industrial Ergonomics*, Vol. 26, pp. 163–176.

Drury, C. G., Ghylin, K. M., and Schwaninger, A. (2007), "Large-Scale Validation of a Security Inspection Model," in P. D. Bust (Ed) *Contemporary Ergonomics 2007*, Taylor & Francis, London

Easterby, R. S. (1967), "Ergonomics Checklists: An Appraisal," *Ergonomics*, Vol. 10, No. 5, pp. 548–556.

Engel, F. L. (1971), "Visual Conspicuity, Directed Attention, and Retinal Locus," *Visual Research*, Vol. 11, pp. 563–576.

Eriksen, C. W. (1990), "Attentional Search of the Visual Field," in *Visual Search*, D. Brogan, Ed., Taylor & Francis, London, pp. 3–19.

Evans, J. R., and Lindsay, W. (1993), *The Management and Control of Quality*, West Publishing Company, Minneapolis/St. Paul, MN.

Fanger, P. O. (1970), *Thermal Comfort, Analyses and Applications in Environmental Engineering*, Danish Technical Press, Copenhagen, Denmark.

Federal Aviation Administration (FAA) (2006), "Line Operations Safety Audits," Advisory Circular AC 120-90, FAA, Washington, DC.

Fox, J. G. (1992), "The Ergonomics Audit as an Everyday Factor in Safe and Efficient Working," *Progress in Coal, Steel and Related Social Research*, Vol. 11, pp. 10–14.

Gale, A. G., Mugglestone, M. D., Purdy, K. J., and McClumpha, A. (2000), "Is Airport Baggage Inspection Just Another Medical Image?," *Proceedings of the SPIE*, Vol. 3981, pp. 184–192.

Ghylin, K.M., Drury, C.G., Batta, R. & Lin, L. (2007). Temporal Effects in a Security Inspection Task: Breakdown of Performance Components. *Proceedings of the Human Factors and Ergonomics Society 51st Annual Meeting*, Oct. 1-5, 2007, 93–97, Baltimore, MD

Givoni, B., and Goldman, R. F. (1972), "Predicting Rectal Temperature Response to Work, Environment, and Clothing," *Journal of Applied Physiology*, Vol. 32, No. 6, pp. 812–822.

Goldberg, J. H., and Gibson, D. C. (1986), "The Effects of Training Method and Type of Checklist upon Visual Inspection Accuracy," in *Proceedings of the Annual International Industrial Ergonomics and Safety Conference*, Louisville, KY, North-Holland, Amsterdam.

Groth, K. M. (1989), "The Modular Work Analysis System (MAS)," in *Recent Developments in Job Analysis, Proceedings of the International Symposium on Job Analysis*, Taylor & Francis, New York, pp. 253–261.

Guélaud, F., Beauchesne, M.-N., Gautrat, J., and Roustang., G. (1975), "Pour une analyse des conditions de travail ouvrier dans l'entreprise," in *Recherche du Laboratoire d'Economie et de Sociologie du Travail C.M.R.S. Aix-en-Provence*, 3rd ed., Librairie Armand Colin, Paris.

Hacker, W., Iwanowa, A., and Richter, P. (1983), "Tätigkeitsbewertungssystem," paper presented at the Psychodiagnostisches Zentrum, Berlin (OST).

Hasslequist, R. J. (1981), "Increasing Manufacturing Productivity Using Human Factors Principles," in *Proceedings of the Human Factors Society 25th Annual Meeting*, Santa Monica, CA, pp. 204–206.

Helmreich, R. L., Wilhelm, J. A., Klinect, J. R., and Merritt, A. C. (2001), "Culture, Error and Crew Resource Management," in E. Salas, C. A. Bowers, and E. Edens, Eds., *Applying Resource Management in Organizations: A Guide for Professionals*, Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 305–331.

Hong, S.-K., and Drury, C. G. (2002), "Sensitivity and Validity of Visual Search Models for Multiple Targets," *Theoretical Issues in Ergonomic Science*, Vol. 3, No. 1, pp. 1–26.

Hou, T.-S., Lin, L., and Drury, C. G. (1993), "An Empirical Study of Hybrid Inspection Systems and Allocation of Inspection Function," *International Journal of Human Factors in Manufacturing*, Vol. 3, pp. 351–367.

Hsiao, Y-L, Joseph, C., Joshi, S., Lapp, J., Pennathur, P. R., and Drury, C. G. (2008), "Posture and Performance: Sitting vs. Standing for Security Screening," *Ergonomics,* Vol. 51, pp. 290–307.

Illuminating Engineering Society (IES) (1993), *Lighting Handbook, Reference and Application*, 8th ed., IES of North America, New York.

International Civil Aviation Organization (ICAO) (1989), *Human Factors Digest No. 1 Fundamental Human Factors Concepts*, Circular 216-AN/131, ICAO, Montreal, Canada.

International Labour Office (ILO) (1999), *Ergonomic Check-points*. ILO, Geneva.

Jamieson, G. H. (1966), "Inspection in the Telecommunications Industry: A Field Study of Age and Other Performance Variables," *Ergonomics*, Vol. 9, pp. 297–303.

Jiang, X., Melloy, B. J., and Gramopadhye, A. (2003), "Evaluation of Best System Performance: Human, Automated and Hybrid Inspection Systems," *Human Factors and Ergonomics in Manufacturing,* Vol. 13, No. 2, pp. 137–152.

Jones, J. C. (1981), *Design Methods: Seeds of Human Factors*, Wiley, New York.

Jones, D. M., and Broadbent, D. E. (1987), "Noise," in *Handbook of Human Factors Engineering*, G. Salvendy, Ed., Wiley, New York.

Karwowski, W., (2005), *Handbook of Standards and Guidelines in Ergonomics and Human Factors (Human Factors/Ergonomics)*, Lawrence Erlbaum Associates, Mahwah, NJ.

Kerlinger, F. N. (1964), *Foundations of Behavioral Research*, Holt, Rinehart and Winston, New York.

Keyserling, W. M., Brouwer, M., and Silverstein, B. A. (1992), "A Checklist for Evaluation of Ergonomic Risk Factors Resulting from Awkward Postures of the Legs, Trunk and Neck," *International Journal of Industrial Ergonomics*, Vol. 9, No. 4, pp. 283–301.

Keyserling, W. M., Stetson, D. S., Silverstein, B. A., and Brouwer, M. L. (1993), "A Checklist for Evaluating Ergonomic Risk Factors Associated with Upper Extremity Cumulative Trauma Disorders," *Ergonomics*, Vol. 36, No. 7, pp. 807–831.

Kirwan, B. (1989), "A Human Factors and Human Reliability Programme for the Design of a Large UK Nuclear Chemical Plant," in *Proceedings of the Human Factors Society 33rd Annual Meeting*, Denver, CO, pp. 1009–1013.

Klimer, F., Kylian, H., Schmidt, K.-H., and Rutenfranz, J. (1989), "Work Analysis and Load Components in an Automobile Plant after the Implementation of New Technologies," in *Recent Developments in Job Analysis*, K. Landau and W. Rohmert, Eds., Taylor & Francis, New York, pp. 331–340.

Klinect, J. R. (2005), Line "Operations Safety Audit: A Cockpit Observation Methodology for Monitoring Commercial Airline Safety Performance," Ph.D. Dissertation, University of Texas at Austin, 2005

Klinect, J. R., Murray, P., Merritt, A., and Helmreich, R. (2003), "Line Operations Safety Audit (LOSA): Definition and Operating Characteristics," in *Proceedings of the 12th International Symposium on Aviation Psychology*, The Ohio State University, Dayton, OH, pp. 663–668.

Kogi, K. (1994), "Introduction to WISE (Work Improvement in Small Enterprises) Methodology and Workplace Improvements Achieved by the Methodology in Asis," in S. McFadden, L. Innes, and M. Hill, Eds., *Proceedings of the 12th Triennial Congress of the International Ergonomics Association*, Vol. 5, Human Factors Association of Canada, Toronto, Ontario, Canada, pp. 141–143.

Kogi, K., and Kuorinka, I., Eds. (1995), *Ergonomic Checkpoints*, ILO Publications, Geneva, Switzerland.

Koli, S., Drury, C. G., Cuneo, J., and Lofgren, J. (1993), "Ergonomic Audit for Visual Inspection of Aircraft," in *Human Factors in Aviation Maintenance: Phase Four, Progress Report*, DOT/FAA/AM-93/xx, National Technical Information Service, Springfield, VA.

Koli, S. T. (1994), "Ergonomic Audit for Non-Repetitive Task," M.S. Thesis, State University of New York at Buffalo, Buffalo, NY.

Lamar, E. S. (1960), "Operational Background and Physical Considerations Relative to Visual Search Problems," in *Visual Search Techniques*, A. Morris and P. E. Horne, Eds., National Research Council, Washington, DC.

Landau, K., and Rohmert, W. (1981), *Fallbeispiele zur Arbeitsanalyse*, Hans Huber, Bern, Switzerland.

Landau, K., and Rohmert, W., Eds. (1989), *Recent Developments in Job Analysis: Proceedings of the International Symposium on Job Analysis*, University of Hohenheim, March 14–15, Taylor & Francis, New York.

Larock, B. (2000), "The Effects of Spatial and Functional Job-Aid Checklists on a Frequently Performed Aircraft Maintenance Task," Master's Thesis, State University of New York at Buffalo, Buffalo, NY.

Larock, B., and Drury, C. G. (2003), "Repetitive Inspection with Checklists: Design and Performance," in *Proceedings of the Ergonomics Society Annual Conference*, Edinburgh, Scotland, April 15–17.

Leitner, K., and Greiner, B. (1989), "Assessment of Job Stress: The RHIA Instrument," in *Recent Developments in Job Analysis, Proceedings of the International Symposium on Job Analysis*, K. Landau and W. Rohmert, Eds., University of Hohenheim, March 14–15, 1989, Taylor & Francis, New York.

Lloyd, C. J., Boyce, P., Ferzacca, N., Eklund, N., and He, Y. (2000), "Paint Inspection Lighting: Optimization of Lamp Width and Spacing," *Journal of the Illuminating Engineering Society,* Vol. 28, pp. 99–102

Ma, J., Pedigo, M., Blackwell, L., Hackworth, C., Holcomb, K., and Gildea, K. and Hiles, J. J., (2011), "The Line Operations Safety Audit Program:Transitioning rom Flight Operations to Maintenance and Ramp Operations," DOT/FAA/AM-11/15, Office of Aerospace Medicine, Washington, DC 20591

Malde, B. (1992), "What Price Usability Audits? The Introduction of Electronic Mail into a User Organization," *Behaviour and Information Technology,* Vol. 11, No. 6, pp. 345–353.

Malone, T. B., Baker, C. C., and Permenter, K. E. (1988), "Human Engineering in the Naval Sea Systems Command," in *Proceedings of the Human Factors Society 32nd Annual Meeting*, Anaheim, CA, Vol. 2, pp. 1104–1107.

Mattila, M., and Kivi, P. (1989), "Job Load and Hazard Analysis: A Method for Hazard Screening and Evaluation," in *Recent Developments in Job Analysis: Proceedings of the International Symposium on Job Analysis*, K. Landau and W. Rohmert, Eds., University of Hohenheim, March 14–15, Taylor & Francis, New York, pp. 179–186.

McClelland, I. (1990), "Product Assessment and User Trials," in *Evaluation of Human Work*, J. R. Wilson and E. N. Corlett, Eds., Taylor & Francis New York, pp. 218–247.

McCormick, E. J. (1979), *Job Analysis: Methods and Applications*, AMACOM, New York.

McCormick, W. T., Mecham, R. C., and Jeanneret, P. R. (1969), *The Development and Background of the Position Analysis Questionnaire*, Occupational Research Center, Purdue University, Lafayette, IN.

McKenzie, R. M. (1958), "On the Accuracy of Inspectors," *Ergonomics*, Vol. 1, No. 3, pp. 258–272.

Miller, R. B. (1967), "Task Taxonomy: Science or Art," in *The Human Operator in Complex Systems*, W. T. Singleton, R. S. Easterby, and D. Whitfield, Eds., Taylor & Francis, London.

Mir, A. H. (1982), "Development of Ergonomic Audit System and Training Scheme," M.S. Thesis, State University of New York at Buffalo, Buffalo, NY.

Morawski, T., Drury, C. G., and Karwan, M. H. (1980), "Predicting Search Performance for Multiple Targets," *Human Factors*, Vol. 22, No. 6, pp. 707–718.

National Institute for Occupational Safety and Health (NIOSH) (1981), *Work Practices Guide for Manual Lifting*, DHEW-NIOSH Publication 81–122, NIOSH, Cincinnati, OH.

Nery, D. (1999), "Audit Tool User Guide: Ergonomic Hazard Management for the Meat Industry in South Australia," WorkCover Corporation, Adelaide, Australia, available: http://www.safework.sa.gov.au/contentPages/docs/meatUserGuide.pdf.

Nodine, C. F., Mello-Thoms, C., Kundel, H. L., and Weinstein, S. P. (2002), "Time Course of Perception and Decision Making during Mammographic Interpretation," *American Journal of Roentgenology*, Vol. 179, pp. 917–923.

Occupational Safety and Health Administration (OSHA) (1990), *Ergonomics Program Management Guidelines for Meatpacking Plants*, OSHA-3121, OSHA, U.S. Department of Labor, Washington, DC.

Ockerman, J. J., and Pritchett, A. R. (1998), "Preliminary Investigation of Wearable Computers for Task Guidance in Aviation Inspection," paper presented at the Second International Symposium on Wearable Computers, Pittsburgh, PA.

Ockerman, J. J., and Pritchett, A. R. (2000), "Reducing Over-Reliance on Task-Guidance Systems," in *Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting*, San Diego, CA.

Ockerman, J. J., and Pritchett, A. R. (2004), "Improving Performance on Procedural Tasks through Presentation of Locational Procedure Context: An Empirical Evaluation," *Behaviour & IT*, Vol. 23, No. 1, pp. 11–20.

Osburn, H. G. (1987), "Personnel Selection," in G. Salvendy, Ed., *Handbook of Human Factors*, Wiley, New York, pp. 911–933.

Overington, I. (1973), *Vision and Acquisition*, Pentech Press, London.

Pai, S. (2003), "Effectiveness of Technological Aids in Checklist Use," Master's Thesis, State University of New York at Buffalo, Buffalo, NY.

Panter, W. (1993), "Biomechanical Damage Risk in the Handling of Working Materials and Tools—Analysis, Possible Approaches and Model schemes," in *Ergonomics Action in the Steel Industry EUR 14832 EN*, O. Berchem-Simon, Ed., Commission of the European Communities, Luxembourg.

Park, J.-K., Boyer, J., Tessler, J., Casey, J., Schemm, L., Gore, R., Punnett, L., and Promoting Healthy and Safe Employment (PHASE) in Healthcare Project Team (2009), "Inter-Rater Reliability of PATH Observations for Assessment of Ergonomic Risk Factors in Hospital Work," *Ergonomics*, Vol. 52, pp. 820–829.

Parsons, K. C. (1992), "The Thermal Audit: A Fundamental Stage in the Ergonomics Assessment of Thermal Environment," in *Contemporary Ergonomics 1992*, E. J. Lovesey, Ed., Taylor & Francis, London, pp. 85–90.

Patel, S., Drury, C. G., and Prabhu, P. (1993), "Design and Usability Evaluation of Work Control Documentation," in *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting*, Seattle, WA, pp. 1156–1160.

Patel, S., Drury, C. G., and Lofgren, J. (1994), "Design of Workcards for Aircraft Inspection," *Applied Ergonomics*, Vol. 25, No. 5, pp. 283–293.

Pearl, A., and Drury, C. G. (1995), "Improving the Reliability of Maintenance Checklists," in *Human Factors in Aviation Maintenance, Phase Four, Progress Report*, DOT/FAA/AM-93/xx, National Technical Information Service, Springfield, VA.

Portillo Sosa, J. (1993), "Design of a Computer Programme for the Detection and Treatment of Ergonomic Factors at Workplaces in the Steel Industry," in *Ergonomics Action in the Steel Industry EUR 14832 EN*, O. Berchem-Simon, Ed., Commission of the European Communities, Luxembourg, pp. 421–427.

Public, L., Wilson, J., and Lowe, E. (2010), "The Operational Demand Evaluation Checklist (ODEC) of Workload for Rail Signalling," *Applied Ergonomics*, Vol. 41, pp. 393–402.

Pulat, B. M. (1992), *Fundamentals of Industrial Ergonomics*, Prentice-Hall, Englewood Cliffs, NJ.

Rasmussen, J. (1987), "Reasons, Causes and Human Error," in *New Technology and Human Error*, J. Rasmussen, K. Duncan, and J. Leplat, Eds., Wiley, New York, pp. 293–301.

Reason, J. (1990), *Human Error*, Cambridge University Press, New York.

Régie Nationale des Usine Renault (RNUR) (1976), "Les profils de ostes: Méthode d'analyse des conditions de travail," in *Collection Hommes et Savoir*, Masson, Sirtés, Paris.

Robson, J. W., and Wolstenhulme, S. (2010), "Ergonomically Auditing Your Scan Room," *Ultrasound*, Vol. 18, pp. 41–44.

Robson, L. S., Macdonald, S., Van Erd, D. L., Gray, G. C., and Bigelow, P. (2010), "Something Might be Missing from Occupational Health and Safety Audits: Findings from a Content Validity Analysis of Five Audit Instruments," *Journal of Occupational and Environmental Medicine*, Vol. 52, pp. 536–543.

Rohmert, W., and Landau, K. (1983), *A New Technique for Job Analysis*, Taylor & Francis, London.

Rohmert, W., and Landau, K. (1989), "Introduction to Job Analysis," *A New Technique for Job Analysis*, Taylor & Francis, London, Part 1, pp. 7–22.

Seeber, A., Schmidt, K.-H., Kierswelter, E., and Rutenfranz, J. (1989), "On the Application of AET, TBS and VERA to Discriminate between Work Demands at Repetitive Short Cycle Tasks," in *Recent Developments in Job Analysis*, K. Landau and W. Rohmert, Eds., Taylor & Francis, New York, pp. 25–32.

SHARE (1990), *Inspecting the Workplace*, SHARE Information Booklet, Occupational Health and Safety Authority, Canberra, Australia.

Simpson, G. C. (1994), "Ergonomic Aspects in Improvement of Safe and Efficient Work in Shafts," in *Ergonomics Action in Mining*, Eur 14831, Commission of the European Communities, Luxembourg, pp. 245–256.

Sinclair, M. A. (1984), "Ergonomics of Quality Control," workshop document, International Conference on Occupational Ergonomics, Toronto, Ontario, Canada.

Sinclair, M. A. (1990), "Subjective Assessment," in *Evaluation of Human Work*, J. R. Wilson and E. N. Corlett, Eds., Taylor & Francis, London pp. 58–88.

Smith, R. T. (2001), "An Ergonomics Lineside Audit Method in Automotive Manufacturing," in D. C. Alexander and R. A. Rabourn, Eds., *Applied Ergonomics*, Taylor & Francis, London, pp. 3–13.

Thomas, L. F., and Seaborne, A. E. M. (1961), "The Sociotechnical Context of Industrial Inspection," *Occupational Psychology*, Vol. 35, pp. 36–43.

Treisman, A. (1986), "Features and Objects in Visual Processing," *Scientific American*, Vol. 255, pp. 114B–125.

U.S. DoD (1989), Human Engineering Design Criteria for Military Systems, Equipment and Facilities, MIL-STD-1472D, U. S. Department of Defense, Washington, DC.

Village, J., Trask, C., Luong, N., Chow, Y., Johnson, P., Koehoorn, M., and Teschke, K. (2009), "Development and Evaluation of an Observational Back-Exposure Sampling Tool (Back-EST) for Work-Related Back Injury Risk Factors," *Applied Ergonomics*, Vol. 40, pp. 538–544.

Wagner, R. (1989), "Standard Methods Used in French-Speaking Countries for Workplace Analysis," in *Recent Developments in Job Analysis*, K. Landau and W. Rohmert, Eds., Taylor & Francis, New York, pp. 33–42.

Wang, M. J., and Drury, C. G. (1989), "A Method of Evaluating Inspector's Performance Differences and Job Requirements," *Applied Ergonomics*, Vol. 20, No. 3, pp. 181–190.

Waters, T. R., Putz-Anderson, V., Garg, A., and Fine, L. J. (1993), "Revised NIOSH Equation for the Design and Evaluation of Manual Lifting Tasks," *Rapid Communications, Ergonomics, 1993*, Vol. 36, No. 7, pp. 748–776.

Wenner, C., and Drury, C. G. (1996), "Active and Latent Failures in Aircraft Ground Damage Incidents," in *Proceedings of the Human Factors and Ergonomic Society 39th Annual Meeting*, pp. 796–799.

Wenner, C., and Drury, C. G. (2000), "Analyzing Human Error in Aircraft Ground Damage Incidents," *International Journal of Industrial Ergonomics,* Vol. 26, No. 2, pp. 177–199.

Westwater, M. G., and Johnson, G. I. (1995), "Comparing Heuristic, User-Centred and Checklist-Based Evaluation Approaches," in *Contemporary Ergonomics 1995*, S. A. Robertson, Ed., Taylor & Francis, London, pp. 538–543.

Wickens, C. D. (1992), *Engineering Psychology and Human Performance*, 2nd ed., Harper-Collins, New York.

Wilkins, J. R., Bean, T. L, Mitchell, G. L., Crawford, J. M., and Eicher, L. C. (1997), "Development and Application of a Pen-Based Computer Program for Direct Entry of Agricultural Hazard Data," *Applied Occupational and Environmental Hygiene*, Vol. 12, No. 2, pp. 105–110.

Wolfe, J. M. (1994), "Guided Search II: A Revised Model of Visual Search," *Psychonomic Bulletin and Review 1994*, Vol. 1, No. 2, pp. 202–238.

Wright, P., and Barnard, P. (1975), "Just Fill in This Form: A Review for Designers," *Applied Ergonomics*, Vol. 6, pp. 213–220.

Yoshida, H., and Ogawa, K. (1991), "Workplace Assessment Guideline—Checking Your Workplace," in *Advances in Industrial Ergonomics and Safety III*, W. Karwowski and J. W. Yates, Eds., Taylor &w Francis, London, pp. 23–28.

# CHAPTER 40

# COST/BENEFIT ANALYSIS FOR HUMAN SYSTEMS INVESTMENTS

## Predicting and Trading Off Economic and Noneconomic Impacts of Human Factors and Ergonomics

**William B. Rouse and Kenneth R. Boff**
**Tennenbaum Institute**
**School of Industrial & Systems Engineering**
**Georgia Institute of Technology**
**Atlanta, Georgia**

## 1 INTRODUCTION

The past decade has been a period of very serious scrutiny of the activities of most enterprises. Business processes have been reengineered and enterprises have been downsized or, more popularly, rightsized. Every aspect of an enterprise now must provide value to customers, earn revenues based on this value, and pay its shares of costs. Aspects of an enterprise that do not satisfy these criteria are targeted for elimination.

This philosophy seems quite reasonable and straight-forward. However, implementation of this philosophy becomes rather difficult when the "value" provided is indirect and abstract. When anticipated benefits are not readily measurable in monetary units and only indirectly affect things amenable to monetary measurement, it can be very difficult to assess the worth of investments in such benefits.

There is a wealth of examples of such situations. With any reasonable annual discount rate, the tangible discounted cash flow of benefits from investments in libraries and education, for example, would be so small as to make it very difficult to justify societal investments in these institutions and activities. Of course, we feel quite justified arguing for such investments. Thus, there obviously must be more involved in such an analysis than just discounted cash flow.

This chapter addresses types of human factors and ergonomics investments that have these intangible characteristics in addition to more tangible attributes. One type is research and development (R&D). This type of investment is often made for the purpose of creating long-term value. It will certainly require years and may take decades before returns are fully realized. It is easy to see how R&D can be difficult to justify in terms of impacts on, for instance, this year's sales and profits or current operational readiness.

Another type of investment with these intangible characteristics involves products and services that enhance human effectiveness. This includes selection,

training, system design, job design, organizational development, health and safety, and, in general, the wide range of things done to assure and enhance the effectiveness of people in organizations ranging from businesses to military units. In particular, investments focused on increasing human potential, rather than direct job performance outputs, are much more difficult to justify than those with near-term financial returns (Rouse et al., 1997b).

This chapter also addresses the complex interaction of these two types of investments, namely, R&D investments in human effectiveness. This is done by building on previous efforts—by the authors and many others—addressing the two elements of this interaction. Investing in R&D to enhance human effectiveness presents a confluence of difficulties related to representing and quantifying benefits as well as attributing costs. Nevertheless, there is a widely shared sense that such investments are socially and economically important. It is difficult, however, to justify particular projects on the basis of such perceptions.

A primary difficulty involves the trade-off between the relatively short-term payoffs of direct improvements in job performance and the inherently long-term benefits of R&D efforts aimed at enhancing human effectiveness. Short-term investments usually involve less uncertainty and fewer risks. In contrast, revolutionary high-payoff innovations usually emerge from much earlier R&D investments. Thus, small, certain, near-term returns compete with large, uncertain, long-term, and potentially very substantial returns. The methodology presented in this chapter enables addressing both types of investments.

In general, several issues underlie the difficulties of justifying the aforementioned types of long-term investments. As just noted, a fundamental issue concerns the associated uncertainties. Not only are the magnitudes and timing of returns uncertain—the very nature and characteristics of returns are uncertain. With R&D investments, for instance, the eventual payoffs from investments are almost always greater for unanticipated applications than for the originally envisioned applications (Burke, 1996). Further, organizations that make the original investments are often unable to take advantage of the eventual returns from R&D (Christensen, 1997).

These findings raise concerns about whether or not the outcomes of R&D will actually be employed. Newly emerging technologies and competitors' initiatives may diminish the value of the outcomes. We assert that R&D should be viewed as the means for creating technology "options" that address the contingent needs of an enterprise (Rouse and Boff, 2001, 2003, 2004). The notion of options, which we formalize in the next section, implies that deployment of the outcome of R&D is contingent on the situation at hand when the decision to exercise an option must be made.

Another central issue relates to the preponderance of intangible outcomes for these types of investments. For example, investments in training may enhance leadership skills of managers or commanders. Investments in organizational development can improve the cohesiveness of "mental models" of management teams or command teams and enhance the shared nature of these models. However, it is difficult to capture fully such impacts in terms of tangible, "bottom-line" metrics.

It is important to differentiate between intangible outcomes and those that are tangible but difficult to translate into monetary benefits or costs. For example, an investment might decrease pollution, which is very tangible, but it may be difficult to translate this projected reduction to estimated economic gain. This is a mainstream issue in economics and not unique to cost/benefit analyses.

A further issue concerns cost/benefit analyses across multiple stakeholders. Most companies' stakeholders include customers, shareholders, employees, suppliers, and communities. Government agencies often have quite diverse sociopolitical constituencies who benefit—or stand to lose benefits—in a myriad of ways depending on investment decisions. For example, government-sponsored market research may be part of a regional economic development plan or may be part of a broader political agenda focused on creating jobs. In general, diverse constituencies are quite likely to attempt to influence decisions in a variety of ways. These situations raise many basic questions relative to the importance of benefits and costs for the different stakeholders.

Yet another issue concerns the difference between assessing cost/benefits and predicting cost/benefits. It is certainly valuable to know whether past investments were justified. However, it would be substantially more valuable to be able to predict whether anticipated investments will later provide benefits that justify the initial investments. Of course, limits of our abilities to predict outcomes are not unique to cost/benefit analysis.

The types of investment problems addressed in this chapter are rife with many uncertainties, intangibles, and stakeholders and associated unpredictability. These issues are explored in this chapter in the context of alternative frameworks for performing cost/benefit analyses. This leads to clear conclusions about how best to methodologically handle these types of investments. Application of the resulting methodology is then illustrated in the context of three investment problems involving technologies for aiding, training, and assuring the health and safety of personnel in military systems.

## 2 COST/BENEFIT FRAMEWORKS

There are a variety of frameworks for scrutinizing and justifying investments, including:

- *Cost/benefit analysis*: methods for estimating and evaluating time sequences of costs and benefits associated with alternative courses of action
- *Cost/effectiveness analysis*: methods for estimating and evaluating time sequences of costs and multiattribute benefits to assure that the greatest benefits accrue for given costs
- *Life-cycle costing*: methods for estimating and evaluating costs of acquisition, operation, and retirement of alternative solutions over their total life cycles

- *Affordability analysis*: methods for estimating and evaluating life-cycle costs compared to expected acquisition, operations, and maintenance budgets over the total life cycle of an alternative investments
- *Return-on-investment analysis*: methods for projecting the ratio, expressed as a percentage, of anticipated free cash flow to planned resource investments

This chapter focuses on cost/benefit in a broad sense that includes many aspects of the other approaches. For more traditional treatments of cost/benefit analysis, as well as worked examples, see Layard and Glaister (1994) and Gramlich (1997).

We should, at the outset, contrast approaches to analysis of investments with those for managing investments. R&D funnels, multistage decision processes, and so forth, are intended to assess progress and evaluate the attractiveness of continued investment. Reviews of these constructs can be found in Cooper et al. (1998b) and Rouse and Boff (2001).

Cost/benefit analyses are very straightforward when one considers fixed monetary investments made now to earn a known future stream of monetary returns over some time period. Things get much more complicated, however, when investments occur over time, some of which may be discretionary, and when returns are uncertain.

Further complications arise when one must consider multiple stakeholders' preferences regarding risks and rewards. Additional complexity is added when returns are indirect and intangible rather than purely monetary. These complications and complexity are more common than are situations where the straightforward cost/benefit analyses are applicable. This section discusses alternative frameworks for addressing cost/benefit analyses and compares these alternatives relative to their abilities to address the issues considered in the Introduction.

## 2.1 Traditional Economic Analysis

The time value of money is the central concept in this traditional approach. Resources invested now are worth more than the same amounts gained later. This is due to the costs of the investment capital that must be paid, or foregone, while waiting for subsequent returns on the investment. The time value of money is represented by discounting the cash flows produced by the investment to reflect the interest that would, in effect at least, have to be paid on the capital borrowed to finance the investment.

Equations (1)–(3) summarize the basic calculations of the discounted cash flow model. Given projections of costs, $c_i$, $i = 0, 1, \ldots, N$, and returns, $r_i$, $i = 0, 1, \ldots, N$, the calculations of net present value (NPV), internal rate of return (IRR), or cost/benefit ratio (CBR) are quite straightforward elements of financial management (Brigham and Gapenski, 1988). The only subtlety is choosing a discount rate, DR, to reflect the current

value of future returns decreasing as the time until those returns will be realized increases:

$$NPV = \sum_{i=0}^{N} \frac{r_i - c_i}{(1 + DR)^i} \tag{1}$$

$$IRR = DR \text{ such that } \sum_{i=0}^{N} \frac{r_i - c_i}{(1 + DR)^i} = 0 \tag{2}$$

$$CBR = \frac{\sum_{i=0}^{N} c_i/(1 + DR)^i}{\sum_{i=0}^{N} r_i/(1 + DR)^i} \tag{3}$$

It is quite possible for DR to change with time, possibly reflecting expected increases in interest rates in the future. Equations (1)–(3) must be modified appropriately for time-varying discount rates.

The metrics in equations (1)–(3) are interpreted as follows:

- NPV reflects the amount one should be willing to pay now for benefits received in the future. These future benefits are discounted by the interest paid now to receive these later benefits.
- IRR, in contrast, is the value of DR if NPV is zero. This metric enables comparing alternative investments by forcing the NPV of each investment to zero. Note that this assumes a fixed interest rate and reinvestment of intermediate returns at the internal rate of return.
- CBR simply reflects the discounted cash outflows divided by the discounted cash inflows, or benefits.

These types of metrics have been successfully applied to a variety of human systems integration investments in automotive, chemical, defense, and pharmaceutical industries (Rouse, 2010). The general conclusion is that, in situations where the investing entity is also the entity that receives the returns on the investment, there is often clear economic value of investing in people. In contrast, when the entity investing is not the entity receiving the returns, the tendency is to view such investments as operating costs and minimize them.

## 2.2 Multiattribute Utility Models

Cost/benefit calculations become more complicated when benefits are not readily transformable to economic terms. Benefits such as safety, quality of life, and aesthetic value are very difficult to translate into strictly monetary values. Multiattribute utility models provide a means for dealing with situations involving mixtures of economic and noneconomic attributes.

Let cost attribute $i$ at time $j$ be denoted by $c_{ij}$, $i = 1, 2, \ldots, L$ and $j = 0, 1, \ldots, N$, and benefit attribute $i$ and time $j$ be denoted by $b_{ij}$, $i = 1, 2, \ldots, M$ and $j = 0, 1, \ldots, N$. The values of these costs and benefits are

transformed to common utility scales using $u(\underline{c}_{ij})$ and $u(b_{ij})$. These utility functions serve as inputs to the overall utility calculation at time $j$ as shown in equation (4) (Keeney and Raiffa, 1976):

$$U(\underline{c}_j, \underline{b}_j) = U[u(c_{1j}), u(c_{2j}), \ldots, u(c_{Lj}),$$
$$u(b_{1j}), u(b_{2j}), \ldots, u(b_{Mj})] \qquad (4)$$

which provides the basis for an overall calculation across time using

$$U(\underline{C}, \underline{B}) = U[U(\underline{c}_1, \underline{b}_1), U(\underline{c}_2, \underline{b}_2), \ldots, U(\underline{c}_N, \underline{b}_N)]$$
$$(5)$$

Note that the time value of benefits depicted in equations (1)–(3) is included in equations (4) and (5) by dealing with the time value of costs and returns explicitly and separately from uncertainty.

An alternative approach involves assessing utility functions for discounted costs and benefits, possibly discounted as represented in equations (1)–(3). With this approach streams of costs and benefits are collapsed across time before the values are transformed to utility scales. The validity of this simpler approach depends on the extent to which people's preferences for discounted costs and benefits reflect their true preferences.

The mappings from $c_{ij}$ and $b_{ij}$ to $u(c_{ij})$ and $u(b_{ij})$, respectively, enable dealing with the subjectivity of preferences for noneconomic benefits. In other words, utility theory enables one to quantify and compare things that are often perceived as difficult to objectify. Unfortunately, models based on utility theory do not always reflect the ways in which human decision making actually works.

Subjective expected utility (SEU) theory reflects these human tendencies. Thus, to the extent that one accepts that perceptions are reality, one needs to consider the SEU point of view when one makes expected utility calculations. In fact, one should consider making these calculations using both objective and subjective probabilities to gain an understanding of the sensitivity of the results to perceptual differences.

Once one admits the subjective, one needs to address the issue of whose perceptions are considered. Most decisions involve multiple stakeholders—in other words, people who hold a stake in the outcome of a decision. It is, therefore, common for multiple stakeholders to influence a decision. Consequently, the cost/benefit calculation needs to take into account multiple sets of preferences. The result is a group utility model as shown in equation (6) (Keeney and Raiffa, 1976; Kirkwood, 1979):

$$U = U[U_1(\underline{C}, \underline{B}), U_2(\underline{C}, \underline{B}), \ldots, U_K(\underline{C}, \underline{B})] \qquad (6)$$

where K is the number of stakeholders.

Formulation of such a model requires that two important issues be resolved. First, mappings from attributes to utilities must enable comparisons across stakeholders. In other words, one has to assume that $u = 0.8$, for example, implies the same value gained or lost for all stakeholders, although the mapping from attribute to utility may vary for each stakeholder. Thus, all stakeholders may, for instance, have different needs or desires for safety and, hence, different utility functions. They also may have different time horizons within which they expect benefits—for example, stakeholders of different generations, some perhaps not yet born, have different time horizons within which they expect to receive benefits. However, once the mapping from attributes to utility is performed and utility metrics are determined, one has to assume that these metrics can be compared quantitatively.

The second important issue concerns the relative importance of stakeholders. Equation (6) implies that the overall utility attached to each stakeholder's utility can differ. For example, it is often the case that primary stakeholders' preferences receive more weight than the preferences of secondary stakeholders. The difficulty of this issue is obvious. Who decides? Is there a superstakeholder, for instance? Do the groups of stakeholders, or their representatives, simply vote on who gets how much weight? Such a procedure has its own theoretical problems that cannot be addressed here.

Beyond these two more-theoretical issues, there are substantial practical issues associated with determining the functional forms of $u(c_{ij})$ and $u(b_{ij})$ and the parameters within these functional relationships. This also is true for the higher level forms represented by equations (4)–(6). As the numbers of stakeholders ($K$), cost attributes ($L$), benefit attributes ($M$), and time periods ($N$) increase, these practical assessment problems can be quite daunting.

## 2.3 Option-Pricing Theory

Many investment decisions are not made all at once. Instead, initial investments are made to create the potential for possible future and usually larger investments involving much greater benefits than likely for the initial investments. For example, investments in R&D are often made to create the intellectual property and capabilities that will support or provide the opportunity to subsequently decide whether or not to invest in launching new products or services. These launch decisions are contingent on R&D reducing uncertainties and risks as well as further market information being gained in the interim between the R&D investment decision and possible launch decision. In this way, R&D investments amount to purchasing options to make future investments and earn subsequent returns. These options, of course, may or may not be exercised.

Amram and Kulatilaka (1999), Boer (1998, 1999), and Luehrman (1998) advocate using option-pricing theory to analyze investments involving such contingent downstream decisions. Option-pricing theory focuses on establishing the value of an option to make an investment decision in an uncertain environment at a later date. Developing option-based models begins with

consideration of the effects sought by the investment and the capabilities needed to provide these effects. In the private sector, desired effects are usually profits, perhaps expressed as earnings per share, and needed capabilities are typically competitive market offerings. Options can relate to which technologies are deployed and/or which market segments are targeted. Purchasing options may involve R&D investments, alliances, mergers, acquisitions, and so on. Exercising options involves deciding which technologies will be deployed in which markets and investing accordingly.

In the public sector, effects are usually couched in terms of provision of some public good such as defense. More specific effects might be expressed in terms of measures of surveillance and reconnaissance coverage, for instance. Capabilities would then be defined as alternative means for providing the desired effects. Options in this example might relate to technologies that could enable the capabilities for providing these effects. Attractive options would be those that could provide given effects at lower costs of development, acquisition, and/or operations.

Option-based valuations are economic valuations. Various financial projections are needed as input to option calculations. Projections needed include:

- Investment to "purchase" option, including timing
- Investment to "exercise" option, including timing
- Free cash flow—profits and/or cost savings—resulting from exercise
- Volatility of cash flow, typically expressed as a percentage

The analyses needed to create these projections are often substantial. For situations where cash flows are solely cost savings, it is particularly important to define credible baselines against which savings are estimated. Such baselines should be choices that would actually be made were the options of interest not available.

The models employed for option-based valuations were initially developed for valuation of financial instruments (Black and Scholes, 1973). For example, an option might provide the right to buy shares of stock at a predetermined price some time in the future. Valuation concerns what such an option is worth. This depends, obviously, on the likelihood that the stock price will be greater than the predetermined price associated with the option.

More specifically, the value of the option equals the discounted expected value of the stock at maturity, conditional on the stock price at maturity exceeding the exercise price, minus the discounted exercise price, all times the probability that, at maturity, the stock price is greater than the exercise price (Smithson, 1998). The net option value equals the option value calculated in this manner minus the cost of purchasing the option.

Thus, there are NPVs embedded in the determination of net option values. However, in addition, there is explicit representation of the fact that one will not exercise an option at maturity if the current market share price is less than or equal to the exercise price. As mentioned earlier, sources such as Amram and Kulatilaka (1999), Boer (1998, 1999), Luehrman (1998), Luenberger (1997), and Smithson (1998) provide a wealth of illustrations of how option values are calculated for a range of models.

It is important to note that the options addressed in this chapter are usually termed "real" options in the sense that the investments associated with these options are usually intended to create tangible assets rather than purely financial assets. Application of financially derived models to nonfinancial investments often raises the issue of the extent to which assumptions from financial markets are valid in the domains of nonfinancial investments. This concern is usually addressed with sensitivity analysis.

The assumptions underlying the option-pricing model and the estimates used as input data for the model are usually subject to much uncertainty. This uncertainty should be reflected in option valuations calculated. Therefore, what is needed is a probability distribution of valuations rather than solely a point estimate. This probability distribution can be generated using Monte Carlo simulation to systematically vary model and input variables using assumed distributions of parameter/data variations. The *Technology Investment Advisor* (Rouse et al., 2000) is an example of a tool available to support these types of sensitivity analyses.

These analyses enable consideration of options in terms of both returns and risks. Interesting "What if?" scenarios can be explored. A question that we have frequently encountered when performing these analyses is, "How bad can it get and have this decision still make sense?" This question reflects a desire to thoroughly understand the decision being entertained, not just get better numbers.

The option value resulting from the above formulation is totally premised on the assumption that waiting does not preempt deciding later. In other words, the assumption is that the decision to exercise an option cannot be preempted by somebody else deciding earlier. In typical situations where other actors (e.g., competitors) can affect possible returns, it is common to represent their impact in terms of changes of projected cash flows (Amram and Kulatilaka, 1999). In many cases, competitors acting first will decrease potential cash flows that will decrease the option value. It is often possible to construct alternative competitive scenarios and determine an optimal exercise date.

A central attraction of this model is the explicit recognition that the purpose of an investment now (i.e., purchasing an option) is to assure the option to make a subsequent investment later (i.e., exercise the option). Thus, for example, one invests in creating new technologies for the option of later incorporating these technologies in product and service lines. The significance of the contingent nature of this decision makes an option-pricing model a much better fit than a traditional discounted cash flow model.

Rouse (2010) includes several examples of human systems integration investments that were framed as real options and shown to have substantial economic value.

In several cases, the cash flow estimates needed for the option-pricing models reflected savings of downstream operating costs once systems were deployed.

However, not all long-term investment decisions have substantial contingent elements. For example, one may invest in training and development to later have the option of selecting among talented managers for elevation to executive positions. There are minimal investments associated with exercising such options—almost all of the investment occurs up front. Thus, option-pricing models are not useful for such decisions.

## 2.4 Knowledge Capital Approach

Tangible assets and financial assets usually yield returns that are important elements of a company's overall earnings. It is often the case, however, that earnings far exceed what might be expected from these "hard" assets. For example, companies in the software, biotechnology, and pharmaceutical industries typically have much higher earnings than companies with similar hard assets in the aerospace, appliance, and automobile industries, to name just a few. It can be argued that these higher earnings are due to, for example, greater knowledge capital among software companies. However, since knowledge capital does not appear on financial statements, it is very difficult to identify and, better yet, project knowledge earnings.

Mintz (1998) summarizes a method developed by Baruch Lev for estimating knowledge capital and earnings. This article in *CFO* drew sufficient attention to be discussed in *The Economist* (1999) and reviewed by Strassman (1999). In general, both reviews applauded the progress represented by Mintz's article but also noted the shortcomings of his proposed metrics.

The key, Mintz and Lev argue, is to partition earnings into knowledge earnings and hard asset earnings. Equation (7) accomplishes this by first projecting normalized annual earnings from an average of three past years and estimates for three future years using readily available information. Earnings from tangible and financial assets are calculated from reported asset values using industry averages of 7 and 4.5% for tangible and financial assets, respectively. Knowledge capital is then estimated by dividing knowledge earnings by a knowledge capital discount rate, as shown in equation (8). Based on an analysis of several knowledge-intensive industries, Mintz and Lev use 10.5% for this discount rate:

$$\text{Knowledge earnings} = \text{normalized annual earnings}$$
$$- \text{earnings from tangible assets}$$
$$- \text{earnings from financial assets} \quad (7)$$

$$\text{Knowledge capital} = \text{knowledge earnings}$$
$$/\text{knowledge capital discount rate} \quad (8)$$

Using this approach to calculate knowledge capital, Mintz compares 20 pharmaceutical companies to 27 chemical companies. He determines, for example, ratios of knowledge capital to book value of 2.45 for pharmaceutical companies and 1.42 for chemical companies.

Similarly, the market value–book value ratio is 8.85 for pharmaceutical companies and 3.53 for chemical companies. Considering this correlation between knowledge capital and market value, Strassman (1999) points out that Mintz's estimates do not fully explain the full excess of market values over book values.

The key issue within this overall approach is being able to partition earnings. While earnings from financial assets should be readily identifiable, the distinction between tangible and knowledge assets is problematic. Further, using industry average return rates to attribute earnings to tangible assets does not allow for the significant possibility of tangible assets having little or no earnings potential. Finally, of course, simply attributing all earnings "left over" to knowledge assets amounts to giving knowledge assets credit for everything that cannot be explained by traditional financial methods.

Nevertheless, the knowledge capital construct appears to have potential application to investments involving, for example, R&D or training and development. The purpose of these two types of investments seems to obviously be that of increasing knowledge capital. Further, companies that make investments for this purpose do seem to create more knowledge capital. The key for cost/benefit analyses is being able to project investment returns in terms of knowledge capital and, in turn, project earnings and separate these earnings into knowledge earnings and hard earnings. Further, one needs to be able to do this for specific investment opportunities, not just the company as a whole.

## 2.5 Comparison of Frameworks

Table 1 provides a comparison of the four frameworks just reviewed. It is important to note that this assessment is not really an apples-to-apples comparison. Multiattribute utility theory provides much more of a general framework than the other three approaches that emphasize financial metrics. Nevertheless, these four approaches represent the dominant alternatives.

Traditional economic analyses are clearly the most narrow. However, in situations where they apply, these analyses are powerful and useful. Most of the investment situations addressed in this chapter do not fit these narrow characteristics. For example, if R&D investments in human effectiveness are viewed within a traditional framework, with typical discount rates, no one would ever invest anything in such R&D. But people do make such investments and, thus, there must be more to it than just NPV, IRR, and CBR. [In fact, Cooper et al. (1998b) have found that companies relying solely on financial metrics for R&D investment decisions tend to be the poorest performers of R&D in terms of subsequent market success.]

One view is that R&D reduces uncertainty and buys time before committing very substantial resources to productization, process development, and so on. Option-pricing theory seems to be a natural extension of traditional methods to enable handling these complications. As noted earlier, several authors have advocated this approach for analyses of R&D investments, for example, Boer (1998, 1999) and Lint and Pennings (1998).

**Table 1  Comparison of Cost/Benefit Frameworks**

| Issue/Framework | Traditional Economic Analysis | Multiattribute Utility Models | Option-Pricing Theory | Knowledge Capital Approach |
|---|---|---|---|---|
| Representation of uncertainties | Focuses on expected revenues and costs without consideration of variances. | Probabilistic uncertainties and stakeholder preferences regarding uncertainties are central to models. | Volatility of returns is a central construct within this model. | Focuses on actual and expected earnings without consideration of variances. |
| Intangible vs. tangible outcomes | All outcomes must be converted to monetary units. | Preferences regarding intangible outcomes can be incorporated. | All outcomes must be converted to monetary units. | All outcomes must be converted to monetary units. |
| Multiple stakeholders in costs/benefits | One-dimensional nature of costs and benefits implies one stakeholder. | Formulations for multiple stakeholders are available and limitations are understood. | One-dimensional nature of costs and benefits implies one stakeholder. | One-dimensional nature of costs and benefits implies one stakeholder. |
| Assessing vs. projecting costs/benefits | Depends on abilities to project monetary costs and benefits. | Depends on abilities to project attributes of utility functions. | Depends on abilities to project monetary costs and benefits. | Difficult to project impact of particular investments. |

The knowledge capital approach provides another, less mathematical, way of capturing the impacts of R&D investments in human effectiveness. The difficulty of this approach, which is probably inherent to its origins in accounting and finance, is that it does not address the potential impacts of alternative investments. Instead, it serves to report the overall enterprise score after the game.

Multiattribute utility models can—in principle—address the full range of complications and complexity discussed thus far. Admittedly, the ability to create a rigorous multiattribute utility model depends on the availability of substantial amounts of information regarding stakeholders' preference spaces, probability density functions, and so on. However, in the absence of such information, a much more qualitative approach can be quite useful, as is discussed later in this chapter.

The value of the multiattribute utility approach also depends on being able to compare overall utilities of alternative investments that, in turn, depends on being able to compare different stakeholders' utilities of the alternatives. This ability to transform a complex, multidimensional comparison into a scalar comparison is laden with assumptions. The saving grace of the approach, in this regard, is that it makes these assumptions quite explicit and, hence, open to testing. This does not, of course, guarantee that they will be tested.

Expected utility calculations serve to show how one alternative is better than another, rather than providing absolute scores. Thus, differences of expected utilities among alternatives are usually more interesting than the absolute numbers. In fact, the dialog among stakeholders that is often associated with trying to understand the sources of expected utility differences can provide crucial insights into the true nature of differences among alternatives.

Overall, one must conclude that multiattribute utility models provide the most generalizable approach. This is supported by the fact that multiattribute

models can incorporate metrics such as NPV, option value, and knowledge capital as attributes within the overall model—indeed, the special case of one stakeholder, linear utility functions, and NPV as the sole attribute is equivalent to the traditional financial analysis. Different stakeholders' preferences for these metrics can then be assessed and appropriate weightings determined. Thus, use of multiattribute models does not preclude also taking advantage of the other approaches—the four approaches therefore can be viewed as complementary rather than competing. For these reasons, the multiattribute approach is carried forward in the remainder of this chapter.

## 3  COST/BENEFIT METHODOLOGY

Cost/benefit analysis should always be pursued in the context of particular decisions to be addressed. A valuable construct for facilitating an understanding of the context of an analysis is the value chain from investments to returns. More specifically, it is quite helpful to consider the value chain from investments (or costs), to products, to benefits, to stakeholders, to utility of benefits, to willingness to pay, and finally to returns on investments. This value chain can be depicted as:

| | | |
|---|---|---|
| investments (costs) | to | resulting products over time |
| products over time | to | benefits of products over time |
| benefits over time | to | range of stakeholders in benefits |
| range of stakeholders | to | utility of benefits to each stakeholder |
| utility to stakeholders | to | willingness to pay for utility gained |
| willingness to pay | to | returns to investors |

The process starts with investments that result—or will result—in particular products over time. Products need not be end products; they might be knowledge, skills, or technologies. These products yield benefits, also over time. A variety of people—or stakeholders—have a stake in these benefits. These benefits provide some level of utility to each stakeholder. The utility perceived—or anticipated—by each stakeholder affects their willingness to pay for these benefits. Their willingness to pay affects their "purchase" behaviors that result in returns for investors.

The central methodological question concerns how one can predict the inputs and outputs of each element of this value chain. This question is addressed elsewhere in some detail for R&D management (Rouse et al., 1997a; Rouse and Boff, 1998) and for human effectiveness (Rouse and Boff, 2003). Briefly, a variety of models have been developed for addressing this need for prediction. These models are very interesting and offer much potential. However, they suffer from a central shortcoming. With few exceptions, there is an almost overwhelming lack of data for estimating model parameters as well as a frequent lack of adequate input data. Use of data from baselines can help, but the validity of these baselines depends on new systems and products being very much like their predecessors. Overall, the paucity of data dictates development of a more qualitative methodology whose usefulness is not totally determined by availability of hard data. The remainder of this section outlines such a methodology.

As indicated in the earlier comparison of four frameworks for addressing cost/benefit analysis, the most broadly applicable of these alternatives are multiattribute utility models. The remainder of this section describes a seven-step methodology that includes the following steps:

- Identify stakeholders in alternative investments.
- Define benefits and costs of alternatives in terms of attributes.
- Determine utility functions for attributes (benefits and costs).
- Decide how utility functions should be combined across stakeholders.
- Assess parameters within utility models.
- Forecast levels of attributes (benefits and costs).
- Calculate expected utility of alternative investments.

It is important to note that this methodology is, by no means, novel and builds upon works by many others related to multiattribute analysis (e.g., Keeney and Raiffa, 1976; Sage, 1977; Hammond et al., 1998; Matheson and Matheson, 1998; Sage and Armstrong, 2000).

### 3.1 Step 1: Identify Stakeholders

The first step involves identifying the stakeholders who are of concern relative to the investments being entertained. Usually this includes all of the people in the value chain summarized earlier. This might include, for example, those who will provide the resources that will enable a solution, those who will create the solution, those who will implement the solution, and those who will benefit from the solution.

### 3.2 Step 2: Define Benefit and Cost Attributes

The next step involves defining the benefits and costs involved from the perspective of each stakeholder. These benefits and costs define the attributes of interest to the stakeholders. Usually, a hierarchy of benefits and costs emerges, with more abstract concepts at the top, for example, viability, acceptability, and validity (Rouse, 1991), and concrete measurable attributes at the bottom.

### 3.3 Step 3: Determine Stakeholders' Utility Functions

The value that stakeholders attach to these attributes are defined by stakeholders' utility functions. The utility functions enable mapping disparate benefits and costs to a common scale. A variety of techniques are available for assessing utility functions (Keeney and Raiffa, 1976).

### 3.4 Step 4: Determine Utility Functions Across Stakeholders

Next, one determines how utility functions should be combined across stakeholders. At the very least, this involves assigning relative weights to different stakeholders' utilities. Other considerations such as desires for parity can make the ways in which utilities are combined more complicated. For example, equation (5) may require interaction terms to assure all stakeholders gain some utility.

### 3.5 Step 5: Assess Parameters of Utility Functions

The next step focuses on assessing parameters within the utility models. For example, utility functions that include diminishing or accelerating increments of utility for each increment of benefit or cost involve rate parameters that must be estimated. As another instance, estimates of the weights for multistakeholder utility functions have to be estimated. Fortunately, there are a variety of standard methods for making such estimates.

### 3.6 Step 6: Forecast Levels of Attributes

With the cost/benefit model fully defined, one next must forecast levels of attributes or, in other words, benefits and costs. Thus, for each alternative investment, one must forecast the stream of benefits and costs that will result if this investment is made. Quite often, these forecasts involve probability density functions rather than point forecasts. Utility theory models can easily incorporate the impact of such uncertainties on stakeholders' risk aversions. On the other hand, information on probability density functions may not be available or may be prohibitively expensive. In these situations, beliefs of stakeholders and subject matter experts can be employed, perhaps coupled with sensitivity analysis (see step 7) to determine where additional data collection may be warranted.

### 3.7 Step 7: Calculate Expected Utilities

The final step involves calculating the expected utility of each alternative investment. These calculations are performed using specific forms of equations (4)–(6). This step also involves using sensitivity analysis to assess, for example, the extent to which the rank ordering of alternatives, by overall utility, changes as parameters and attribute levels of the model are varied.

### 3.8 Use of the Methodology

Some elements of the cost/benefit methodology just outlined are more difficult than others. The overall calculations are quite straightforward. The validity of the resulting numbers depends, of course, on stakeholders and attributes having been identified appropriately. It further depends on the quality of the inputs to the calculations.

These inputs include estimates of model parameters and forecasts of attribute levels. As indicated earlier, the quality of these estimates is often compromised by lack of available data. Perhaps the most difficult data collection problems relate to situations where the impacts of investments are both uncertain and very much delayed. In such situations, it is not clear which data should be collected and when they should be collected.

A recurring question concerns the importance that should be assigned to differences in expected utility results. If alternative $A$ yields $U(A) = 0.648$ and alternative $B$ yields $U(B) = 0.553$, is $A$ really that much better than $B$? In fact, are either utilities sufficiently great to justify an investment?

These questions are best addressed by considering past investments. For successful past investments, what would their expected utilities have been at the time of the investment decisions? Similarly, for unsuccessful past investments, what were their expected utilities at the time? Such comparisons often yield substantial insights.

Of course, the issue is not always $A$ versus $B$. Quite often the primary question concerns which alternatives belong in the portfolio of investments and which do not. Portfolio management is a fairly well-developed aspect of new product development (e.g., Cooper et al., 1998a; Gill et al., 1988). Well-known and recent books on R&D/technology strategy pay significant attention to portfolio selection and management (e.g., Roussel et al., 1991; Matheson and Matheson, 1998; Boer, 1999; Allen, 2000). In fact, the conceptual underpinnings of option-pricing theory are based on notions of market portfolios (Amram and Kulatilaka, 1999).

Most portfolio management methods rely on some scoring or ranking mechanism to decide which investments will be included in the portfolio. Expected utility is a quite reasonable approach to creating such scores or ranks. This is particularly useful if sensitivity analysis has been used to interactively explore the basis and validity of differences among alternatives.

A more sophisticated view of portfolio management considers interactions among alternatives in the sense that synergies between two alternatives may make both of them more attractive (Boer, 1999; Allen, 2000). Also, correlated risks between two alternatives may make both of them less attractive. A good portfolio has an appropriate balance of synergies and risks.

In principle, at least, the notions of portfolio synergy and risk can be handled within multiattribute utility models. This can be addressed by adding attributes that are characteristics of multiple rather than individual alternatives. In fact, such additional attributes might be used to characterize the whole portfolio. An important limitation of this approach is the likely significant increase in the complexity of the overall problem formulation. Indeed, this is an issue in general when multi-attribute utility models are elaborated to better represent problem complexities.

Beyond these technical issues, it is useful to consider how this cost/benefit methodology should affect decision making. To a very great extent, the purpose of this methodology is to get the right people to have the right types of discussions and debates on the right issues at the right time. If this happens, the value of people's insights from exploring the multiattribute model usually far outweighs the importance of any particular numbers.

The practical implications of this conclusion are quite simple. Very often, decision making happens within working groups who view computer-generated, large-screen displays of the investment problem formulation and results as they emerge. Such groups perform sensitivity analyses to determine the critical assumptions or attribute values that are causing some alternatives to be more highly rated or ranked than others. They use "What if ..?" analyses to explore new alternatives, especially hybrid alternatives.

This approach to investment decision making helps to substantially decrease the impact of limited data being available. Groups quickly determine which elements of the myriad of unknowns really matter—where more data are needed and where more data, regardless of results, would not affect decisions. A robust problem formulation that can be manipulated, redesigned, and tested for sanity provides a good way for decision-making groups to reach defensible conclusions with some level of confidence and comfort.

## 4  THREE EXAMPLES

Human effectiveness concerns enhancing people's direct performance (aiding), improving their potential to perform (training), and assuring their availability to perform (health and safety). These are central issues in human systems integration. Investments in human effectiveness also have the potential of increasing returns on other investments by, for example, enabling people to take full advantage of new technologies.

Three examples of aiding, training, and health and safety investments are discussed in this section: VCATS (aiding), DMT (training), and PTOX (health and safety). These examples focus on enhancing human effectiveness and human systems integration in military systems, particularly Air Force systems. The applicability of these technologies, and the relevance of the following analysis of the impacts of these technologies, to other military services and to nonmilitary problems should also be readily apparent.

## 4.1 Visually Coupled Targeting and Acquisition System

The Visually Coupled Targeting and Acquisition System (VCATS) provides aid to military aircraft pilots. VCATS includes a helmet-mounted tracker and display (HMT/D), associated signal-processing sensor/transducer hardware, interchangeable panoramic night vision goggle with head-up display (PNVG-HUD), and extensive upgrades to the aircraft's operational flight program software (Rastikis, 1998). VCATS enables the pilot to cue and be cued by on-board and off-board systems, sensors, and weapons as well as be spatially and temporally coupled with the control processes implemented with the HMT/D and PNVG-HUD. The system is particularly effective in helping pilots to cue weapons and sensors to targets, maintain "ownship" formation situation awareness, and avoid threats via provision of a real-time, three-dimensional portrayal of the pilots' tactical and global battlefield status. In general, VCATS enables pilots to acquire targets and threats faster. This results in improvements in terms of (1) how far, (2) how quickly, and (3) how long—for both initial contacts and countermeasures.

To a great extent, the case for advanced development has already been made for VCATS and current support is substantial. However, the transition from advanced development to production involves assuring that the options created by VCATS and validated by combat pilots are exercised. The case has also been argued for ongoing investments in basic research and exploratory development to assure that VCATS has future technology options, particularly for migration to multirole fighter aircraft. The maturity of the program should help in making this case in terms of benefits already demonstrated. However, in the current budget climate, there is also substantial risk that VCATS research may be viewed as essentially "done." This raises the potential for negative decisions regarding further investments.

## 4.2 Distributed Mission Training

Distributed mission training (DMT) involves aircraft, virtual simulators, and constructive models that, collectively, provide opportunities for military pilots to gain experiences deemed important to their performance proficiency relative to anticipated mission requirements (Andrews, 2000). The desired training experiences are determined from competencies identified as needed to fulfill mission requirements. These competency requirements are translated to training requirements stated in terms of types and durations of experiences deemed sufficient to gain competency.

The case to be made for DMT involves investments to address research issues and technology upgrades of near-term capabilities. The primary options-oriented argument is that investments in R&D in DMT will create contingent possibilities for cost savings in training due to reduced use of actual aircraft. More specifically, DMT options, if exercised, will provide cash flows of savings that justify the investments needed to field this family of technologies.

A much more subtle options-oriented argument concerns the training experiences provided by DMT that

could not otherwise be obtained. Clearly, the opportunity to have relevant training experiences must be better than not having these experiences. The option, therefore, relates to proficiency vs. possible lack of proficiency.

As straightforward as this may seem, it quickly encounters the difficulty of projecting mission impacts—and the value of these impacts—of not having proficient personnel. One possible approach to quantifying these benefits is to project the costs of using real aircraft to gain the desired proficiencies. While these costs are likely to be prohibitive, and thus never would be seriously considered, they nevertheless characterize the benefits of DMT.

## 4.3 Predictive Toxicology

Predictive toxicology (PTOX) is concerned with projecting the impacts on humans from exposure to operational chemicals (individual and mixtures). The impact can be characterized in terms of the possibility of performance decrement and consequent loss of force effectiveness, possible military and civilian casualties, and potential long-term health impacts. Also of concern are the impacts of countermeasures relative to sustaining immediate performance and minimizing long-term health impacts [Office of Science and Technology Policy (OSTP), 1998].

The case to be made involves investment in basic research and exploratory development programs, with longer term investment in an advanced development program to create deployable predictive toxicology capabilities. The requisite R&D involves developing and evaluating models for predicting performance and health impacts of operational chemicals. Advanced development will focus on field sensing and prediction, termed deployment toxicology. The nature of the necessary models is strongly affected by the real-time requirements imposed by deployment.

## 4.4 Applying the Methodology

The remainder of this section primarily addresses steps 1–4 of the cost/benefit methodology in the context of these three examples related to human effectiveness aspects of human systems integration. These steps constitute the "framing" steps of the methodology, rather than the "calculation" steps. Appropriate framing of cost/benefit analyses is critical to subsequent calculations being meaningful and useful.

### 4.4.1 Step 1: Identify Stakeholders

This step involves identifying people, usually types of people, and organizations that have a stake in costs and benefits. All three of the examples involve three classes of stakeholders: warfighters, developers, and the public. A key issue concerns the relative importance of these three types of stakeholders. Some would argue that warfighter preferences dominate decisions. Others recognize the strong role that developers, and their constituencies, play in procurement decisions. Yet another argument is that the dominating factor is value to the public, with the other stakeholders being secondary in importance.

Warfighters as stakeholders include military personnel in general, especially for PTOX. Warfighters of particular importance include aircraft pilots, personnel who support flight operations, and military commanders. Developers as stakeholders include companies and their constituencies, for example, stockholders, employees, and communities. Several agents, including Congress, the executive functions within the military services, and the military procurement establishment, represent the public's interests. Pilots and other military personnel are users of the technologies of interest, developers are the providers, and the public's agents are the customers for these technologies. There are obvious trade-offs across the interests of users, providers, and customers.

### 4.4.2 Step 2: Define Benefit and Cost Attributes

Benefits and costs tend to fall in general classes. Example benefits for military organizations and contractors include:

- Enhanced impact → increased lethality, survivability, and availability
- Enhanced operability → decreased response time and increased throughput
- Enhanced design → new techniques and larger pool of experienced people
- Increased opportunities → new tactics and countermeasures

Example cost attributes applicable to military procurement include:

- Investment costs → capital investments and R&D costs
- Recurring costs → operating and G&A costs
- Time costs → time from development to fielding to competent use
- Opportunity costs → other costs/benefits foregone

These general classes of benefits and costs can be translated into specific benefit and cost attributes for the three classes of stakeholders in VCATS, DMT, and PTOX. Benefits for warfighters (users) include enhanced performance (e.g., response time), confidence in performance, and health and safety in varying combinations for the three examples. Costs for these stakeholders include learning time and changing their ways of doing things to assure compatibility between new and legacy technologies.

Benefits for companies and their constituencies (providers) include R&D funds received, subsequent intellectual property created, and competitive advantages that result. Also important are jobs and economic impacts in the community. Direct costs include bid and proposal costs as well as opportunity costs. Less direct costs include, for instance, economic development resources and incentives provided to the companies by their communities.

The primary benefit sought by the public's agents (customer) is mission performance/dollar. It can easily be argued for all three examples that mission performance is increased. Unfortunately, it is difficult to attach a value to this increase. For example, what is the value of being able to generate 5% more sorties per time period? The answer depends on whether more sorties are needed.

Few would argue with the importance of successfully meeting mission requirements. However, if the types of innovations represented by these examples enable exceeding mission requirements, what are such increases worth? This is a politically sensitive question. If better performance is of substantive value, why wasn't this level of performance specified in the original requirements?

A good way to avoid this difficulty is to take mission requirements as a given and determine how much money could be saved in meeting these requirements by adopting the technologies in question. For example, could requirements be met with fewer aircraft, pilots, and support personnel? As shown in Table 2, the cost savings due to these decreases can be viewed as benefits of the technologies. It also might be possible for VCATS, DMT, or PTOX to enable meeting mission requirements with less-capable systems, rather than just fewer systems. This possibility provides substantial opportunities for increased benefits due to these technologies.

### Table 2  Public Benefits/Costs for Three Examples

|  | VCATS | DMT | PTOX |
|---|---|---|---|
| Benefits | Fewer aircraft and associated personnel to meet mission requirements due to better performance and fewer aircraft losses | Fewer aircraft and associated personnel to meet mission requirements due to better performance, fewer aircraft losses, and fewer aircraft for training | Fewer personnel to meet mission requirements and decreased medical costs due to fewer people affected by toxic materials, fewer people lost to toxic effects, fewer people to care for people affected, and decreased downstream medical costs |
| Costs | Initial investment (option price) for proposed R&D costs and later contingent investment (exercise price) for subsequent fielding of technology | Initial investment (option price) for proposed R&D costs and later contingent investment (exercise price) for subsequent fielding of technology | Initial investment (option price) for proposed R&D costs and later contingent investment (exercise price) for subsequent fielding of technology |

Note that this philosophy amounts to trying to provide a given level of defense for the least investment. Another approach might be to attempt to provide the most defense per investment dollar. However, this immediately begs the question of how much defense is enough. Unlike the business world where value is defined by the marketplace and, hence, can provide a basis for optimization [see Nevins and Winner (1999) for a good example], there is no widely agreed-upon approach to measuring military value and optimizing accordingly.

The rationale for the benefits indicated in Table 2 for each of the three examples include:

- VCATS enables pilots to compete with threats, increase the number of wins versus losses, and counter threats (e.g., missiles) in ways that they could not do otherwise. Consequently, it must be possible to meet *fixed* mission requirements with fewer aircraft and associated infrastructure. These benefits can be translated into financial returns in terms of cost avoidance.

- DMT provides opportunities to practice behaviors that would not otherwise be practiced, for the most part due to the costs of practice. This decreases the probability of not performing acceptably given inadequate training. DMT also provides training experiences that would not otherwise be possible. For example, in the DMT environment, pilot "kills" actually disappear. In contrast, field exercises often "reuse" kills because of the costs of getting adversaries into the exercise in the first place.

- PTOX enables larger proportions of deployed forces to be fully functional, be less dependent on medical surveillance or medication, and have earlier intervention, before the onset of problems. In principle, this should enable reducing the size of the deployed force, which is critical for increasingly likely expeditionary military missions (Fuchs et al., 1997).

- PTOX also provides cost avoidance due to downstream health impacts. The ability to predict the "body burden" of toxicity during deployment should enable removing personnel from risk once the burden is approaching predetermined limits. These capabilities are likely to also be very important for nonmilitary operations such as disaster clean-up.

It is *not* essential that the savings indicated in Table 2 actually occur. For example, it may be that the number of aircraft is not decreased, perhaps due to factors far beyond the scope of these analyses. However, one can nevertheless attribute to these technologies the benefits of having provided opportunities to meet mission requirements in less costly manners. Technologies that provide such opportunities are valuable; the extent of this value is the extent of the opportunities for savings.

This argument puts all three examples on common ground. The benefits of all alternative technologies can be expressed as reduced costs to meet requirements. From an options-pricing perspective, these savings can be viewed as free cash flow returned on investments in these technologies. The "option price" is the R&D costs. The "exercise price" is the subsequent costs of fielding the technologies. Thus, assuming costs savings can be projected (albeit with substantial volatility), the option values of investing in these technologies can be calculated.

### 4.4.3 Step 3: Determine Stakeholders' Utility Functions

Different stakeholders' preferences over the benefit and cost attributes will vary substantially with specific situations. However, there is a small family of functional relationships that captures most, if not all, expressed preferences (Keeney and Raiffa, 1976). Thus, while context-specific tailoring is needed, it can be performed within a prescribed (and preprogrammed) set of functions, both within and across stakeholders. Similarly, alternative parameter choices can be prescribed in terms of choices of weightings.

An important aspect of cost/benefit analyses, as advocated in this chapter, is the likely nonlinear nature of utility functions. In particular, diminishing returns and aspiration levels tend to be central to stakeholders' "preference spaces." In other words, while linear functions imply that incremental increases (or decreases) of attributes always yield the same incremental changes in utility, nonlinear functions lead to shifting preferences as attributes increase (or decrease). Figure 1 portrays a range of example utility functions.

To illustrate how these types of relationships can be employed to represent the preferences of users, providers, and customers, the general forms of each type of stakeholder's utility function are shown in equations (9)–(11):

$$U_{\text{user}} = U[u(\text{performance}), u(\text{confidence}),$$
$$u(\text{cost of change})] \tag{9}$$

$$U_{\text{provider}} = U[u(\text{resources}), u(\text{advantage}),$$
$$u(\text{cost of pursuit})] \tag{10}$$

$$U_{\text{customer}} = u(\text{option value}) \tag{11}$$

where, as noted earlier, users are primarily concerned about impacts of investments on their performance, their confidence in their performance, and the costs of changing their ways of performing; providers are concerned with the investment resources supplied to develop the technologies in question, the competitive advantages created by the intellectual property created, and the costs of pursuing the investment opportunities; and, finally, customers are focused on the financial attractiveness of the investments as reflected in the option values of the alternatives which are based on projected cash flows (i.e., costs savings), volatility of cash flows, magnitudes of investments required, and time periods until returns are realized.
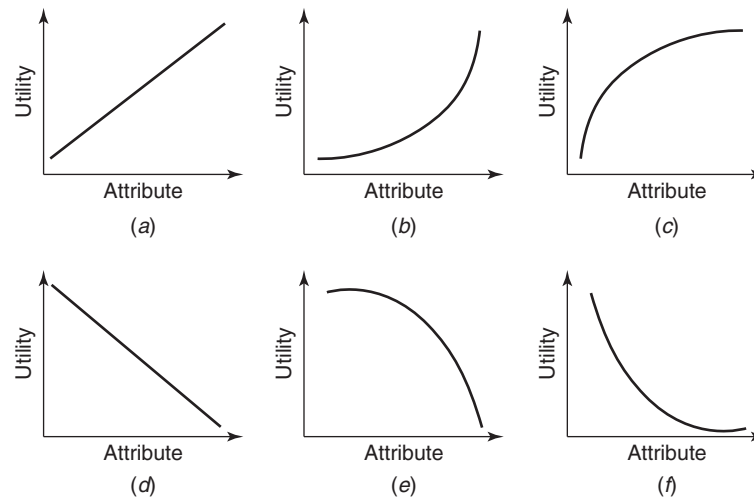
**Figure 1**  Example utility functions: (*a*, *d*) inear relationships; (*b*, *e*) accelerating relationships; (*c*, *f*) diminishing relationships.

Considering the elements of equations (9)–(11), the appropriate functional forms from Figure 1 are likely to be as follows:

- *u*(performance) is an accelerating returns function (Figure 1*b*):
  - VCATS is least concave since relatively modest performance improvements are of substantial utility
  - DMT is moderately concave since training on otherwise untrained tasks must produce substantial improvements to yield high utility
  - PTOX is most concave since major decreases in performance risk are needed to assure high utility increases of personnel availability
- *u*(confidence) is a linear function (Figure 1*a*) since greater confidence is always better, but there are unlikely to be significant thresholds
- *u*(cost of change) is an accelerating decline function (Figure 1*f*) since low to moderate costs are easily sustained while larger costs present difficulties
- *u*(resources) is an accelerating returns function (Figure 1*b*) since moderate to large resources are needed to make opportunities attractive
- *u*(advantage) is a linear function (Figure 1*a*) since greater advantage is always better, but there are unlikely to be significant thresholds
- *u*(cost of pursuit) is an accelerating decline (Figure 1*f*) since low to moderate costs are easily sustained while larger costs present difficulties
- *u*(option value) is a linear function (Figure 1*a*) since customers will inherently gain the expected value across a large number of investments

It is important to note the importance of this last assumption. If customers', that is, the public's, utility function were not linear, it would be necessary to

entertain assessing the specific form of their function. Unlike users and providers, the public is not so easily identified and interviewed.

With the identification of the stakeholders (step 1) and framing of the cost/benefit attributes (step 2), the process of determining the form of stakeholders' utility functions (step 3) can draw upon considerable standard "machinery" of decision analysis. The specific versions of the functional forms discussed above are likely to vary with VCATS, DMT, and PTOX. However, the overall formulation chosen is quite general.

### 4.4.4  Step 4: Determine Utility Functions Across Stakeholders

Another important aspect of the utility functions is their typical lack of alignment across stakeholders. Specifically, either different stakeholders care about different things or possibly they care about the same things in different ways. For example, customers may be very price sensitive while users, who seldom pay prices themselves, are usually much more concerned with impacts on their job performance.

For the types of investment problems considered in this chapter, preferences typically differ across time horizons and across people with vested interests in different investment opportunities. Thus far in the formulation of the three examples, the stakeholders do not have attributes in common. However, they are nevertheless likely to have competing preferences since, for example, the alternative providing the greatest performance impact may not have the largest option value.

Differing preferences across stakeholders are often driving forces in pursuing cost/benefit analyses. These differing preferences can be aggregated, and traded off, by formulating a composite utility function such as

$$U = U\left[U_{\text{user}}, U_{\text{provider}}, U_{\text{customer}}\right] \qquad (12)$$

Often equation (12) will be linear in form with weights assigned to component utility functions to reflect the relative importance of stakeholders. Slightly more complicated are multi-linear forms which include products of component functions, for example, $U_{user} \times U_{customer}$. Multilinear formulations tend to assure that all stakeholders gain nonzero utility because, otherwise, zero in either term in a product yields zero overall.

Considering trade-offs across stakeholders, it is important to note that the formulation of the analysis can often be usefully expanded to include a broader set of stakeholders. These additional stakeholders may include other entities that will benefit by advances of the technologies in question, although they may have little or no stake in the immediate application for the technology. It is also quite possible that stakeholders such as "the public" have multiple interests, for example, military effectiveness and public safety from toxic risks.

Broadening the analysis in this way is likely to have differing impacts on the assessment for the three examples due to the natures of the technologies and issues being pursued. The three examples differ in this regard in the following ways:

- VCATS addresses a rather esoteric set of issues from the public's perspective.
- DMT addresses an issue with broad general support from the public, but narrower specific constituencies.
- PTOX addresses strong cross-cutting health and safety issues of substantial concern to the public.

These differences suggest that PTOX would gain a larger $\Delta U$ than DMT, and DMT would in turn gain a larger $\Delta U$ than VCATS, by broadening the number of stakeholders and issues. Quite simply, the "spin-off" benefits of PTOX are likely to be perceived as much greater by a larger number of stakeholders.

However, if the formulation is further broadened to consider the likelihood that the desired technologies will emerge elsewhere if investments are not made in these efforts, the $\Delta U$ impacts are likely be the opposite. PTOX research and development are being pursued by several agencies. DMT has broad applicability for both military and nonmilitary applications and consequently is being pursued by other parties. VCATS, in contrast, is highly specialized and is unlikely to emerge from other sources.

These two possibilities for broadening the formulation, in terms of stakeholders and issues, clearly illustrate the substantial impact of the way in which cost/benefit assessments are framed. If the framing is too focused, important spin-off benefits may not be included. On the other hand, framing the analysis too broadly may raise issues that are difficult to quantify, even roughly, and include stakeholders whose preferences are difficult to assess. Of course, many modeling efforts face such difficulties (Sage and Rouse, 1999).

### 4.4.5  Steps 5–7: Calculation of Overall Cost/Benefit

The remaining steps of the cost/benefit methodology involve assessing parameters of utility functions, forecasting levels of attributes, and calculating expected utilities. Performing these steps obviously depends on having data on stakeholders' preferences and projected/targeted attribute levels. Discussion of such data is well beyond the scope of this chapter—and, in light of the nature of the examples, it would be difficult to publish the requisite data.

The needed data can, in many instances, be quite difficult to compile. It can be particularly difficult to relate returns on human effectiveness investments to organizational impacts. Relationships between human and organizational performance are needed. These relationships should answer the following types of questions:

- How do improvements in human performance (e.g., via aiding) translate to increased organizational impacts? Specifically, how does a 2-second improvement in pilot response time due to VCATS affect mission performance?
- How do improvements in human potential to perform (e.g., via training) translate to actual performance and consequent increased organizational impacts? Specifically, how does increased practice via DMT impact subsequent performance and, in turn, translate to improved mission performance?
- How do improvements in human availability to perform (e.g., via health and safety) translate to actual performance and consequent increased organizational impacts? Specifically, how does prevention of toxic exposure, due to PTOX, affect immediate unit performance and thereby affect mission performance?

These can be difficult questions. However, they are not inherently cost/benefit questions. Instead, they are fundamental system design questions (Sage and Rouse, 1999). If answers are possible, then cost/benefit analyses are more straightforward.

For the VCATS, DMT, and PTOX examples, it may be possible to translate human performance improvements to organizational impacts via mission models. Such models are typically used to determine, for example, the "logistics footprint" needed to support a targeted sortie generation rate or, as another illustration, the combat wins and losses likely with competing defensive measures and countermeasures. Such models can be applied, perhaps with extensions, to project the impacts of faster responses due to VCATS, improved task performance due to DMT, and increased personnel availability due to PTOX.

It is important to note, however, that even if such projections are not available, the multiattribute methodology presented here can still be employed. However, the validity of cost/benefit assessments and predictions will then depend upon subjective perceptions of attribute levels and the relative importance of attributes. Any limitations of this more subjective approach reflect underlying limitations of knowledge rather than inherent limitations of the methodology.

Once $U[U_{\text{user}}, U_{\text{provider}}, U_{\text{customer}}]$ is fully specified, both functionally and in terms of parameters of these functions, one is in a position to project attribute levels (e.g., option values), calculate the expected utility of the alternative investments (e.g., VCATS, DMT, and PTOX), and perform sensitivity analyses. This provides the basis for making investment decisions. There are several ways that these cost/benefit assessments can be used to inform decision making.

The most common way of using expected utility cost/benefit assessments is to rank order alternative investments in terms of decreasing $U[U_{\text{user}}, U_{\text{provider}}, U_{\text{customer}}]$ and then allocate investment resources from highest ranked to lowest ranked until resources are exhausted. This approach allows the possibility of alternatives with mediocre $U_{\text{customer}}$ making the cut by having substantial $U_{\text{user}}$ and $U_{\text{provider}}$. To avoid this possibility, one can rank order by $U[U_{\text{user}}, U_{\text{provider}}, U_{\text{customer}}]$ all alternatives with $U_{\text{customer}} > U_{\text{co}}$, which implies a minimum acceptable option value.

If resources are relatively unconstrained, one can, invest in all alternatives for which $U_{\text{user}} > U_{\text{uo}}$, $U_{\text{provider}} > U_{\text{po}}$, and $U_{\text{customer}} > U_{\text{co}}$. This reflects situations where all stakeholders prefer investment to no investment. Of course, one can also rank order these alternatives by $U[U_{\text{user}}, U_{\text{provider}}, U_{\text{customer}}]$ to determine priorities for investment. However, if resources are truly unconstrained, this rank ordering will not change the resulting investment decisions.

### 4.5 Summary

The three examples discussed in this section have portrayed a cross section of human effectiveness investments to enhance human systems integration, ranging from aiding to training to health and safety investments. The discussion has shown how this range of investment alternatives can be fully addressed with an overarching multiattribute utility, multistakeholder cost/benefit formulation. The stakeholder classes of user, provider, and customer are broadly applicable. The classes of attributes discussed also have broad applicability.

These examples have also served to illustrate the merits of a hybrid approach. In particular, option value theory has been used to define the issue of primary interest to customers, assuring that investments make financial sense, and this issue has then been incorporated into the overall multiattribute formulation. This enabled including in the formulation a substantial degree of objective rigor as well as important subjective attributes and perceptions. As a result, rigor is not sacrificed but instead is balanced with broader, less quantifiable considerations.

It is useful to note that the knowledge capital construct was not employed in the formulation for these three examples, despite the intuitive appeal of the notion that investments in human effectiveness increase knowledge capital (Davenport, 1999). While the formulation reported here could have included increases of knowledge capital as possible benefits, there is no basis for predicting such impacts. Subjective estimates could, of course, be employed. However, this construct is not defined with sufficient crispness to expect reliable estimates from subject matter experts.

The discussion of these examples of human effectiveness investments has served to illustrate the value of an overall cost/benefit formulation. The generality of this formulation allows it to be applied to analyses of a wide variety of human system integration investment decisions. The types of information needed to support such analyses are defined by this formulation. While the availability of information remains a potential difficulty, this formulation nevertheless substantially ameliorates the typical problems of comparing ad hoc analyses of competing investments. Also of great importance, this formulation enables cross-stakeholder comparisons and trade-offs that, for the lack of a suitable methodology, are usually ignored or resolved in ad hoc manners.

## 5  CONCLUSIONS

It is difficult to make the case for long-term investments that will provide highly uncertain and intangible returns. This chapter has reviewed alternative ways to characterize such investments and presented an overall methodology that incorporates many of the advantages of these alternatives. This methodology has been illustrated in the context of R&D investments in human effectiveness aspects of human systems integration.

Central to the cost/benefit analysis methodology presented is a multiattribute, multistakeholder formulation. This formulation includes nonlinear preference spaces that are not necessarily aligned across stakeholders. The nonlinearities and lack of alignment provide ample opportunities for interesting trade-offs.

It is important to stress the applicability of this methodology to nearer term human effectiveness investments, which may or may not involve R&D. While the time frame will certainly affect choices of attributes—for instance, option values may not be meaningful for near-term investments—the overall cost/benefit methodology remains unchanged. This chapter focused on long-term R&D investments because such analyses are the most difficult to frame and perform.

It is also useful to indicate that cost/benefit analysis, as broadly conceptualized in this chapter, can be a central element in assessment activities related to life-cycle costing (e.g., affordability) and program/contract management [e.g., earned value management (EVM, 2000)]. For the former, attributes reflecting life-cycle costs can easily be incorporated. For the latter, costs and benefits can be tracked and compared to original projections. This does, of course, require that benefits be attributable to ongoing processes and not just outcomes.

This cost/benefit methodology, when coupled with appropriate methods and tools for predicting attribute levels (Sage and Rouse, 1999), can enable cost/benefit predictions and, thereby, support investment decision making. Using attributes such as option values and potentially knowledge capital can make it possible to translate the intuitive appeal of R&D and human

effectiveness investments into more tangible measures of value.

Note also that the methodology includes many of the elements necessary to developing a business case for human effectiveness investments. Markets (stakeholders), revenues (benefits), and costs are central issues in business case development and in this methodology. However, this methodology also supports valuation of investments with broader constituencies (e.g., the public) and ranges of issues (e.g., jobs created) than typically considered in business cases.

Finally, we have also found that use of the methodology presented here provides indirect advantages in terms of causing decision-making groups to clarify and challenge underlying assumptions. This helps decision makers avoid being trapped by common delusions which could mislead them relative to likely cost/benefits (Rouse, 1998).

## REFERENCES

Allen, M. S. (2000), *Business Portfolio Management: Valuation, Risk Assessment, and EVA Strategies*, Wiley, New York.

Amram, M., and Kulatilaka, N. (1999), *Real Options: Managing Strategic Investment in an Uncertain World*, Harvard Business School Press, Boston.

Andrews, D. H. (2000), "Distributed Mission Training," in *International Encyclopedia of Ergonomics and Human Factors*, W. Karwowski, Ed., Taylor and Francis, Philadelphia, PA.

Black, F., and Scholes, M. (1973), "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, pp. 637–659.

Boer, F. P. (1998, September–October), "Traps, Pitfalls, and Snares in the Valuation of Technology," *Research Technology Management*, pp. 45–54.

Boer, F. P. (1999), *The Valuation of Technology: Business and Financial Issues in R&D*, Wiley, New York.

Brigham, E. F., and Gapenski, L. C. (1988), *Financial Management: Theory and Practice*, Dryden, Chicago, IL.

Burke, J. (1996), *The Pinball Effect: How Renaissance Water Gardens Made the Carburetor Possible and Other Journeys through Knowledge*, Little, Brown, Boston.

Christensen, C. M. (1997), *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*, Harvard Business School Press, Boston.

Cooper, R. G., Edgett, S. J., and Kleinschmidt, E. J., (1998a), *Portfolio Management for New Products*, Addison-Wesley, Reading, MA.

Cooper, R. G., Edgett, S. J., and Kleinschmidt, E. J. (1998b), "Best Practices for Managing R&D Portfolios," *Research Technology Management*, Vol. 41, No. 4, pp. 20–33.

Davenport, T. O. (1999), *Human Capital: What It Is and Why People Invest It*, Jossey-Bass, San Francisco.

Economist, (1999, June 12), "A Price on the Priceless: Measuring Intangible Assets," *Economist*, pp. 61–62.

EVM (2000), Earned Value Management Center, http://evms.dcmdw.dla.mil/.

Fuchs, R., McCarthy, J., Corder, J., Rankine, R., Miller, W., and Gawron, V. (1997), "United States Air Force Expeditionary Forces," Air Force Scientific Advisory Board, Washington, DC.

Gill, B., Nelson, B., and Spring, S. (1996), "Seven Steps to New Product Development," in *The PDMA Handbook of New Product Development*, M. D. Rosenau, Jr., Ed., Wiley, New York, Chapter 2.

Gramlich, E. M. (1997), *A Guide to Cost-Benefit Analysis*, Waveland, Prospect Heights, IL.

Hammond, J. S., Keeney, R. L., and Raiffa, H. (1998), *Smart Choices: A Practical Guide to Making Better Decisions*, Harvard Business School Press, Boston, MA.

Keeney, R. L., and Raiffa, H. (1976), *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, Wiley, New York.

Kirkwood, C. W. (1979), "Pareto Optimality and Equity in Social Decision Analysis," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 9, No. 2, pp. 89–91.

Layard, R., and Glaister, S., Eds. (1994), *Cost-Benefit Analysis*, Cambridge University Press, Cambridge, UK.

Lint, O., and Pennings, E. (1998), "R&D as an Option on Market Introduction," *R&D Management*, Vol. 28, No. 4, pp. 279–287.

Luehrman, T. A., (1998, July–August), "Investment Opportunities as Real Options," *Harvard Business Review*, pp. 51–67.

Luenberger, D. G. (1997), *Investment Science*, Oxford University Press, Oxford, UK.

Matheson, D., and Matheson, J. (1998), *The Smart Organization: Creating Value through Strategic R&D*, Harvard Business School Press, Boston, MA.

Mintz, S. L. (1998, February), "A Better Approach to Estimating Knowledge Capital," *CFO*, pp. 29–37.

Nevins, J. L., and Winner, R. I. (1999, April), "Ford Motor Company's Investment Efficiency Initiative: A Case Study," Paper No. P-3311, Institute for Defense Analyses, Alexandria, VA.

Office of Science and Technology Policy (OSTP) (1998, August), "A National Obligation: Planning for Health Preparedness for and Readjustment of the Military, Veterans, and Their Families after Future Deployments," Presidential Review Directive No. 5, Office of Science and Technology Policy, Washington, DC.

Rastikis, L. (1998), "Human-Centered Design Project Revolutionizes Air Combat," *CSERIAC Gateway*, Vol. 9, No. 1, pp. 1–6.

Rouse, W. B. (1991), *Design for Success: A Human-Centered Approach to Designing Successful Products and Systems*, Wiley, New York.

Rouse, W. B. (1998), *Don't Jump to Solutions: Thirteen Delusions That Undermine Strategic Thinking*, Jossey-Bass, San Francisco.

Rouse, W. B. (2010), *The Economics of Human Systems Integration. Valuation of Investments in People's Training and Education, Safety and Health, and Work Productivity*, Wiley, New York.

Rouse, W. B., and Boff, K. R. (2001), "Strategies for Value: Quality, Productivity, and Innovation in R&D/Technology Organizations," *Systems Engineering*, Vol. 4, No. 2, pp. 87–106.

Rouse, W. B., and Boff, K. R. (2003a), "Value Streams in Science & Technology: A Case Study of Value Creation and Intelligent Tutoring Systems," *Systems Engineering*, Vol. 6, No. 2, pp. 76–91.

Rouse, W. B., and Boff, K. R. (2003b), "Cost/Benefit Analysis for Human Systems Integration: Assessing and Trading Off Economic and Non-Economic Impacts of HIS," in *Handbook of Human Systems Integration*, H. R. Booher, Ed., Wiley, New York, Chapter 17.

Rouse, W. B., and Boff, K. R. (2004), "Value-Centered R&D Organizations: Ten Principles for Characterizing, Assessing & Managing Value," *Systems Engineering*, Vol. 7, No. 2, pp. 167–185.

Rouse, W. B., Boff, K. R., and Thomas, B. G. S. (1997a), "Assessing Cost/Benefits of R&D Investments," *IEEE Transactions on Systems, Man, and Cybernetics—Part A*, Vol. 27, No. 4, pp. 389–401.

Rouse, W. B., Kober, N., and Mavor, A., Eds. (1997b), *The Case for Human Factors in Industry and Government*, National Academy Press, Washington, DC.

Rouse, W. B., Howard, C. W., Carns, W. E., and Prendergast, E. J. (2000), "Technology Investment Advisor: An Options-Based Approach to Technology Strategy.

Information ● Knowledge ● Systems," *Management*, Vol. 2, No. 1, pp. 63–81.

Roussel, P. A., Saad, K. N., and Erickson, T. J. (1991), *Third Generation R&D: Managing the Link to Corporate Strategy*, Harvard Business School Press, Cambridge, MA.

Sage, A. P. (1977), *Systems Methodology for Large-Scale Systems*, McGraw Hill, New York:.

Sage, A. P., and Armstrong, J. (2000), *An Introduction to Systems Engineering*, Wiley, New York.

Sage, A. P., and Rouse, W. B., Eds. (1999), *Handbook of Systems Engineering and Management*, Wiley, New York.

Smithson, C. W. (1998), *Managing Financial Risk: A Guide to Derivative Products, Financial Engineering, and Value Maximization*, McGraw-Hill, New York.

Strassman, P. A. (1999, September), "Does Knowledge Capital Explain Market/ Book Valuations?," *Knowledge Management*, available: www.strassman.com/pubs/km.

# CHAPTER 41

# METHODS OF EVALUATING OUTCOMES

**Paula J. Edwards**
**HIMformatics**
**Atlanta, Georgia**


**François Sainfort, Julie A. Jacko, and Molly McClellan**
**University of Minnesota**
**Minneapolis, Minnesota**


**Thitima Kongnakorn**
**United BioSource Corporation**
**Lexington, Massachusetts**

## 1 INTRODUCTION

An underlying goal of all human factors research is to produce research that is compelling and meaningful. Meister (2004) suggests that this is accomplished in part by selecting appropriate research objectives and through careful experimental design. In Chapter 11, designing an experiment to collect reliable, valid data relevant to examining the identified research objectives was discussed. In this chapter we focus on another aspect of sound experimental design: selecting appropriate measures and outcomes to collect during an experiment and selecting appropriate methods of analyzing the experimental outcomes and drawing conclusions from them.

As demonstrated in this chapter, selecting the appropriate data outcomes to collect and the appropriate methods of analyzing the outcomes go hand in hand.

Certain characteristics of the data collected, such as the measurement type, drive which evaluation methods may be used to analyze the data, which in turn influence the types of conclusions that can be drawn from the results of that analysis. Other characteristics of the data, such as level of objectivity and specificity, affect the conclusions that can be made from the analysis and how or where those conclusions can be applied.

To help human factors researchers select the appropriate data outcomes and analysis methods for their research, we define characteristics of various outcome data and describe analysis and evaluation methods frequently used to analyze those outcomes. We begin by characterizing outcomes along a number of dimensions and discussing the implications these characteristics have on selecting both appropriate outcome measures

and appropriate evaluation method(s). Next, a variety of methods of evaluating both structured and unstructured outcome data are described. For structured data, statistical analysis methods frequently used in human factors are presented. For each method, the purpose, assumptions, methods, and results are described as well as guidelines for interpreting the results and drawing conclusions. Next, several methods of analyzing unstructured data, such as content analysis, are presented. The chapter concludes with an example of applying methods of analyzing structured and unstructured data to the analysis of survey data. Applying this knowledge of outcome characteristics and evaluation methods should enable human factors researchers to produce research outcomes and conclusions that provide compelling and meaningful insight into the field of human factors.

## 2   TYPES OF OUTCOMES

The types of outcome data that are produced in human factors research are as varied as the humans they strive to measure and analyze. To choose the appropriate data to collect to support the research objectives, we must first understand the nature and characteristics of the data. Outcome data and measures can be classified along a variety of dimensions. The dimensions most relevant to selecting appropriate outcomes and analysis methods are (1) level of structure, (2) level of objectivity, (3) specificity, (4) measurement type, and (5) multiplicity.

### 2.1   Level of Structure

In the dictionary (Merriam-Webster, 2010), *structure* is defined as "something arranged in a definite pattern of organization." Depending on the research methods used, the resulting data may be structured, unstructured, or a combination of the two. The level of structure is one of the most significant factors driving the methods appropriate for analyzing your data.

Based on this definition of structure, *unstructured data* are not arranged in a definite pattern. Unstructured data include descriptions, observation notes, answers to open-ended questions, video and audio recordings, and pictures. Field studies typically generate large amounts of detailed data that reflect the richness of the work being observed (Wixon, 1995). In a literature review, Kujala (2003) found that the authors of research papers involving field studies felt that the data gathered were invaluable and helped them in understanding the customer/user's needs, and the participants responses were positive. Field studies usually obtain unstructured data in the form of notes on observed actions and recordings of observed activities. Whereas the unstructured, or qualitative, data resulting from these observations contain invaluable detail on the activity being observed, the raw, unstructured data generally use more subjective methods of analyzing outcomes. Additionally, because unstructured data are so rich in detail, it can be more difficult to present findings clearly with this type of data. Because of their disadvantages, researchers often use unstructured data to produce structured data, as in the case of coding or classifying types of activities observed, or create figures and tables to summarize and communicate

the details in a more structured manner. These methods of analyzing unstructured data are addressed in more detail in Section 5.

*Structured data*, conversely, have a definite pattern in the way they are collected and stored. For example, consider a survey that asks: What factor is most important to reducing errors? If the question were open ended, the answer would be unstructured, since the subject could write in anything that comes to mind. On the other hand, if the question were multiple choice, the answer would have structure since all participants would have to choose from a limited number of available choices. Structured data typically lend themselves to quantification and are often referred to as *quantitative data*. Other examples of structured data include category classifications, rating scales, counts of events, and times/durations. When the data are structured, a wider range of analysis methods are available for evaluating the outcomes. In the case of structured data, the data items are sometimes also referred to as *measures*.

### 2.2   Level of Objectivity

*Objectivity* is "expressing or dealing with facts or conditions as perceived without distortion by personal feelings, prejudices, or interpretations" (Merriam-Webster, 2010). When it comes to human factors outcomes, the level of objectivity is a continuum with objective and subjective at either extreme and considerable gray area in between. At one end of the spectrum we have *objective data*, which are recorded "without the aid or expression of the subject whose performance is being recorded" (Meister, 2004, p. 81). For purely objective data, task performance is recorded manually or automatically, with minimal human involvement in the measurement. For example, time to complete a task, missed targets, height, and other physical measures are objective. *Subjective data*, on the other hand, are based on the subject or experimenter's opinions, values, and interpretations. Subjective measures rely on human perception, cognition, judgment, and experience (Wickens et al., 2003). Examples of subjective measures include responses from interviews/questionnaires, verbal reports of activities or thought processes, self-ratings, and other personal judgments. To illustrate the continuum of objectivity, consider verbal reports collected during an experiment. The validity of verbally reported thought sequences depends on the time interval between the occurrence of a thought and the verbal report, where the highest validity is seen in concurrent, think-aloud verbalizations and the lowest is retrospective reports where participants create inferential biases in the reported information (Ericsson, 2006).

So which is better, objective or subjective? The answer, of course, depends on the research objectives. Although the level of objectivity has little effect on which methods may be used to analyze the data, it has a great effect on the validity and the generalizability of the research and the conclusions that can be drawn from the research. In human factors, the subjects' subjective opinion of their performance, confidence, or workload is sometimes of as much or more interest than their objective performance on the task. In many cases, it is

useful for human factors researchers to collect *both* subjective and objective measures. If the results of both the objective and subjective measures support a particular conclusion, there is greater confidence in the validity of that conclusion. Therefore, when selecting outcome measures, researchers should consider whether objective measures, subjective measures, or a combination are best suited to supporting their research objectives and the conclusions they hope to draw from their research.

### 2.2.1 Preference-Based Measures

Within the set of subjective measures is a subset of measures, referred to as preference-based measures, which indicate a subject's likes or dislikes based on his or her experience and values. Survey questions that ask what a subject likes, prefers, or values or that ask the subject to rate value or importance of items are examples of preference-based measures. Similarly, comparisons and choices among options presented to subjects also capture preferences.

These measures are worth distinguishing from other subjective measures because they are notoriously difficult to measure reliably. This is caused in part by the nature of values and preferences and the uncertainty present in applying values to different sets of options and choices. People tend not to know their preferences in an unfamiliar situation, especially those with lasting consequences, and even when they are known, those preferences tend to be labile (Fischhoff et al., 1988). Often, preferences are "constructed" as people learn about or experience options. When preferences are known, they are extremely difficult to measure. Research has shown that people tend to construct their preferences during the process of elicitation and the method used to elicit their preferences can affect their final expressed preferences (Slovic, 1995). For example, studies in health care have shown that the use of different methods of eliciting preferences (Chapman and Elstein, 2000), physician's explanations of treatment alternatives (Mazur and Hickam, 1994), and framing of information about alternatives (positive, negative, or neutral) (Llewellyn-Thomas et al., 1995) affected patients' stated treatment preferences. Additional research has furthermore demonstrated that almost half of health care patients have dominant preferences when faced with discrete choices and that both past experiences and the complexity of the decision task influenced their dominant preferences (Scott, 2002). Therefore, when preference-based subjective measures are of interest, special care should be taken during experimental design to ensure the validity of these measures. The researcher must take steps to ensure that the way the preference-based measures are collected does not bias the results.

### 2.3 Specificity

Specificity indicates whether a measure refers to a particular task, industry, or situation, as opposed to being applicable to a variety of areas. Again, this characteristic is a continuum ranging from specific to generic. *Specific measures* are tailored to measuring a phenomenon of interest but have little generalizability to other phenomena. At the other end of the scale, *generic measures* are generalizable to a variety of tasks or situations. However, it can be quite difficult to define generic measures that are sensitive enough to capture the phenomenon of interest in a variety of situations that may be quite different. For example, in human–computer interaction (HCI), target highlight time (THT) is a measure used to quantify the salience of feedback received from the interface as the user completes a drag-and-drop task. THT, a specific measure for drag-and-drop tasks, is much more sensitive than total task time to measuring reaction time to feedback because task time is strongly affected by a number of other factors. Task time is a generic measure, and although it is less sensitive in this case, it has the advantage of being generalizable to other computer tasks. So task time could be used to compare drag-and-drop performance to performance on a different computer task, such as point and click.

Another example relates to measuring health outcomes resulting from various treatments. A generic measure would capture a range of health status dimensions (e.g., physical function, mental health, social function). On the other hand, a disease-specific measure will zoom in on specific symptoms and implications related to a disease of interest (e.g., lower back pain).

The specificity of a measure affects the conclusions drawn from the data. If very specific measures are used, the results may not generalize to a broad enough range of situations, thereby limiting the scope and applicability of the conclusions. Conversely, if the measures are too broad, they may not be sensitive enough to detect the phenomenon of interest through the statistical analysis methods used. Therefore, it is crucial to reconcile the specificity of the measures with the research objectives as the study measures are chosen. In some cases, researchers choose to collect a combination of specific and generic measures. In selecting generic measures, it is recommended that researchers look at industry standard metrics and measures used in related research to enable comparability across studies.

### 2.4 Measurement Type

In structured data, the measurement type determines to a large degree the statistical methods that may be used to analyze the data. The measurement type indicates the amount of information the measure contains with respect to the value being measured (Sheskin, 2007). The four measurement types, in order from least to greatest information provided, are nominal/categorical, ordinal/rank order, interval, and ratio. These categories are defined as follows (Sheskin, 2007; Argyrous, 2000):

1. *Nominal/categorical*: indicates the category to which an item belongs. The category may be represented as a number or text, but even if the category is represented as a number, it cannot be manipulated mathematically in a meaningful way. In human factors, nominal data would include gender, race, and part number. For each of these measures, it is not possible to add or rank order them, since the name/number is used solely for identification purposes.

2. *Ordinal/rank order*: indicates rank orders using numbers but does not provide information on the magnitude of difference between two ranks. An example of ordinal data from human factors is participant rankings of the preferred tool or method. If the participant is asked to rank three proposed tools, from the one they prefer most to the one they prefer least, those rankings (1–3) do not indicate if the participant strongly prefers the first ranked system to the second or only slightly prefers the first to the second.

3. *Interval*: indicates both order and magnitude of difference between values by providing a number along a scale that has intervals of equal distance between equal values on the scale. This means that a difference of 10 between two values represents the same magnitude of change regardless of whether the initial value was small or large. However, interval scales arbitrarily assign a zero score rather than having a true zero. For example, consider a question that asks participants to rate on a scale from 1 to 10 how hard they had to work physically to complete a task. A one-point increase in "work" is the same one-point increase in work whether the initial rating is 3 or 7. However, there is no true measurement of zero work; 1 on the scale was selected arbitrarily to represent an extremely low workload.

4. *Ratio*: like interval measures, provides a number along a scale that indicates both order and magnitude of difference between values. Unlike interval measures, however, ratio measures have a true zero point. Because there is a true zero point, it is possible to compare scores by taking their ratios. For example, age has a true zero point, so one can meaningfully say that a person who is 40 is twice as old as someone who is 20. In contrast, it is not appropriate to compare ratios of work in the previous example. Since the true zero point for work is unknown, saying that a task with a work rating of 10 is twice as much work as a task with work rating of 5 is inaccurate.

The measurement type has an enormous influence on the methods that can be used to analyze the data. Table 1 indicates which type of analysis is appropriate for each measurement type. Several of these analysis methods are discussed in further detail in Section 4.

## 2.5 Dimensionality

Another important characteristic of structured outcome data is the degree of dimensionality. Some outcomes can be captured directly with one overall measure (e.g., the time required to perform a given task, or how one person feels about his or her own health at a point in time as measured by a global five-point Likert scale with five possible answers: excellent, very good, good, fair, or poor). However, many outcomes need to be captured through a number of dimensions. In the case of multiple dimensions, a complex measurement task often then consists of appropriately aggregating the various dimensions into one overall quantitative measure of the outcome of interest. For example, although performance on a job clearly contains many different aspects, one may wish to combine the performance on different aspects into an overall assessment of global performance. To represent overall performance adequately, one needs to understand the potential relationships between the various elements.

Many scales are constructed based on different items and thus appear to be inherently multidimensional. However, it is important to differentiate between a scale and an index. A *scale* typically is comprised of multiple items whose values are caused by an underlying construct (or latent variable) (Bollen, 1989). On the other hand, an *index* consists of several cause indicators or individual variables that together determine, or at least strongly relate to or influence, the level of the construct of interest (Devellis, 2003). Thus, in a scale, the items that comprise the scale typically correlate to each other, and multiplicity of items increases the overall reliability of the scale. On the other hand, for an index, constituent variables may be independent of each other (e.g., physical function versus social function, both important cause indicators of overall health). Creating an index may or may not be important in terms of analyzing and understanding outcomes. In terms of analyzing multiple outcomes, however, an index is desirable if one overall metric is desired to represent the outcomes. Furthermore, an index allows for unidimensional analytical approaches to be used, whereas multidimensional outcomes typically require multivariate analysis methods.

## 2.6 Summary: Selecting Appropriate Outcome Data

As demonstrated in the preceding sections, a variety of outcome characteristics influence the methods used to analyze outcomes and the conclusions that can be

**Table 1  Analyses Appropriate for Each Measurement Type**

| Measurement Type | Compare Counts/ Frequencies | Greater Than/ Less Than Comparisons | Ratio Comparisons | Mathematical Operations | Statistical Analysis |
|---|---|---|---|---|---|
| Nominal | × | | | | |
| Ordinal | × | × | | | |
| Interval | × | × | | × | × |
| Ratio | × | × | × | × | × |

drawn from those outcomes. Therefore, selecting the appropriate outcome data to collect begins with the definition of the research objectives. However, it is important to consider the research objectives in broad terms when developing a data collection and analysis plan for human factors research. Instead of just thinking in terms of the phenomenon to be studied or the intervention to be evaluated, also think about the goals for communicating the outcomes before determining which outcome data to collect. If these objectives and goals are not well established up front, it is unlikely that you will by chance collect data that supports an ill-defined research objective.

Once the research objectives are established and well understood, the next step is to select what data outcomes to collect. It is always useful to begin by looking at the domain literature to identify frequently used data and measures and any gaps in those outcomes. Using measures consistent with other research is useful in that it is a prerequisite for having results that are comparable with other studies. However, do not be afraid to bridge any gaps that exist in the literature by creating new measures. For example, there may be a need for a new, more sensitive measure of a phenomenon of interest or a more generalizable measure that enables comparisons across multiple related tasks.

Next consider the types of conclusions that should be drawn to support the research objectives. What level of objectivity is required? What specificity? What analysis method(s) will enable drawing those conclusions? Do the required analysis methods impose any restrictions on the structure or measurement type of the outcome data? All of these questions should be answered to develop a data collection and analysis plan for the research study. Addressing these topics up front helps ensure that the outcome data collected will be valid and credible and will support the research objectives identified. Of course, this does not guarantee that the outcomes will always produce the expected results—human factors data are always full of surprises!

## 3  MEASUREMENT OF OUTCOMES

Measurement is a fundamental activity of science. As Krantz et al. (1971, p. 1) explain: "When measuring some attribute of a class of objects or events, we associate numbers (or other familiar mathematical entities, such as vectors) with the objects in such a way that the properties of the attribute are faithfully represented as numerical properties." Although this process can be relatively straightforward for physical measures such as length or density, it can be very difficult for psychosociological constructs, such as stress or health status. Whether one measure is created or multiple measures are used, fundamental psychometric properties need to be tested properly before using the measurement system created. These fundamental properties are described briefly in the next section. In addition, we describe briefly methods that can be used when multidimensional outcomes need to be aggregated into an overall scale or an index.

### 3.1  Psychometric Properties

There are two overall fundamental properties in measurement: reliability and validity. Ghiselli et al. (1981) consider reliability a fundamental issue in psychosociological measurement and, in the context of developing scales, define it as the proportion of variance attributable to the true score of the latent variable. Although several terminologies exist, there are essentially three types of validity in scale development: content validity, criterion-related validity, and construct validity. *Content validity* refers to the extent to which a set of items selected in a scale covers the content domain. *Criterion-related validity* refers to the extent to which the scale created relates to an existing criterion or "gold standard." Finally, *construct validity* refers to the extent to which a scale "behaves" as it is expected to, according to theoretical relationships with existing constructs, where these relationships have been formulated prior to the development of the scale. A number of techniques exist to ascertain the reliability and validity of scales (see, e.g., Devellis, 2003).

### 3.2  Multidimensional Outcomes

As mentioned above, many (complex) evaluation problems are by nature multidimensional and require the construction of a scale or an index. Scales and indices differ in fundamental ways and require very different techniques for their development. Devellis (2003) provides a structured eight-step process as a guideline for scale development:

1. Determine clearly what to measure.
2. Generate an item pool.
3. Determine the format for measurement.
4. Have the initial pool reviewed by experts.
5. Consider the inclusion of validation items.
6. Administer the items to a development sample.
7. Evaluate the items.
8. Optimize the scale length.

As part of the final step, data reduction techniques for creating scales are commonly used and are described in greater detail in Section 4.3.

As opposed to creating a scale, multiattribute utility theory (MAUT) can be applied directly to create an index. Edwards and Newman (1982, p. 10) distinguish "four different classes of reasons for evaluations: curiosity, monitoring, fine tuning, and programmatic choice.... These reasons for evaluation share two common characteristics that make MAUT applicable to them all. The first is that, implicitly or explicitly, all require comparison of something with something else.... The second characteristic is that [entities to be evaluated] virtually always have multiple objectives." Thus, MAUT is applicable to many situations where multiple outcomes need to be aggregated into an overall index.

These multiple objectives lead to the identification of what Keeney and Raiffa (1976) call *evaluators* or *attributes* and purport to describe completely the
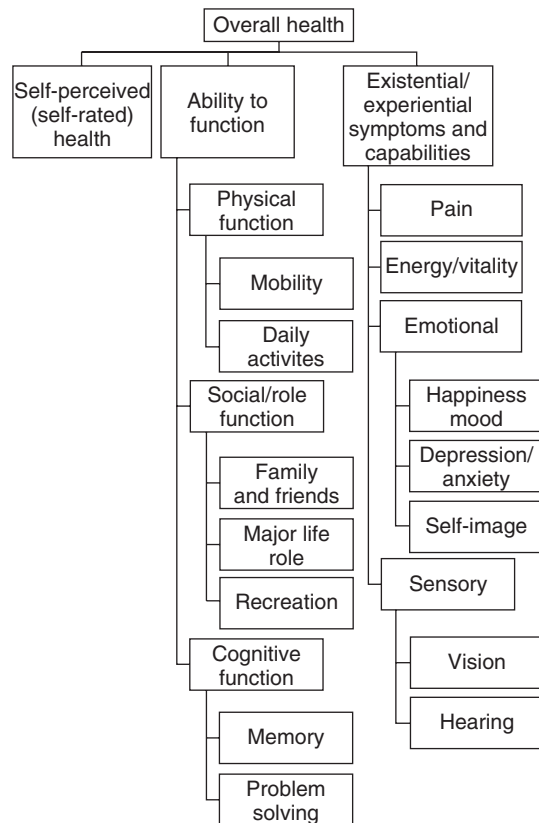
**Figure 1** Hierarchy of attributes to measure overall health. [Reprinted with permission from Fryback (1998) by the National Academy of Sciences. Courtesy of National Academies Press, Washington, DC.]

consequences of any of the possible actions or entities to be evaluated. According to Keeney and Raiffa, each attribute itself must be comprehensive and measurable, and the set of attributes describing the consequences must be complete, operational, decomposable, nonredundant, and minimal (p. 50). Often, these attributes can be structured meaningfully into a hierarchy. On top of the hierarchy is the all-inclusive objective, which indicates the reason for being interested in the problem in the first place. Figure 1 illustrates such a hierarchy in the context of creating an index comprising multiple outcomes to measure overall health.

MAUT provides a way of aggregating the information describing each entity on the multiple attributes into a summary measure or index. As described by Edwards and Newman (1982, p. 79), "the goal of MAUT is to come up with one number for each [entity] of evaluation, expressing in highly concentrated form how well that [entity] does on all evaluative dimensions. But whether that much compression is appropriate depends very much on the purpose of the evaluation." MAUT is widely used to combine multiple outcomes for at least three reasons. First, it is very appealing and convenient

to have a summary index. It is especially useful to have a summary index when the objective of the evaluation is to monitor changes, to compare alternatives, or to assign another quantity in proportion to the index score (e.g., assigning monetary rewards as a function of performance). Second, as pinpointed by von Winterfeldt and Edwards (1986), in practice, the additive and multiplicative models are the only workable ones; they are relatively simple and therefore easily accepted forms of aggregation. Third, MAUT is grounded in theory: The standard decomposition theorem for additive and multiplicative utility functions is seen as a determinant, to provide theoretical justification for the development of such evaluative models.

## 4  ANALYSIS OF STRUCTURED OUTCOME DATA

Both unstructured and structured outcome data result from human factors research. However, our discussion of analysis methods will begin by focusing on analyzing structured outcome data, or measures, since the results from these analytical methods are frequently the primary focus of research results presented in the human factors literature. There are a variety of statistical and graphical methods for exploring and analyzing structured data. In this section we review methods that are frequently used by human factors researchers. The methods are grouped according to the objective of the analysis. For each method we discuss when it is appropriate to use the method, the type of results produced by the method, and how to draw conclusions from the results. Delving into the statistical details of these methods is beyond the scope of this chapter, so suggested statistics reference(s) are also provided for each method. The decision model presented in Figure 2 is provided to help researchers decide which statistical tests are appropriate based on their analysis objectives and characteristics of the data.

### 4.1  Exploring Your Data

As Box et al. (1978) point out, when doing statistical analysis on outcome data, it is important not to forget what you know about the subject matter in your field. One way to build subject matter knowledge is to explore your data prior to completing any statistical tests. Human factors data can be quite different from data found in other domains. One of the distinguishing characteristics of human factors data is that it tends to be very noisy. This is especially true when the population being studied is very heterogeneous. Because the people being studied vary in physical and mental capabilities, expertise, and other factors, their performance on the same task will naturally vary. To compound this, even the same person acting under varying environmental factors may perform differently. The noise inherent in human factors data can reduce the power of statistical methods to detect effects. This is why it is very important to get to know your data before you start running statistical tests. By exploring and getting to know your data, you develop an initial understanding of potential occurrences and trends that can be used to validate and spot potential problems in the statistical analysis.
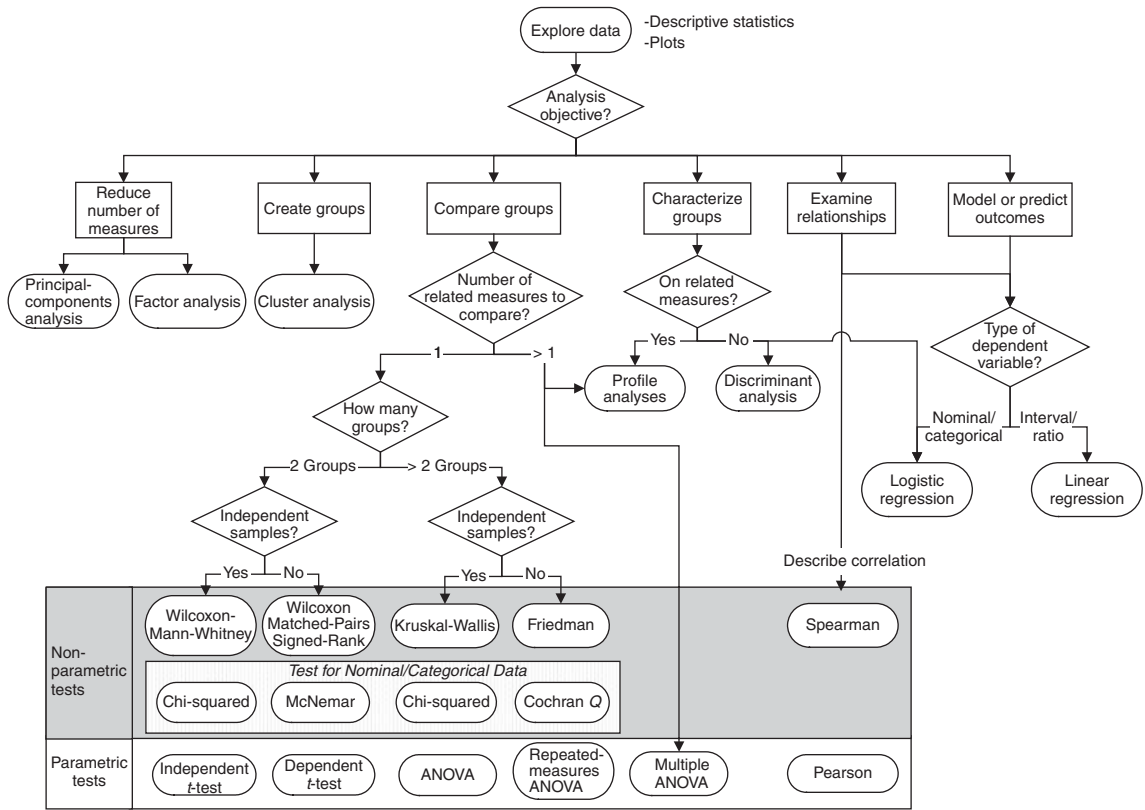
**Figure 2** Decision model for analyzing structured measures.

### 4.1.1 Exploring the Data Distribution

There are several graphical techniques useful for getting to know your data. For example, box plots (Figure 3) and histograms (Figure 4) may be used to get an overall feel for the distribution of the data. Both of these types of plots illustrate the minimum value, maximum value, and overall distribution and variability of the data. Additionally, a box plot indicates quartiles and the median value and highlights data points that are outliers or extreme values.

Identifying extreme values and outliers is important, as these data points are far outside the expected range of values, which is determined based on the variability of the data, measured by the standard deviation. These data points should be examined since they may be caused by a data collection or related error. For example, the value may be the result of a data entry error. In some cases, an unexpected event may have occurred during the trial, causing the data to be invalid. For example, in a task that is being timed, if the subject starts the task, then pauses to ask the experimenter a question, the task time for that trial may be artificially inflated and may need to be excluded from the data analysis. In another example, if all of the extreme values and outliers are from the same participant, there may be some characteristic of that participant that is causing the person to perform very

differently from the others. For example, one participant may have more experience with the experimental task, causing the person to perform much better than the other participants. The researcher needs to be aware of the cause of this difference and make an educated decision about whether or not it is appropriate to include the participant in the data analysis. If participant(s) or trials are excluded from analysis, the reason for excluding them should be presented when reporting the study findings, to ensure that correct conclusions are drawn from the study.

### 4.1.2 Descriptive Statistics

Descriptive statistics are used to convey information about the central tendency and dispersion of the data set. One frequently reported measure of central tendency is the *mean*, or average, value. However, one must be careful when reporting means because extreme values, which should be identified in your data exploration, can distort the mean. For example, if a researcher wanted to examine the net worth of a group of 10 people selected at random and Warren Buffet happened to be one of those people, Warren Buffet's net worth (in the billions) would distort the mean, making it in the hundreds of millions even if the net worth of all the others in the group was less than $100,000. Because the mean is vulnerable to
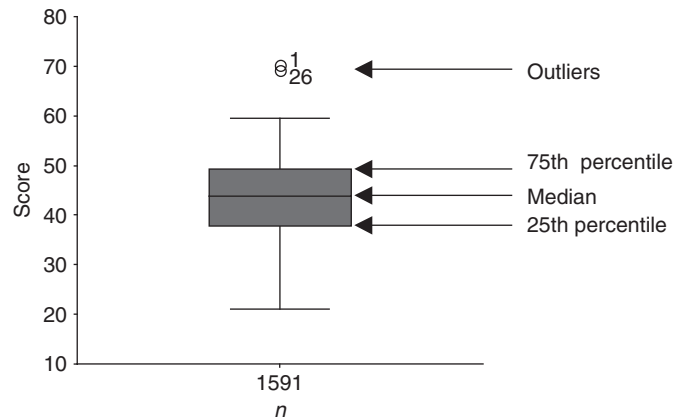
**Figure 3**   Box plot.
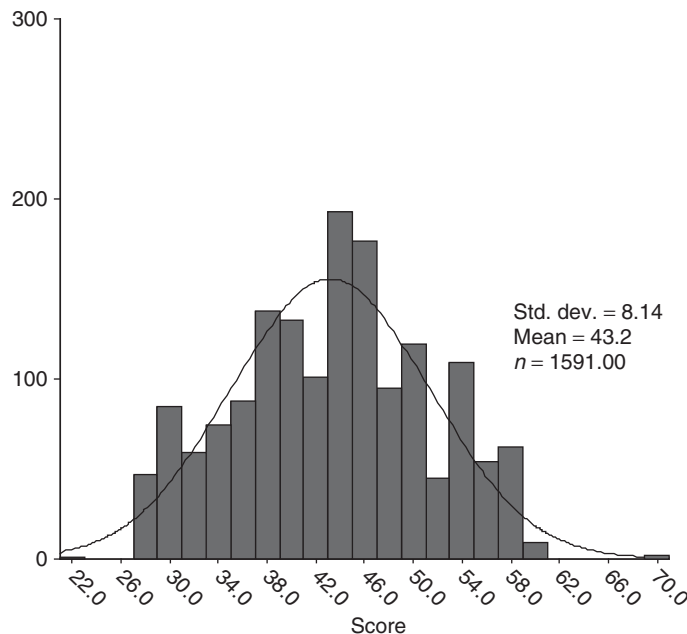


**Figure 4**   Histogram.

this distortion, it is often useful to look at the median and mode as well. The *median* is the value that splits the data set in half, with 50% of the values greater than that value and 50% less than that value. To find the median, the values are arranged in order from least to greatest; the middle value is the median. If there is an even number of values, the median is the average of the two middle values. The *mode*, on the other hand, is the value that occurs most frequently.

   In addition to examining measures of central tendency, examine the dispersion of the data as well. Dispersion is important for two primary reasons. First, dispersion, or variability, affects the power of statistical tests to detect significant effects, which is discussed in more detail later in the chapter. Second, in human factors, researchers are sometimes interested in reducing variability to gain more consistent performance over time. As a measure of dispersion, researchers frequently report the *variance* ($\sigma^2$) or *standard deviation* ($\sigma$). These two values are directly related, as the standard deviation is the square root of the variance. The variance is calculated based on the square of the difference between each value and the mean and therefore increases as more values are a greater distance from the mean. Statistical packages and spreadsheet programs calculate this value, so the mathematical equation is not provided here. However, any statistics book provides the mathematical definitions of these measures.

When discussing dispersion, it is also often worthwhile to examine the *range* of the values, defined by the *minimum* and *maximum values*, and *percentiles*. The *X* percentile is defined as the value at which *X*% of the values fall at or below that value. The range and quartiles (the 25th, 50th, and 75th percentiles) are depicted in a box plot, as discussed in Section 4.1.1.

*References for further details*: Sheskin (2007), Field (2005).

## 4.2  Statistical Analysis Methods

A variety of statistical analysis methods can be used to make inferences about structured outcome measures. The purpose of these statistical methods is to help you understand and draw conclusions from the outcome measures. The appropriate statistical test to use depends on the analysis objective and on characteristics of the data. The reader is again referred to the decision model in Figure 2 for assistance in choosing the appropriate method(s). The methods discussed in this section may be used to accomplish the following objectives:

- *Comparing groups'* *t*-test, Wilcoxon–Mann–Whitney, Wilcoxon matched-pairs signed-rank, chi-squared, analysis of variance (ANOVA), repeated-measures ANOVA, Kruskal–Wallis, Friedman
- *Characterizing groups*: profile analyses, discriminant analysis
- *Creating groups*: cluster analysis
- *Describing and modeling relationships*: correlation, linear regression, logistic regression

Since this is a book on human factors and ergonomics, not statistics, formulas and detailed mathematical explanations are not provided for these statistical methods. Since most statistical packages provide functions that automate these calculations, the focus is on understanding conceptually how each method works, when it is appropriate to use it, and how to understand and interpret the results in order to draw conclusions. Before delving into the methods, several general statistical concepts need to be reviewed.

### 4.2.1  General Statistical Concepts

In inferential statistical methods, the researcher seeks to determine whether or not phenomena observed in the data are caused by random variation in the data. If it is not caused by random variation, we can make inferences about the nature and cause of those phenomena. In making these inferences, several general statistical concepts play a role both in designing experiments and in interpreting and communicating results. In this section we review these concepts in the context of an example experiment in which a researcher wants to compare the time to complete a task using the current method to the time using a proposed new method. The researcher collects the data and computes the mean time to complete the task under each method. The new method has a lower mean time, but how can they be sure that the lower time is actually caused by using the new method instead of just by chance?

***Type I and II Errors***   Inferential statistical methods typically frame the research question as a hypothesis, which is then tested to determine whether or not it is true. In our example, the researcher's hypothesis would be that there is a difference between the mean times of the two methods. When testing a hypothesis, researchers are vulnerable to making two types of errors. First, if the researcher concludes that the hypothesis is true when the result is actually caused by chance, he or she has made a *type I error*. In our example, the researcher would make a type I error if he or she concluded that the new method reduced the time to complete the task when in reality the lower mean time was only caused by chance variation in the data. To avoid these errors, researchers typically limit the type I error rate ($\alpha$). In human factors research, $\alpha$ is usually set at 5%, which means that type I errors will occur at most in 5 of every 100 tests.

In the second type of error, *type II error*, the researcher concludes that the result was caused by chance when, in reality, the hypothesis is true. If our researcher concluded that there was no difference in the times for the two methods when the proposed method is actually faster, they would commit a type II error. The likelihood of making a type II error ($\beta$) is an indicator of the *power* ($1 - \beta$) of the test, its ability to detect a difference when one actually exists. The type I error rate ($\alpha$), the sample size ($n$), the mean, and the variance of the data determine the power of the test (Kutner et al., 2004). For researchers designing experiments, understanding this relationship between $\alpha$ and the power of the test is crucial. Since the industry practice is to control $\alpha$ at no more than 0.05, the only way to increase the power of a test is to collect more data, increasing $n$, or control the experiment to reduce variability in the data, which can be difficult with human factors data, as mentioned in the discussion on experimental design in Chapter 11. Some statistical packages, such as MiniTab, provide calculators that estimate the sample size needed to achieve the desired power given an estimate of the variance and mean. When designing an experiment, it is highly recommended that you estimate the power to ensure that enough data are collected to achieve the desired power for statistical testing.

***Experimentwise Error***   In situations where more than two groups are being examined or several, possibly related, dependent measures are being examined, researchers need to be careful to manage the *experimentwise error* rate to ensure the validity of their results. Experimentwise error is the combined type I error rate for all the statistical tests being performed. Take a simple example. A researcher is comparing the task time of three experimental groups. If the researcher uses the *t*-test to compare each test to the other tests, three *t*-tests are completed, comparing group 1 to 2, 1 to 3, and 2 to 3. If the researcher sets the type I error rate ($\alpha$) to 0.05, each test has a 0.95 chance that there will be no type I errors. Since there are three tests, each of which is assumed to be independent, the experimentwise probability that there is no type I error is $0.95 \times 0.95 \times 0.95 = 0.857$. This means that the experimentwise error rate ($1 - 0.857$) is 0.143. In other words, 14.3% of the time there will be a type I error!

However, statisticians are aware of this phenomenon, so tests are available that account for it and control the experimentwise error to the desired $\alpha$ level. For example, this is why ANOVA is used to compare more than two groups instead of multiple $t$-tests, as demonstrated in the example. ANOVA examines all of the groups together to determine whether any of them differ significantly at the $\alpha$ level of experimentwise error. Once it is established that there is a significant difference at the experiment level, post hoc comparisons are completed to determine which pairs are different. These post hoc comparisons adjust to account for the number of comparisons being performed.

***p-Values***   Researchers generally provide $p$-values when reporting the results of statistical analyses. The $p$-value indicates the likelihood of making a type I error. In other words, the $p$-value is the probability that the researcher will conclude that the hypothesis is true when the result is actually caused by chance. To continue the task time comparison example, if the statistical test resulted in $p = 0.25$, it would indicate that there is a 25% chance that the difference in the mean task time is simply caused by random variation in the data. This is clearly higher than the 5% $\alpha$ threshold that is generally accepted, so in this case, the researcher would conclude that there is not evidence to support the conclusion that there is a difference in the task times. This means that either there truly is no difference in the task times or the test did not have enough power to detect the difference, in which case the researcher could collect additional data to increase the power of the test. On the other hand, if the $p$-value were 0.03, it would indicate only a 3% chance that the difference in times is due to chance, and the researcher could conclude that the new process affects the task time.

***Confidence Intervals***   When comparing groups, the $p$-value indicates whether or not there is a difference between the groups but gives no indication of the magnitude of this difference. Confidence intervals (CIs) provide this information, making the results of the study easier for practitioners to interpret and apply. CIs indicate the magnitude for a value such as the mean or the mean difference between two groups. Because the mean is calculated from a sample of data, it merely provides an estimate for the actual population mean. Consequently, with a different sample of data, the estimate of the mean, although close to the first mean, is unlikely to be exactly the same. A CI provides a range in which the actual mean falls. The confidence interval is for a certain $\alpha$ level, typically $\alpha = 0.05$, and is referred to as the $(1 - \alpha)$-level CI. The interval is calculated using the mean and the variance (Wu and Hamada, 2009). In a 95% confidence interval, there is a 95% chance that the actual mean falls within the upper and lower bounds of the interval. How does knowing this help us?

In our example, the researcher might want to understand the magnitude of the difference in task time between the current method and the method proposed. If the 95% CI for the difference in task time (time$_{current}$)$-$(time$_{proposed}$) was found to be 33–45 s, it would indicate that the proposed process saves 33–45 s on the task.

***Statistical versus Practical Significance***   Just because an analysis indicates that a result is statistically significant does not mean that it has practical significance as well. For example, in the CI presented previously, we reported that the time savings achieved by using the proposed method is 33–45 s. If the task normally takes 60 min to complete, this time savings amounts to a mere 1% saving, so it may not be cost-effective to change to the new method. On the other hand, if the task normally takes 5 min, the time savings amounts to an 11–15% savings, a much more practically significant result. This illustrates why it is important for researchers to understand and report practical significance as well as statistical significance. Reporting practical significance makes it easier for people less familiar with statistics and human factors methods to understand and apply the results, increasing the use and impact of the results.

***Parametric and Nonparametric Tests***   In inferential statistics, many tests are classified as parametric or nonparametric. This distinction is made because *parametric tests* are based on certain assumptions about characteristics of the distribution of the data, whereas *nonparametric tests* make no such assumptions about the distribution of the data (Sheskin, 2007; Sprent and Smeeton, 2007). For example, when comparing two independent samples, the $t$-test, a parametric test, or the Mann–Whitney $U$-test (a nonparametric test) may be used to analyze the data. Which should the researcher choose?

In general, parametric tests are appropriate for interval and ratio data, whereas nonparametric tests are used for categorical/nominal and/or ordinal/rank-order data. With interval and ratio data, it is always preferable to use parametric tests because they have more statistical power. However, when using parametric tests, it is important to verify that the underlying assumptions of the test are met. Although many of these tests are robust enough to handle some departures from the assumptions (Sheskin, 2007; Newton and Rudestam, 1999), large departures from the assumptions may make the test inappropriate for use with the given data set. For example, the $t$-test assumes that the data being analyzed are characterized by a normal distribution (normality assumption) and that the variance of the underlying population is homogeneous (homogeneity of variance assumption). In the discussion of each test, the assumptions are listed and, if appropriate, ways to validate those assumptions. However, since many parametric tests assume that the data or the error between the data and a model of the data are normally distributed, it is worth recalling the normal distribution. Figures 3 and 4 are a box plot and histogram, respectively, of data that resemble a normal distribution. The reference line in the histogram displays how data with a normal distribution would be shaped—with a large number of values in the middle near the mean and progressively fewer occurrences as you move farther away from the mean. In the box plot, the line representing the median is located in the center of the box and the box and lines are fairly balanced. All of these things indicate that the
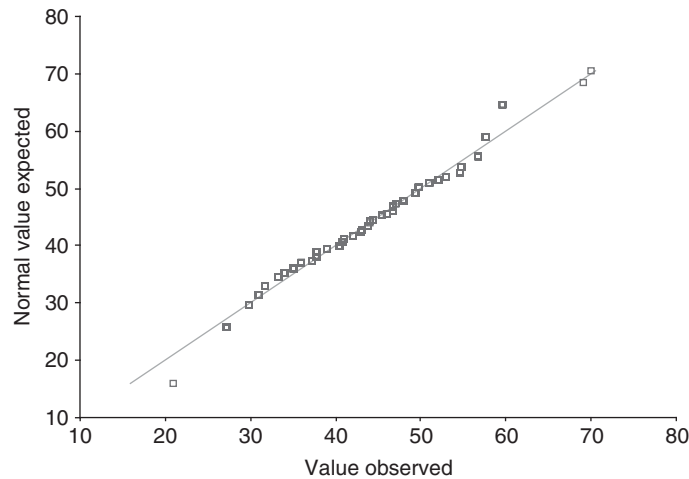
**Figure 5** Normal probability plot (normally distributed data).

data resemble a normal distribution. To confirm this, examine the normal probability ($Q$–$Q$) plot for these data (Figure 5). Since most of the data points fall along a straight line, this also indicates that the data resemble a normal distribution. In this case, we could conclude that the assumption of normality is met.

When significant violations of the parametric test assumptions are observed, the researcher has two alternatives. One alternative is to transform the measure ($y$) using one of the power transformations such as $\log(y)$ or $1/y$, so that the transformed measure meets the test assumptions. However, when using a transformation, researchers should take care to ensure that the results of the analysis are interpretable. This means that the results of the analysis on the transformed measure must still be meaningful (e.g., if $y$ = death rate, then $1/y$ = survival rate and the $1/y$ transformation has an interpretable meaning). Also, the researcher must take care to present the results in a manner that is clear, even to those with limited statistical knowledge. Many statistics books (e.g., Kutner et al., 2004; Newton and Rudestam, 1999; Wu and Hamada, 2009) provide details on how to select the appropriate data transformation. Refer to one of these books for more information on this topic.

The second alternative when a parametric test is not appropriate is to use an equivalent nonparametric test. In our $t$-test example, if the data set significantly violated the normality and homogeneity of variance assumptions, the researcher could instead use the Mann–Whitney test to compare the two groups. This test is based strictly on the rank order of the data points, so it makes no underlying assumptions about the distribution of the data or its variance. However, because it is based on the ranks, it sacrifices the additional information provided in the interval/ratio data, making it less powerful. This tradeoff between the power provided by parametric tests and the absence of data distribution assumptions in nonparametric tests is crucial for researchers when selecting the appropriate test for their data.

Note that there is some debate in the applied statistics community over whether a parametric or a nonparametric test should be used when there are departures from the parametric test assumptions. However, as Sheskin (2007) demonstrates with examples in his book, frequently, when both a parametric test and its nonparametric counterpart are applied to the same data set, they result in the same or similar conclusions. Therefore, the prudent researcher should use the decision process illustrated in Figure 6 when trying to decide whether a parametric or a nonparametric test is appropriate.

### 4.2.2 Comparing and Creating Groups

Now that the review of general statistical concepts is complete, we can focus on the methods used to accomplish the researcher's analysis goals. In human factors research, researchers are frequently interested in examining groups of participants, items, or events. For pre-existing groups, the researcher may want to compare groups on some dependent measure of interest or characterize those groups based on a number of different measures. In other cases, the researcher may be interested in creating new groups of related participants, items, or events to develop a better understanding of relationships and patterns in the data.

Of these three general analysis goals related to groups, comparing two or more experimental groups on a dependent measure of interest is the one most frequently seen in human factors research. For example, a researcher may be interested in examining worker efficiency, measured by task time, when using each of several available tools. After designing the experiment and collecting the data, as described in Chapter 11, the researcher must use the appropriate statistical test to analyze the data and draw conclusions about the influence of each tool on task time. In the first part of this section we review statistical tests used to compare two groups (e.g., $t$-test, Mann–Whitney $U$-test) and to compare two or more groups (e.g., ANOVA, Friedman test).
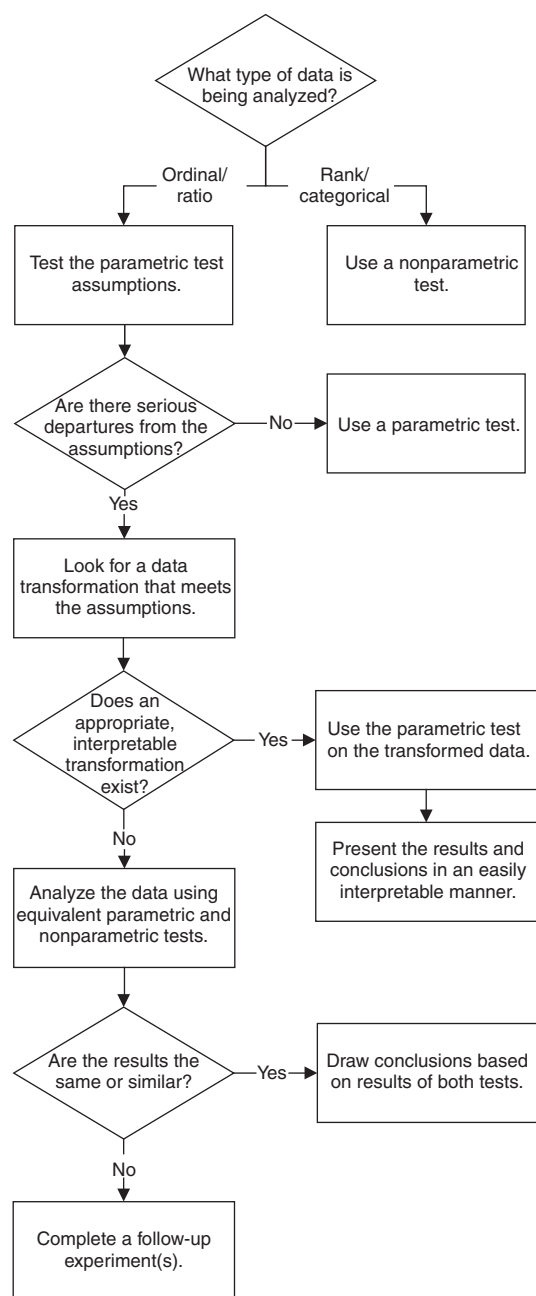
**Figure 6** Decision model for choosing between a parametric and a nonparametric test.

After the review of methods for comparing groups, two methods are presented for characterizing groups. The first is profile analyses, which characterize each group on a set of related measures. Profile analyses may also be used to compare the profiles of each group to determine whether or not they differ. The second method is discriminant analysis, which examines independent measures for items in each group to determine which measures can be used to separate the groups. These identified discriminant measures are then used to define rules for classifying or predicting which group a new item will belong to given its discriminant measures. The section concludes with cluster analysis, an exploratory method used to create groups. Cluster analysis examines characteristics of individual items and creates groups such that items in each group are highly similar to each other and highly different from items in other groups.

***Comparing Two Groups***   The tests presented here are used to determine if the variable of interest differs between two groups. The appropriate test to use depends on three factors. First, are the two groups dependent or independent? Referring to the experimental design concepts in Chapter 11, within-subjects (repeated-measures) designs are dependent because each subject receives each experimental condition. This means that the values for each condition are related since they come from the same participant. In contrast, in between-subjects designs different sets of randomly selected participants receive each experimental condition and are assumed to be independent of each other. The second factor in selecting the appropriate test is whether or not the data adequately meet the assumptions of the test, as described in the "Assumptions" section for each test. The third factor is measurement type of the dependent measure. Table 2 identifies these factors and the appropriate test for each. Note that since the McNemar test is not used frequently compared to the other tests, it is not discussed in detail in this chapter. However, more information about this test may be found in Sheskin (2007) or Sprent and Smeeton (2007).

***t*-Test**   The *t*-test is a parametric test used to compare means of a dependent measure for two groups. There are two forms of the test: (1) the *dependent* or *paired t-test*, used when the same subjects received both experimental conditions, and (2) the *independent t-test*, used when different subjects are assigned to each experimental condition. Both tests are used to examine the question: Do the means of the dependent measure differ in the two populations represented by the groups? If the difference between the groups is small, it may only be caused by chance variation in the data, but if the difference is large, we may conclude that there is, indeed, a significant difference between the two groups. The primary difference in the two tests is how they examine the values.

**Table 2  Tests for Comparing Two Groups**

|  | Parametric | | |
|---|---|---|---|
|  | Interval/ Ratio | Ordinal/ Rank | Categorical |
| Dependent groups | Dependent (paired) *t*-test | Wilcoxon signed-rank test | McNemar test |
| Independent groups | Independent *t*-test | Mann–Whitney *U*-test | Chi-squared test |

*Assumptions* The assumptions of both forms of the *t*-test are that (1) the dependent variable of each underlying population is normally distributed and (2) the two populations have equal population variances (i.e., homogeneity of variance). The independent *t*-test also assumes that the each group is selected randomly from the population it represents (i.e., the two groups are independent of each other). Note that the assumption of homogeneity of variance can be tested using Hartley's $F_{max}$-test for homogeneity of variance.

*Methods and Results* Both forms of the *t*-test calculate a test (*t*) statistic that is used to determine the *p*-value. Conceptually, the *t*-statistic is the ratio of the difference in the means to the standard error of the difference. The difference in the two tests is that the independent *t*-test examines the means of each group [$\text{mean}(Y_1) - \text{mean}(Y_2)$], while the dependent *t*-test examines the mean difference of the pairs of results for each participant [$\text{mean}(Y_{i1} - Y_{i2})$, where $i$ = participant number]. When completing a *t*-test using a statistical package, the results typically provide the *t*-statistic and the associated *p*-value. The *p*-value indicates the likelihood that the difference in the means occurred by chance. Therefore, the smaller the *p*-value is, the greater evidence the test provides that the two groups are indeed different.

*Drawing Conclusions* In human factors literature, *p*-values less than or equal to 0.05 are usually considered significant. Therefore, if the *p*-value is 0.05 or less, the researcher can conclude that evidence indicates that there is a significant difference in the dependent variable between the two groups (experimental conditions). However the *p*-value does not indicate the size of the difference between the groups; therefore, it is recommended that the researcher also consider and report the size of the difference when drawing conclusions. This can be done by calculating a 95% CI on the difference between the population means. Again, most statistical packages calculate the CI, so it is quite easy to include this additional information in the results. This helps ensure that the results have practical as well as statistical significance.
*References for further details*: Sheskin (2007), Lomax (2007).

### Wilcoxon–Mann–Whitney Test
The Wilcoxon–Mann–Whitney test is a nonparametric test used to examine differences between groups on ordinal/rank data where the two groups are independent. This test is approximately the same as the Mann–Whitney *U*-test, which yields comparable results using slightly different computations. This test examines the question: Do the two groups have different median values (i.e., do the mean ranks for the two groups differ)? If the underlying population for the two groups is the same, we would expect both groups to have a similar distribution of ranks from low to middle to high. Conversely, if the underlying populations are different, we would expect one group to be concentrated in the low ranks and the other in the high ranks.

*Assumptions* This test assumes that each group is selected randomly from the population it represents (i.e., the two groups are independent of each other). It also assumes that the two groups come from populations with similarly shaped distributions, although no assumptions are made about what that shape is.

*Methods and Results* In this test, the results from the two groups are combined, the data are sorted from least to greatest, and the data are ranked overall. Then the sum of the ranks for each group is calculated. Most statistical packages calculate automatically the *p*-value by comparing the sum of the ranks for the group with the smallest sum of ranks to the relevant critical value. If the *p*-value is not provided, it can be determined by comparing the smaller sum of ranks to critical values, typically provided in the appendix of statistics books.

*Drawing Conclusions* As with the *t*-test, the *p*-value indicates the likelihood that the observed differences in rank sums are from chance as opposed to being due to differences between the underlying populations for each group. A significant *p*-value (e.g., $p \leq 0.05$) provides evidence that the median of one group is lower than the median of the other group. In other words, the group with the lower rank sum has lower values of the dependent variable than those of the other group.
*References for further details:* Sheskin (2007), Argyrous (2000), Sprent and Smeeton (2007).

### Wilcoxon Matched-Pairs Signed-Rank Test
The Wilcoxon matched-pairs signed-rank test is a nonparametric test used to examine differences between groups when the two groups are *dependent*. This is in contrast to the Wilcoxon–Mann–Whitney test, which is used to examine two *independent* groups. The Wilcoxon matched-pairs signed-rank test is the nonparametric equivalent of the dependent (paired) *t*-test. The test is based on the ranks of the differences between scores for each participant. The test examines the question: Is the median value for the difference between the two scores equal to zero (i.e., does the experimental condition cause a difference in the scores)?

*Assumptions* This test assumes that participants are selected randomly from the population represented and that each subject receives both experimental conditions. It also assumes that original scores for each participant are ordinal/ratio data and that the distribution of scores for each experimental condition comes from populations with similarly shaped distributions, although no assumptions are made about what that shape is.

*Methods and Results* In this test, the ranks are generated based on the difference between the ordinal/ratio score obtained in each experimental condition for each participant. Therefore, the original ordinal/ratio score is needed. First the difference for each participant is calculated ($Y_{1i} - Y_{2i}$). Then the absolute value of the difference scores are ranked in order from least to greatest. These difference ranks are split into two groups: the positive ranks, which are those ranks where the difference score was positive ($Y_{1i} - Y_{2i} > 0$), and the

negative ranks, where the difference score was negative ($Y1_i - Y2_i < 0$). Next the sum of the positive ranks is calculated and compared to the sum of the negative ranks. The smaller of the two rank sums is used as the Wilcoxon's $T$-statistic, which is compared to the expected value of the rank sum if there was no difference between the groups. When using a statistical package to complete this test, the researcher is shielded from these calculations. The package usually calculates and reports the positive and negative rank sums, Wilcoxon's $T$-statistic, and the $p$-value.

*Drawing Conclusion* As with the Wilcoxon–Mann–Whitney test, the $p$-value indicates the likelihood that the difference observed between the smallest rank sum observed (Wilcoxon's $T$) and the rank sum expected are by chance, as opposed to being due to differences between the scores in each group. A significant $p$-value (e.g., $p \leq 0.05$) provides evidence that the median difference in scores is not zero. To determine whether experimental condition 1 or 2 resulted in higher differences, we must examine the positive and negative rank sums. If the difference scores were calculated as $Y_{1i} - Y_{2i}$, a larger positive rank sum indicates that condition 1 resulted in higher scores than condition 2. This is logical since a larger positive rank sum indicates that more participants had a positive difference in scores ($Y_{1i} - Y_{2i}$) and/or that the absolute value of the positive differences was larger than those observed for participants with negative differences.

*References for further details:* Sheskin (2007), Argyrous (2000), Sprent and Smeeton (2007).

## Chi-Squared Test

The chi-squared test is a nonparametric test used to examine differences between groups using nominal/categorical data. For example, it could be used to determine if there were significant differences in the number of males and females in two experimental groups. This test can be used to examine two or more groups.

## Comparing More Than Two Groups

The tests presented here are used to determine if the dependent measure differs between more than two groups of interest. Similar to the case in which two groups are being examined, the appropriate test to use depends on the answers to four questions: (1) Are the groups dependent or independent? (2) Do the data adequately meet the assumptions of the parametric test? (3) What is the measurement type of the dependent measure? (4) Are multiple, correlated measures being compared across groups? Table 3 identifies these factors and the appropriate test for each. Note that since the Cochran $Q$-test is not used frequently compared to the other tests, it is not discussed in detail in this chapter. However, more information about this test may be found in Sheskin (2007) or Sprent and Smeeton (2007).

## ANOVA Test

The ANOVA test is used to determine if two or more independent groups differ on an interval/ratio-dependent measure. This test answers the question: Is there a difference in the mean for at least two of the groups? ANOVA is closely related to the

**Table 3  Tests for Comparing More Than Two Groups**

| | Parametric: Interval/Ratio | Nonparametric Ordinal/Rank | Nonparametric Categorical |
|---|---|---|---|
| Dependent groups | Repeated-measures ANOVA test | Kruskal–Wallis test | Cochran $Q$-test |
| Independent groups | ANOVA test | Friedman test | Chi-squared test |
| Multiple, correlated measures | Multiple ANOVA test | | |

$t$-test and is preferred for examining more than two groups because it controls the experimentwise error rate. A variety of ANOVA procedures exist to support analyzing a variety of experimental designs, such as those with two grouping factors or using a mixed design. For simplicity, we address only one-factor ANOVA here since the concepts used in this procedure extend to more advanced ANOVA procedures.

*Assumptions* ANOVA assumes that the data being analyzed comprise a randomly selected sample. In the model on which the ANOVA is based, the error terms are normally distributed with mean zero. The variances of the data in each group are approximately equal (homogeneity of variance). Note that ANOVA is based on the general linear model, so these assumptions are the same as many of the assumptions for linear regression. See "Linear Regression" in Section 4.2.3 for more details on testing these assumptions.

*Methods and Results* ANOVA determines if the mean is equal for all the groups. To compare the groups, an ANOVA table is constructed which breaks down the sources of the variation in the dependent measure. For the one-factor ANOVA, there are two possible sources of variation: the independent measure used for grouping or random variation. The ratio of the variation accounted for by the independent measure [mean square treatment (MSTr)] to the random variation [mean square error (MSE)] is the $F$-statistic. The $F$-statistic is compared to a threshold value to determine the likelihood ($p$-value) that the differences in the means of the groups are due to random variation or actual differences in the mean of the underlying population. Most statistical packages report the full ANOVA table, including the breakdown of the variation (in terms of sum of squares, degrees of freedom, and the mean square for each source of variation), the $F$-statistic, and the $p$-value.

Results from the ANOVA $F$-test indicate only whether or not there is a difference between at least two of the groups. To determine which groups are actually different, post hoc tests, or paired comparisons, must also be performed. However, completing these multiple comparisons increases the experimentwise error rate ($\alpha$). Several post hoc comparison methods exist that are designed to control the experimentwise error rate. Two

commonly used methods are the Bonferroni and the Tukey. In both of these methods, the $t$-test is used to compare the means for each pair of groups. To control the error rate, the critical value to which the $t$-statistic is compared is adjusted to be more stringent and to account for the number of comparisons being made. Most statistical packages allow you to choose which post hoc method to use. The Bonferroni method is more conservative—it is more robust than other methods but as a result has less power to detect differences. The Tukey method is more sensitive but is not as robust when departures from test assumptions occur. When using statistical packages to complete paired comparisons, of interest in the output results are the mean difference between the groups, the $p$-value, and *the 95% CI on the mean difference.*

Because ANOVA analysis is based on the general linear model, it is important to complete a residual analysis as part of this procedure to ensure that the data meet all of the assumptions of the model. See "Linear Regression" in Section 4.2.3 for more details on how to validate these assumptions.

*Drawing Conclusions* To draw conclusions, first look at the $p$-value for the ANOVA $F$-test. If the $p$-value is sufficiently low, we may conclude that there is a difference among the groups and examine the results of the post hoc tests to determine which groups differ and the direction of that difference. When examining post hoc test results, first look at the $p$-values for each pair to identify which groups have means that differ significantly. Once these have been identified, examine the 95% CI on the mean difference to determine the direction of the difference based on the sign of the mean [e.g., positive (A – B) indicates group A > group B] and the magnitude of the difference.

*References for further details*: Kutner et al. (2004), Sheskin (2007), Argyrous (2000), Wu and Hamada (2009).

**Repeated-Measures ANOVA Test** The repeated-measures, or within-subjects, ANOVA test is used to determine if two or more dependent groups differ on an interval/ratio-dependent variable of interest. This test is used to analyze data from repeated-measures and blocked experimental designs where each participant receives each experimental condition. This test answers the question: Is there a difference in the mean for at least two of the experimental conditions? The repeated-measures ANOVA is related conceptually to the dependent $t$-test and is closely related to the ANOVA used to examine multiple, independent groups.

*Assumptions* This test assumes that the data are a randomly selected sample. In the ANOVA model on which the test is based, the error terms are normally distributed with mean zero. In contrast to the ANOVA test, the repeated-measures ANOVA assumes sphericity instead of homogeneity of variance. Sphericity is a more complex concept related to the underlying variance and covariance of the populations being examined. For more information on this assumption, see Sheskin (2007).

*Methods and Results* Like ANOVA, this test examines whether the means for all groups are equal. The primary difference in the repeated-measures ANOVA is that it acknowledges and accounts for the variation between participants, in effect comparing each participant's performance across groups. To do this, this method decomposes the variation in the dependent variable into three possible sources of variation: (1) the independent measure used for grouping, (2) the participant completing the trial, or (3) random variation. In this case, two $F$-statistics may be calculated, one for the grouping measure (as in ANOVA) and one for the blocking variable, the participant. As with ANOVA, most statistical packages report the full ANOVA table, including the breakdown of the variation (in terms of sum of squares, degrees of freedom, and the mean square for each source of variation), the $F$-statistics, and the $p$-value. Upon completing the repeated-measures ANOVA, a post hoc test should be completed using the methods described in the section on ANOVA to examine the differences between groups in more detail.

*Drawing Conclusions* Drawing conclusions for repeated-measures ANOVA is essentially the same as that for ANOVA, so refer to that section for more detail. The primary difference is that, in addition to drawing conclusions about differences in the dependent measure related to the grouping measure, you can also draw conclusions about whether or not the participants varied significantly on the dependent measure. If there are differences among participants, further investigation may be warranted to try to determine if another characteristic of the participant is at the root of these differences.

*References for further details:* Kutner et al. (2004), Sheskin (2007), Wu and Hamada (2009), Lomax (2007).

**Effects of Measurement Errors on ANOVA** Liu and Salvendy (2009) recently demonstrated the effects of measurement errors on psychometric measurements in ergonomics studies. For ANOVA, they highlight the possibility of overestimating the error sum of squares. This can occur in the commonly used fixed-effect, single-factor balanced design, single-factor for repeated-measures, and factorial ANOVA. Liu and Salvendy (2009) suggest that these measurement errors reduce the statistical power of hypothesis testing; in other words they reduce the probability of getting a statistically significant result for a real effect in a population.

*References for further details*: Liu and Salvendy (2009).

**Multivariate ANOVA (MANOVA) Test** MANOVA is conceptually similar to ANOVA, except that MANOVA examines differences among groups on a set of correlated dependent measures. MANOVA takes steps to manage the experimentwise error by accounting not only for the number of groups being compared, as in ANOVA, but also for the number of dependent measures being examined. If the MANOVA test is significant, it indicates that there is a significant difference among the groups on at least one of the measures. Once this conclusion is made, use ANOVAs and the appropriate post hoc tests to analyze each of the dependent measures

individually to determine which dependent measures vary and between which groups those measures vary. Due to the complexity of this test, the reader is referred to the references for further details for additional information regarding specific assumptions, methods, and results of MANOVA.

*References for further details:* Johnson and Wichern (2007), Johnson (1998), Field (2005).

**Kruskal–Wallis Test**    The Kruskal–Wallis test is the nonparametric equivalent of ANOVA. It is appropriate for ordinal data and examines differences in ranks between independent groups. It is an extension of the Wilcoxon–Mann–Whitney test and is used to analyze more than two groups. As in the Wilcoxon–Mann–Whitney test, this test examines the question: Do the mean ranks for the groups differ? If there are true differences between the groups, we would expect one or more groups to be concentrated in the low ranks and the others to be concentrated in the high ranks.

*Assumptions*    Kruskal–Wallis assumes that the groups represent a random sample and are independent of each other. It also assumes that the groups come from populations with similarly shaped distributions, although no assumptions are made about what that shape is.

*Methods and Results*    As in the Wilcoxon–Mann–Whitney test, the data from all groups are combined and sorted from least to greatest, and the data are ranked overall. Then the sum of the ranks for each group is calculated. The test statistics, $H$, is calculated based on these sums. Statistical packages calculate the $H$-statistic and associated $p$-value by comparing $H$ to the relevant critical value. As with the ANOVA test, a significant $p$-value for this test only indicates that two or more groups vary in their median rank. To determine which groups differ, pairwise comparisons must be completed. Kruskal–Wallis uses the Wilcoxon–Mann–Whitney test to compare each pair of groups. To control experimentwise error $(\alpha)$, the Bonferroni method is typically used to adjust $\alpha$. For this method, divide the target $\alpha$ by the number of comparisons $(j)$. In other words, if your target is $\alpha = 0.05$ and you are completing four paired comparisons, the target $\alpha'$ for each comparison is $0.05/4 = 0.0125$. This means that $p$-values greater than 0.0125 would be rejected in the paired comparisons.

*Drawing Conclusions*    If the $p$-value for the Kruskal–Wallis test is sufficiently low, we conclude that there is a difference among the groups and examine the results of the post hoc tests to determine which groups differ. In the post hoc tests, if any pairs differ significantly (i.e., they have a $p$-value below the Bonferroni revised threshold,$\alpha'$), we may conclude that those pairs of groups differ. To determine the directionality of the difference, compare the sum of the ranks for each group. The group with the higher rank sum can be inferred to have higher values of the dependent measure than those of the other group.

*References for further details:* Sheskin (2007), Lomax (2007), Sprent and Smeeton (2007).

**Friedman Test**    The Friedman test is the nonparametric equivalent of the repeated-measures ANOVA. Like the Kruskal–Wallis test, the Friedman test uses ordinal data to examine whether the mean ranks differ for the groups. However, the Friedman test is used for dependent groups, such as those found in repeated-measures designs.

*Assumptions*    This test assumes that the data being analyzed comprise a randomly selected sample.

*Methods and Results*    To analyze differences among the groups, the Friedman test examines the rank of each group within a participant's results. This is in contrast to the Kruskal–Wallis test, which assigns an overall rank to pooled scores for all groups. Therefore, the Friedman test results in a set of rankings for each participant. After the ranks are assigned, the ranks for each group are summed. These rank sums are then used to calculate the Friedman test statistic $(\chi_r^2)$. When this test is run using statistical packages, the test statistic and the associated $p$-value are typically reported. As in the previous tests examining multiple groups, if the Friedman test is significant, post hoc comparison tests are required to determine which pairs of groups differ. One method of accomplishing this is to use the Wilcoxon matched-pairs signed-rank test to compare each pair of groups. To control the experimentwise error $(\alpha)$, the Bonferroni method is typically used to adjust $\alpha(\alpha' = \alpha/\text{number of tests})$, as described in the Kruskal–Wallis test.

*Drawing Conclusions*    The first step in drawing conclusions is to examine the $p$-value of the Friedman test. If $p$ is small (e.g., $p < 0.05$), we may conclude that there is a difference among the groups. If there is a difference in the groups, examine the post hoc test results to determine which groups differ. In the paired comparisons, identify any pairs that differ significantly (i.e., they have a $p$-value less than the Bonferroni revised threshold, $\alpha'$). For the pairs that differ, compare the sum of the ranks for each group in the pair to determine which group has higher values. The group with the higher rank sum can be inferred to have higher values of the dependent measure than those of the other group.

*References for further details:* Sheskin (2007), Lomax (2007), Sprent and Smeeton (2007).

**Chi-Squared Test**    The chi-squared $(\chi^2)$ test is a nonparametric test used to examine differences between groups on a nominal/categorical measure. The test is typically employed in association with a contingency or crosstab table. A contingency table is an $r \times c$ table where there is a row for each of the $r$ groups and a column for each of the $c$ possible values of the nominal measure being examined. Each cell represents the frequency with which the given row–column pair occurred. The chi-squared test examines whether or not there is a relationship between the group and the nominal measure. For example, is the number of females the same for all groups, or do a disproportionate proportion of women fall into a particular group?

*Assumptions* This test assumes that categorical data are used and that the $r \times c$ categories are mutually exclusive. That is, a participant/object will occur in only one cell in the contingency table. It also assumes that the data are from a random sample. An additional assumption is that the expected frequency of each cell is at least 1 and that no more than 20% of the cells have an expected value of 5 or less.

*Methods and Results* The chi-squared test compares the expected frequency of a category to the frequency of that category observed for each of the $r \times c$ categories in the table. The expected frequency is calculated based on the assumption that there is no relationship between the grouping measure and the nominal measure, so it is based on the proportion of overall participants who fall into that particular group. So in our example of examining gender differences among the groups, if you have 80 participants who are evenly divided among four experimental groups, we would expect each group to be 50% female, which translates to an expected frequency of 10 females [20 participants per group $\times$ 0.5 (expected percentage of females)]. The $\chi^2$-statistic is calculated based on the difference between the frequency observed and the frequency expected. When using a statistical package to complete this test, results typically include the contingency table with observed and expected frequencies (or proportions), the $\chi^2$-statistic, degrees of freedom, and the $p$-value. *Note:* The degrees of freedom are calculated as $(r-1)(c-1)$.

*Drawing Conclusions* Again, we begin drawing conclusions by examining the $p$-value. If $p$ is sufficiently small, evidence indicates that there is a significant difference between the expected and observed frequencies in the cells. This in turn indicates that there is a relationship between the grouping measure and the nominal measure. However, $p$ does not provide an indication of how they are related. Which cells differ from the expected frequencies? To determine this, most sources recommend completing paired comparisons in the form of $2 \times 2$ contingency tables for the subsets of the larger $r \times c$ table. As mentioned previously, these multiple comparisons can inflate the experimentwise error rate, $\alpha$. To control the error rate in the chi-squared test, we again use the Bonferroni method, which divides the target $\alpha$ by the number of comparisons ($j$).

Now we know which cells differ but still cannot draw any conclusions regarding the magnitude of the differences, which indicate the strength of the relationship. There are a number of measures of association or correlation that indicate the size of the effect that can be examined. These include the phi coefficient, for $2 \times 2$ tables only, and Cramer's phi, for tables larger than $2 \times 2$. Phi coefficients range from $-1$ to 1, with the absolute value indicating the strength of the effect. Cramer's phi ranges from 0 to 1 with a similar interpretation. Cohen (1988) suggests the following recommendations for interpreting the size of an effect:

- *Small effect:* between 0.10 and 0.30
- *Medium effect:* between 0.30 and 0.50
- *Large effect:* greater than or equal to 0.50

*References for further details:* Sheskin (2007), Argyrous (2000), Lomax (2007).

***Characterizing and Creating Groups*** In some cases we are interested in understanding the characteristics of known groups or in using the characteristics of items to create new groups. Several approaches are available for accomplishing either of these goals. To characterize a group based on a number of correlated measures, profile analyses are appropriate. On the other hand, discriminant analysis can be used to examine groups on a number of independent measures to determine which measures most distinguish groups from each other. These discriminating characteristics can then be used to classify or predict which group new items might fall into. When there are only two groups, logistic regression is sometimes used to predict classification of items into the two groups. Sometimes, however, we are simply interested in creating groups of items that are similar to each other on a number of characteristics. In this case, cluster analysis proves useful. In the following section we describe profile analyses, discriminant analysis, and cluster analysis. Logistic regression is described in Section 4.2.3.

**Profile Analyses** Profile analyses are used to examine the means of several measures for two or more groups. The measures are typically correlated. For example, they may be the results of several tests given to a participant. The profile for each group is a plot of the mean of each measure (see Figure 7). Profile analyses attempt to answer the following questions: Are the profiles for two groups parallel (e.g., is the difference in the means the same for all measures)? If they are parallel, are they coincident (e.g., is the mean for both groups the same)? Do the profiles show any trends?

*Assumptions* Profile analyses assume that each group's measures are independent of those of the other groups and normally distributed. It also assumes that the measures being used to create the profile use the same scale/units.

*Methods and Results* Profile analyses use a number of statistical tests, based on those discussed previously, to compare the difference in the mean values of the two groups for each measure. The primary difference between the two is that the statistical methods used in profile analysis take into account the number of measures being examined and their covariance, similar to MANOVA. To determine whether or not two profiles are parallel, the $T^2$-statistic is used to test if the difference between two groups in the means for each measure is the same.

*Drawing Conclusions* Looking at the profiles in Figure 7, we would intuitively suspect that the profiles for groups 1 and 2 could be parallel. If the $T^2$-statistics for these groups produced a sufficiently small $p$-value, we could conclude that this is indeed the case.

*Reference for further details:* Johnson and Wichern (2007).
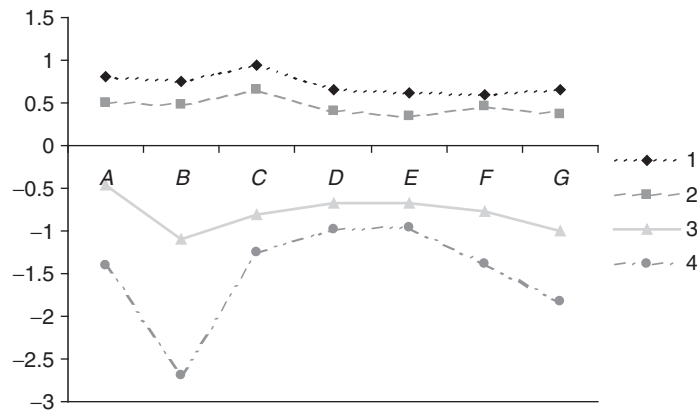
**Figure 7**  Profile plot for four groups.

**Discriminant Analysis**  Discriminant analysis is used to explore features of items (e.g., participants, events) in existing groups to determine which features differentiate items in one group from the other group(s). In many cases, these discriminant features, represented by measures, can be used to develop rules for classifying new items into one of the existing groups. For example, a researcher who was examining an assembly task might group participants in a study into two groups: those who completed the task successfully and those who did not. This researcher could use discriminant analysis to determine which participant and environmental characteristics differentiate those who completed the task successfully from those who did not. These discriminant features could, in turn, be used to develop rules for predicting whether or not the task will be completed successfully in certain situations. Discriminant analysis develops rules used to predict which group a given item will belong to given a set of measures describing that item. In general, the goals of discriminant analysis are (1) to identify and describe the features most distinctly and separate items in known groups and (2) to derive rules used to optimally sort or classify new items into one of the given groups. In this section we examine the simple case of discriminating between two groups, although the method can be used for more than two groups.

*Assumptions*  Discriminant analysis assumes that the populations of the groups being examined have a multivariate normal distribution with equal covariance.

*Methods and Results*  The first step in discriminant analysis is to define a set of discriminant rules based on the means and variance–covariance of the measures for the two groups being examined. Although in many cases they result in the same rules, four different methods may be used to develop these rules: the likelihood rule, the linear discriminant function rule (Fisher's method), the Mahalanobis distance rule, and the posterior probability rule. The linear discriminant function rule (Fisher's method) is the best known. This method

finds a linear transformation ($Y$) of the measures ($X$'s) such that the $Y$'s of the two groups are separated as much as possible. The discriminant rule compares the $Y$ of the new item to the midpoint between the two groups. If the value is greater than the midpoint, it is assigned to the group with larger values of $Y$, and vice versa.

Once the discriminant rules are established, it is important to assess the adequacy of these rules by estimating the probability of classifying an item correctly. As researchers, we want to choose the rule that classifies items correctly a sufficiently high portion of the time. After all, a rule that only works 50% of the time is not very useful; we might as well flip a coin to assign an item to a group. Of the three methods that may be used, the simplest is to use resubstitution estimates. For this method, create predicted classifications for the original data based on the discriminant rules and compare how frequently the prediction is correct. The method is problematic because it tends to overestimate the probability of correct classification. For large data sets a second method of assessing adequacy of the rules is to use holdout data. In this method, a subset of the data available is withheld when the discriminant rule is created. The discriminant rule is used to predict classifications for the holdout group; then the percentage of correct classifications is examined. For smaller data sets, this method should not be used since all of the data are required to create the best possible discriminant rule. The most accurate method is known as cross-validation, jackknifing, or Lachenbruch's holdout method. This method is more complex than the other two, so refer to one of the references for further details on how to complete this method of estimating the likelihood of a correct classification.

*Drawing Conclusions*  The first step in drawing conclusions is to examine the probability of classifying an item correctly. If this percentage is sufficiently high considering the type of data and the purpose for which the discriminant analysis is intended, we may conclude that the discriminant rule is an adequate predictor for classification. Next, examine the coefficient/weight of

each measure in the discriminant rule. The measures with higher coefficients factor more heavily in attempting to separate the two groups; therefore, those measures are better for discriminating between groups.

*References for further details*: Johnson and Wichern (2007), Johnson (1998).

**Cluster Analysis** Clustering is an exploratory method used to examine a set of ungrouped items (e.g., participants, events) and to identify any naturally occurring clusters, or groups, of items. This is in contrast to discriminant analysis, in which there were preexisting groups and the goal was to define rules for classifying new items into those preexisting groups. In cluster analysis we start with a set of items (e.g., participants) and a number of descriptive measures related to those items (e.g., a variety of behavioral measures related to task performance). The primary goal is to identify groups of items (participants) that are highly similar to each other on the descriptive measures (behavior) but highly different from the items in the other groups. In other words, we want to minimize the distance between items within a group and to maximize the distance between groups.

Once these groups are defined, the methods identified previously for comparing groups can be used to examine differences between the groups on the descriptive measures used to create the groupings. Although this can provide interesting insights, it is often more interesting to examine the groups for differences on measures that were not used to create the groupings. For example, if behavioral measures were used to group participants, it might be interesting to examine the resulting groups for differences in demographic variables such as age, education, and experience level. Note that cluster analysis is exploratory in nature and is more of an art than a science. There are a number of arbitrary judgments that the researcher must make, such as the appropriate clustering method and number of clusters. Therefore, validating that the resulting groups are meaningful and useful is crucial to successful cluster analysis.

*Methods and Results* Cluster analysis uses one of several available algorithms to identify groups of items based on a number of similarity and/or dissimilarity measures. Before using an algorithm, always explore the data using scatterplots and other graphical representations to see if there are any apparent natural groupings. If obvious groupings are observed, they can be used to validate the number of clusters and assignment of items to those clusters after the clustering algorithm is run. They may also be helpful for selecting the appropriate clustering algorithm. There are two primary classes of algorithms for identifying clusters in a data set: nonhierarchical and hierarchical. Nonhierarchical methods begin with a given number of clusters and an initial set of seed points around which the clusters will be built. The disadvantage of these methods is that the researcher must make an initial guess at the number of groups and location of the initial seed points. If the data are already thoroughly understood, it may be possible to set these effectively, but frequently this is not the case. Therefore, hierarchical clustering algorithms are more widely recommended.

Frequently used hierarchical clustering algorithms begin with each item in a single group and merge the closest groups successively. These algorithms, known as *agglomerative hierarchical methods*, attempt to merge the most similar items into a group while keeping the groups dissimilar. When using these methods, several different linkage methods can be used to build the clusters, and these methods vary in how they measure dissimilarity, the distance between two clusters. Two frequently used methods are nearest neighbor and furthest neighbor. In nearest neighbor, at each step the two closest items/clusters are merged and the dissimilarity between two clusters is measured as the distance between the two closest members. Both methods result in a series of merged clusters, and it is up to the researcher to decide which number of clusters is appropriate. This can be done by examining a dendrogram, or hierarchical tree diagram (Figure 8). The dendrogram depicts each successive merging step through the lines used to join items (cases). The horizontal distance of the line indicates the distance between the two merged clusters. Therefore, we want to choose the number of clusters that results in the smallest horizontal lines connecting the members, indicating that the members are quite close, and the largest horizontal lines connecting the identified clusters, indicating the clusters are far apart. In Figure 8 it appears that selecting three clusters will accomplish this goal.

If a dendrogram does not present an obvious choice of groups, it is sometimes useful to try a different linkage method to see if it will produce more distinct groups. As this process demonstrates, choosing the appropriate linkage method and number of groups is a subjective judgment made by the researcher. This is the reason that cluster analysis is considered exploratory in nature. However, there are statistical tests that can also be used to validate the selected number of clusters. See the references identified at the end of this section for details on these procedures. Once the appropriate number of clusters is identified, each item is assigned to a cluster and further analysis can be performed on each cluster (group).

*Drawing Conclusions* Before drawing conclusions as to the results of a cluster analysis, we first determine if the clusters are meaningful and useful. Examining the dendrogram gives a good indication of whether or not the groups created are distinct. If they are not distinct, it is unlikely that the groups will be useful. Another way is to examine two- and three-dimensional scatterplots of the measures used to create the groups using a different symbol or color to represent items in each group. If the items form distinct visual clusters, it is an indication that the grouping is good. After confirming visually that the groups are distinct, it is also useful to use the statistical methods described previously to compare the groups on the measures used to create the groups. This provides statistical evidence that the groups indeed differ from each other on these measures. Once we are confident that the groups are meaningful, it is often interesting to compare the groups on measures that were not used to create the groups. This can sometimes give insight into complex relationships between

Rescaled distance cluster combine
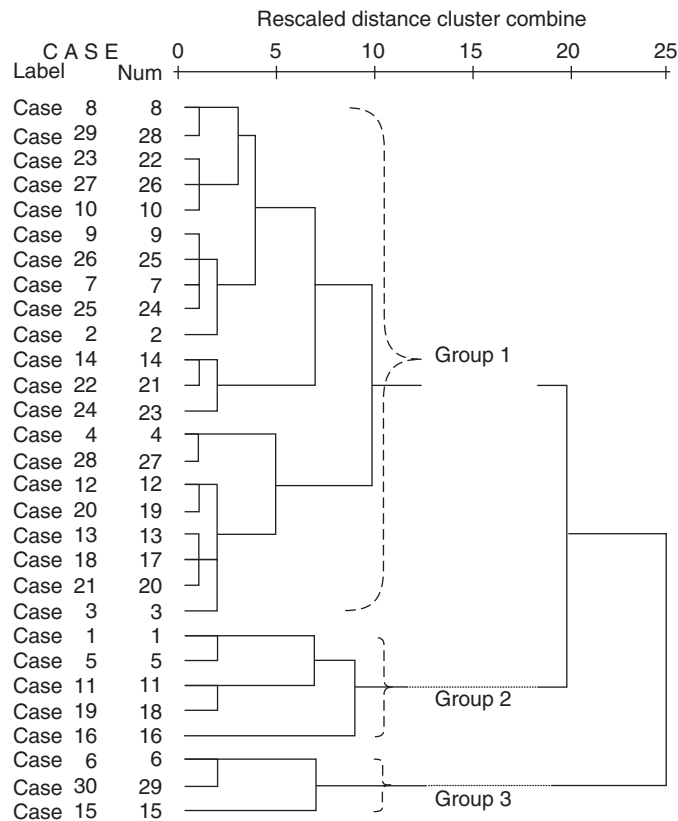


**Figure 8**   Dendrogram from cluster analysis.

measures used in cluster analysis and other measures that might lead to interesting directions for further research and analysis.

*References for further details*: Johnson and Wichern (2007), Johnson (1998).

### 4.2.3   Examining and Modeling Relationships

In previous sections we have focused on examining and in some cases creating groups. However, frequently, researchers are interested in examining relationships between certain measures, independent of specific groups. For example, a researcher might want to examine the relationship that several independent measures have with a dependent measure. Alternatively, we might want to examine relationships among independent variables to determine if some redundant variables can be excluded from future experimental designs and/or analysis. Although a number of advanced techniques for modeling performance are presented in Part 6, the simple modeling methods presented here focus on understanding and characterizing relationships among a number of measures. In fact, the insights gained through these simple modeling methods can inform the development of more complex models.

***Correlation Using Pearson and Spearman Coefficients***   Correlation is used to examine the relationship between two ordinal and/or interval/ratio measures. Correlation measures are used to examine the question: What is the degree of relationship or association between two measures? Correlation analysis is used to determine the strength of the relationship between the two variables—it does not imply causality in the relationship. Two commonly used measures of correlation are the Pearson product–moment coefficient, a parametric measure used for interval/ratio data, and the Spearman rank-order coefficient, a nonparametric measure used for ordinal data.

*Assumptions*   Both the Pearson and Spearman tests assume that the data used comprise a random sample. The Pearson coefficient additionally assumes that both variables use an interval/ratio scale and that the two variables have a bivariate normal distribution. In a bivariate normal distribution, both variables and the linear combination of the variables are normally distributed. The latter half of the assumption means that the Pearson coefficient assumes that the relationship between the two variables is linear.

*Methods and Results*   The Pearson coefficient examines the degree to which a linear relationship exists between the two measures of interest. The coefficient ($r$) ranges from −1 to 1 and the absolute value of $r$ ($|r|$) is an

indicator of the strength of the relationship. The closer $|r|$ is to 1, the stronger the relationship between the two variables. The sign of $r$ indicates the direction of the linear relationship, with positive values indicating a direct relationship and negative values indicating an inverse relationship. Here, $r^2$ represents the proportion of the variance in one measure accounted for by the other measure. When using the Pearson coefficient, create a scatterplot of the two variables to make sure that the assumption of a linear relationship holds. If a curvilinear relationship exists, $r$ will be 0 even if a relationship between the two variables does actually exist. Additionally, large sample sizes are best for this analysis. When small sample sizes are used, several factors, such as the presence of outliers and restrictions on the range of one of the variables, can distort the value of $r$. For this reason, in experiments with a small number of observations, even if $r$ is large, the $p$-value may indicate that the correlation is not significant. Conversely, in an experiment with a large sample, the $p$-value may indicate a significant correlation even though $r$ is relatively small.

In contrast to the Pearson method, Spearman's rank-order coefficient determines the degree to which a monotonic relationship exists between the two variables rather than a linear relationship. The Spearman coefficient accomplishes this by examining the relationship between two ordinal variables by analyzing the ranks of the two variables. Each participant is ranked on each variable, and the coefficient ($r_S$) is calculated based on the difference in the ranks for each participant. Once $r_S$ is calculated, the interpretation of the correlation coefficient and $p$-value are the same as the interpretation using the Pearson coefficient ($r$).

*Drawing Conclusions*   To draw conclusions, first look at the $p$-value and correlation coefficient ($r$). If $p$ is large, it does not necessarily indicate that no relationship exists since, as mentioned previously, $p$ is sensitive to the number of observations. For example, when using the Pearson coefficient, if $p$ is large, $r$ is large, and the number of observations is small, it might be advisable to collect more data. The additional data would make $r$ less vulnerable to distortions and reduce the threshold used to calculate $p$. In the case where $p$ is small enough to be declared significant, you must still examine $r$ to determine the strength of the relationship. As mentioned previously, Cohen (1988) provides suggestions for interpreting the size of a correlation effect based on $r$: small effect ($0.10 \leq r < 0.30$), medium effect ($0.30 \leq r < 0.50$), and large effect ($r \geq 0.50$). For an example of correlation analysis, see analysis 1 in the case study presented later in Figure 14.

*References for further details:* Kutner et al. (2004), Sheskin (2007), Lomax (2007).

**Linear Regression**   Linear regression is used to examine the effect that certain measures (predictors) have on a dependent measure when the dependent measure consists of interval/ratio data. Linear regression models predict the value of the dependent measure based on the values of predictors. A linear regression explores two primary questions: (1) What effect do the predictors

examined have on the dependent measure? (2) Given a set of values for the predictors, what is the value predicted for the dependent measure? For example, we might use linear regression to examine the factors that influence task completion time for a computer task. The measures used as predictors may be of any measurement type, but nominal data must be recoded to use a series of 0–1 indicators, where 1 indicates that a given category is present. For example, if job type is a predictor with three possible values (administrative, managerial, engineering), create three new variables, one for each job type, and assign a 1 to the variable representing the participant's job type and 0 to the remaining two variables. These three variables should be used in the analysis instead of the original nominal variable. There are two forms of linear regression: simple linear regression, which uses only one predictor in the model, and multiple linear regression, which includes multiple predictors.

*Assumptions*   Linear regression is based on a general linear model and therefore must meet the assumptions of that model. The first assumption is that the relationship between the predictors and the dependent measure is linear. This assumption can be validated by examining a scatterplot of the data and ensuring that the plot resembles a straight line. If the scatterplot between a predictor and the dependent measure is curvilinear and resembles an upright or inverted U, linear regression may still be used if both linear and quadratic components ($X$ and $X^2$) are included in the model. Additionally, multiple linear regression assumes that the predictors are *not* highly correlated. When two or more predictors are correlated, multicollinearity exists and this causes the regression model to be imprecise. There are several tests for multicollinearity; see Kutner et al. (2004) for more details. When multicollinearity exists, drop redundant predictors from the model or use the data reduction techniques discussed in Section 4.3 to develop a set of independent predictors based on the predictors correlated.

The next set of assumptions is related to the error terms, or residuals, of the model. The residuals are calculated by subtracting the observed value of the dependent measure from the value predicted. The linear model assumes that residuals are independent and randomly distributed. It also assumes that at each value of a predictor ($X$), the error terms are normally distributed with a mean of zero and, for all values of $X$, the variance of the error terms is the same (homogeneity of variance). To check several of these assumptions, examine scatterplots of the residuals. If there is a pattern to the residuals, it indicates that they are not random. This usually means that the model is a bad fit. This can be caused if predictors are missing from the model or if the underlying relationship between the dependent variable and one or more of the variables is not linear. If the distribution of points is not centered around zero, it indicates that the mean of the errors is not zero and that the model is consistently overestimating (mean $< 0$) or underestimating (mean $> 0$) the value predicted. If the scatterplot is funnel shaped, as illustrated in Figure 9, it indicates that the error variances change with the value of $X$, and
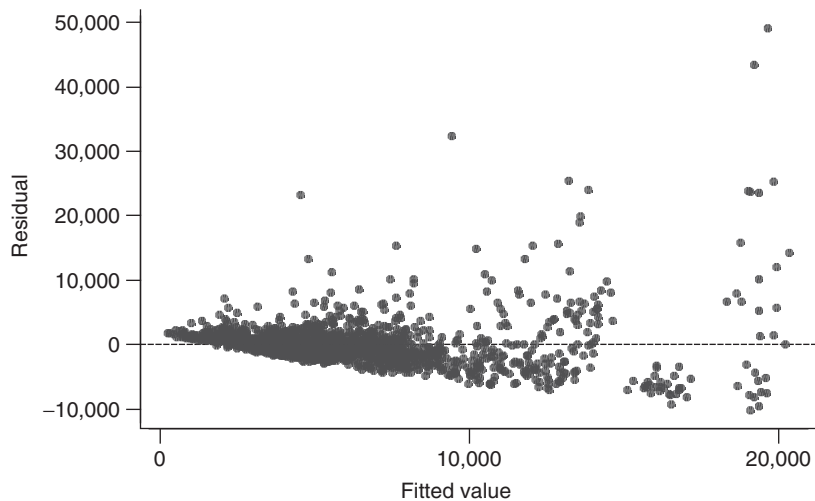
**Figure 9** Residual plot where the homogeneity of the variance assumption is violated.
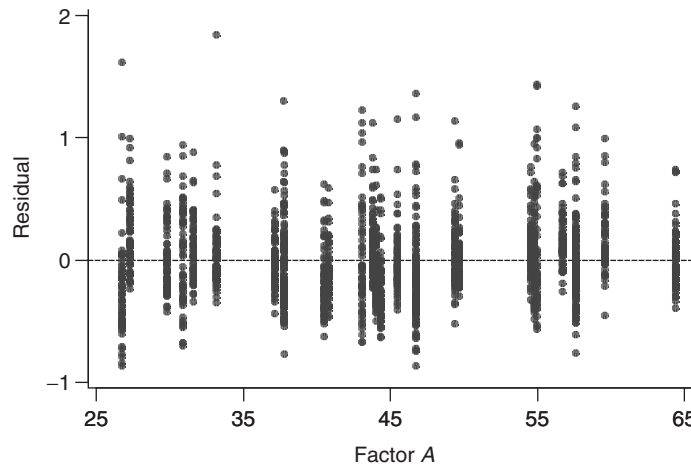


**Figure 10** Good residual plot that meets the error distribution assumptions.

the homogeneity of the variance assumption is violated. In contrast, Figure 10 depicts a good residual plot that supports the assumptions that errors are randomly distributed with mean zero and with homogeneous variance. When the homogeneity of the variance assumption is violated, the violation can often be resolved by using a transformation of the dependent variable, as discussed in "Parametric and Nonparametric Tests" in Section 4.1.1.

To test the assumption that the error terms are normally distributed, create a normal probability plot of the residuals. If the points on this plot are in a straight line, the distribution of residuals is close to a normal distribution, as illustrated in Figure 11. When running the analysis, make sure that the option to create these plots is selected so that the statistical package will generate the residual graphs automatically as part of the linear regression analysis.
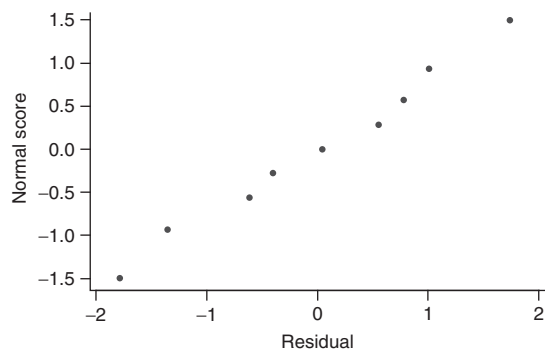


**Figure 11** Normal probability plot of residuals when error terms are normally distributed.

*Methods and Results*   Linear regression attempts to model the relationship between the dependent measure ($Y$) and a number of predictor measures or factors ($X_i$). Multiple regression generates a model in the following format: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i$. In linear regression, the goal is to develop a parsimonious model with sufficient explanatory power. In other words, we want to gain the strongest match between the model and the data with the least number of predictors. A model with fewer predictors is preferred because it is simpler to interpret and use. In general, the Pareto principle applies to model building—a small number of factors account for a large portion of the variability. Therefore, we can often reduce the size of the model without sacrificing a significant portion of its explanatory power.

Several methods may be used to build a model from the predictors, including forward elimination, backward elimination, stepwise selection, and best subsets. Stepwise selection and best subsets are generally the preferred methods. In stepwise selection the predictor that contributes most to the explanatory model is added to the model. This continues, and as additional predictors are added, the predictors added previously are checked to make sure that they are still significant given the addition of the new predictor. The process ends when all of the predictors in the model are significant and all predictors excluded from the model are not significant. In the best subset method of model selection, all possible model combinations are examined and compared using Mallow's $C_p$-statistic, which indicates the explanatory power of the model adjusted to account for the number of variables. This adjustment factor ensures that more parsimonious models will be favored over larger models when the explanatory power of both models is similar. When comparing $C_p$ statistics, choose a model with a $C_p$-value that is low and close to 1 plus the number of predictors.

After the model has been selected, run the linear regression for the model. The results of the regression include the coefficient and $p$-value for each predictor and an $r^2$ and adjusted $r^2$ for the model. These two values are indicators of the proportion (0–100%) of the variation in the data that is explained by the model, so large values of $r^2$ are desirable. The adjusted $r^2$ adjusts the $r^2$-value to penalize models with a larger number of variables; therefore, it is the preferred measure of model fit.

*Drawing Conclusions*   After selecting a model and running the linear regression, first examine the $p$-values for each predictor included in the model to ensure that all predictors are statistically significant. Look at the $r^2$- and adjusted $r^2$-values to ensure that the model's explanatory power is sufficient for the intended use of the model. "Sufficient" will depend on the intended use. For example, in a situation where the goal is to tightly control the dependent measure within a certain range of tolerances, a higher explanatory power may be required. Next, examine the residual plots to ensure that none of the assumptions have been violated. Also, in cases where the explanatory power of the model is low, examining plots of the residuals versus the possible predictor variables may provide insight into how to improve the model. A pattern in one of these plots might indicate that the linear model is not adequately capturing an underlying relationship. For example, if the residuals make a U-shaped pattern, including a quadratic component ($X_2^1$) in the model might improve the fit.

After validating that the model meets the necessary assumptions and has sufficient explanatory power, the model can be interpreted and applied. The coefficient for each predictor indicates the size and direction of the effect that a predictor has on the dependent measure. Be careful when comparing the coefficients of several predictors to understand their relative contribution to the dependent measure. Differences in the scales used to measure those factors can skew the comparison. For example, if predictors A and B are in the model of dependent measure C and the range of values for these measures are 1–5, 100–1000, and 0–1000, respectively, A's coefficient may be larger than B's simply because of the difference in scale. To make comparisons of the size of the effect of several factors when different scales are used, it may be useful to run the regression using standardized values of the predictors. However, interpretation of the coefficient in terms of the effect of an incremental change in a given factor can be more complex and less generalizable (Lomax, 2007), since the standardized value is based on the variance of the specific sample.

Note that when drawing conclusions about the effects of predictors, it is best to limit them to the range of values examined for each predictor. It can be dangerous to extrapolate these trends to values far outside the range observed in the experiment since the underlying relationship between the predictor and the dependent measure may differ when the predictor is at levels extremely different from those examined.

*Effects of Measurement Errors on Linear Regression*   Measurement errors can also negatively impact ergonomics studies utilizing both simple linear regression and multiple linear regression. Remember from the previous section on linear regression assumptions that it assumes that at each value of a predictor ($X$) the error terms are normally distributed with a mean of zero and, for all values of $X$, there is homogeneity of variance. Liu and Salvendy (2009) point out that when $X$ is measured with errors, it causes the slope coefficient to be biased toward zero. For further information on measurement errors in linear regression see Liu and Salvendy (2009), Carroll and Stefanski (1995), and Fuller (1987).

*References for further details:* Kutner et al. (2004), Wu and Hamada (2009), Lomax (2007).

**Logistic Regression**   A logistic regression is used to examine the effect that certain measures (predictors) have on a dependent measure when the dependent measure is nominal/categorical. Frequently, the dependent measure is binary (e.g., yes/no, true/false, 0/1). Conceptually, logistic regression is similar to linear regression except that, instead of modeling and predicting the value of the dependent measure, logistic regression models

the likelihood that the dependent measure will have a certain value (e.g., yes). Logistic regression explores two primary questions: (1) What effect do the predictors have on the probability that the event represented by the dependent variable will occur? (2) Given a set of values for the predictors, what is the probability that the event will occur? For example, a logistic regression could examine the likelihood that a worker will complete a task successfully given a certain set of tools and worker characteristics.

*Assumptions*   Logistic regression assumes that the dependent variable is categorical/nominal.

*Methods and Results*   The logistic regression analyzes the data to provide a model of how each predictor ($X_i$) affects the probability ($\pi$) that the event (e.g., Y = "yes") will occur. The model takes the format

$$\pi = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i}}$$

where $X_i$ is the value of the $i$th predictor and $\beta_i$ is the coefficient representing the effect that the $i$th predictor has on the likelihood ($\pi$) that the event will occur. Note that the equation representing the power to which $e$ is raised is the same as the equation used in a linear regression. As in linear regression, the goal in building a logistic regression is to create a parsimonious model with sufficient explanatory power. Therefore, the first step is to determine which of the available predictors to include in the model. Although several methods of variable selection are available, the backward stepwise method is used most frequently (Kutner et al., 2004; Field, 2005). This method begins by including all predictors in the model and at each step removes the predictor that has the smallest effect on model fit. This continues until all the predictors in the model meet an established threshold of significance, typically $p \leq 0.05$. Once the model is developed, examine the model's chi-squared statistic, derived from the likelihood ratio, to ensure that the model has an adequate fit to the data.

When using a statistical package to complete a logistic regression, the model output includes the estimated coefficient and odds ratio for each factor and a $p$-value indicating the level of significance of that factor as well as the model chi-squared statistics. It also reports a classification table that reports the number of cases in which the value predicted matched the value observed and the percentage of correct classifications. It is also useful to have the package save the predicted probability that the event will occur and the predicted value of the dependent measure, since these can both be used to examine how well the model fits the data. The predicted value is determined by comparing the predicted probability to an established threshold (usually, 0.50). If the probability is greater than the threshold, the predicted value is "yes," the event will occur; otherwise, it is "no."

*Drawing Conclusions*   The first step in drawing conclusions is to determine if the model has an adequate fit. If the model chi-squared statistics has a
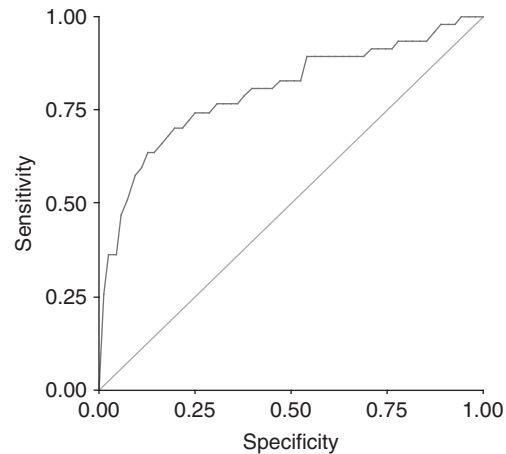


**Figure 12**   ROC curve for a logistic regression.

sufficiently small $p$-value and the percentage of correct classifications is high, we can conclude that the model has an adequate fit. Unfortunately, in logistic regression there is no value equivalent to the $r^2$ used in linear regression to estimate the degree of fit between the model and the data. However, when the response being modeled is binary (yes/no or 0/1), a receiver operating characteristic (ROC) curve may be used to estimate the degree of fit. An ROC curve (Figure 12) indicates how well the model predicts the dependent measure by comparing the true positive rate (i.e., predicted = yes and actual = yes) to the false-positive rate (i.e., predicted = yes and actual = no). If chance were used to predict the dependent variable, the prediction would be correct on average 50% of the time. The diagonal line in the graph indicates this chance fit between the predicted and actual values. The area under the ROC curve represents the fit of the model. Therefore, if there is substantially more area under the ROC curve than under the diagonal line, the fit is good. A statistical test is used to test whether the area under the curve is significantly different from 50% and provide a 95% CI on the area. The closer the area is to 1 (100%), the better the model fit.

Once the model is deemed adequate, the results of the model may be interpreted. The model may be used for prediction by plugging in a set of values for the factors in the model to determine the likelihood that the event will occur given those values. Alternatively, the model coefficients for each predictor may be examined to determine that predictor's effect size. Unfortunately, this is not as straightforward as effect estimation in linear regression, since in logistic regression a unit increase in $X_i$ multiplies the odds of the event by $e^{\beta i}$. Therefore, to interpret the effect, we examine the odds ratio ($e^{\beta i}$), which is the ratio of the probability of the event with a unit increase in $X_i$ to the probability of the event without the unit increase in $X_i$. The odds ratio should be compared to a baseline of 1 (100%). So an odds ratio of 1.25 indicates a 25% increase in the event's likelihood, and an odds ratio of 5 would indicate

a fivefold increase in the odds. Conversely, an odds ratio of less than 1 would indicate that an increase in $X_i$ reduces the likelihood of the event. By examining the odds ratio for each factor in the model, we can identify those factors that have the greatest positive and negative impact on the dependent variable being examined.

*References for further details:* Kutner et al. (2004), Johnson (1998), Field (2005).

## 4.3   Data Reduction Techniques

Several of the methods described in Section 4.2 assume that the measures being examined (e.g., predictors in linear regression, dependent measures in MANOVA) are independent of each other. However, frequently in human factors research a number of the collected measures are correlated but capture different aspects of a given phenomenon. This is often the case in complex concepts such as decision satisfaction, where it is difficult for one measure to capture every facet of the concept. When examining these correlated measures independently, it can be difficult to isolate the independent components of the phenomenon that are driving the underlying variation in those measures. When this is the case, data reduction techniques can be useful to isolate those independent components that explain the variation, thereby reducing the number of measures to be analyzed and aiding in interpretation of the data.

For example, data reduction techniques are often used to examine questionnaire data, as demonstrated in a study by Sainfort and Booske (2000) which examined postdecision satisfaction. See their case study in Figure 14. In this study, after completing a decision task, participants were asked to complete a questionnaire related to their satisfaction with their decision. The questionnaire presented 10 statements that addressed several aspects of decision satisfaction (e.g., "My decision is sound," "More information would help") and asked the participant to rate each statement on a scale from 1 (strongly disagree) to 5 (strongly agree). Obviously, the answers to a number of these statements are expected to be correlated. By examining the correlations among the measures and using factor analysis to isolate the underlying constructs affecting the variation in answers, Sainfort and Booske were able to identify four underlying dimensions of decision satisfaction: self-efficacy, satisfaction with choice, usability of information, and adequacy of information. Thus, in this example, factor analysis enabled the researchers to reduce a set of 10 highly correlated measures to four relatively independent measures and, as a result, gained insight into the underlying structure of the highly abstract concept of postdecision satisfaction.

In this section we examine two data reduction techniques: principal-component analysis and factor analysis. These two methods are conceptually very similar but differ in the mathematics used to accomplish the results. However, research has demonstrated that the two methods yield highly similar results, especially when sample sizes are large (Fava and Velicer, 1992). In general, sample sizes of at least 160 are recommended, but in some cases larger sample sizes (e.g., 300 or more) are necessary to ensure stability of the components

(Guadagnoli and Velicer, 1988). These two data reduction techniques are often not an end in and of themselves. Typically, the new component/factor variables created as a result of this analysis are used as input for other analysis techniques discussed in previous sections, such as regression or group comparisons.

### 4.3.1   Principal-Component Analysis

The primary objectives of principal-component analysis (PCA) are to reduce the dimensionality of a data set by discovering the true dimensionality of the data and identify new and meaningful underlying variables that represent the true dimensionality (Johnson, 1998). PCA is typically used as an exploratory technique to help researchers understand the data, especially the correlation structure of the data. PCA results in a set of new variables, principal components, which are uncorrelated and account for as much of the variability in the original data as possible.

*Methods and Results*   The focus of PCA is to explain the variability in the measures through the components identified. To accomplish this, PCA produces an orthogonal transformation of the measures into a number of principal components, which constitute a linear combination of the measures being examined and depend solely on the covariance of those measures. The method begins by identifying the component that accounts for the largest portion of the variation, as indicated by the eigenvalue. This continues, with each component accounting for less of the overall variation.

One of the first steps in PCA is determining the appropriate number of components. In other words, how many "true" dimensions are represented by the data? There are several methods for determining this number, although it is more of an art than a science. The first method is to establish a minimum threshold, usually 1, for a component's eigenvalue. Alternatively, we could examine a scree plot, which plots the eigenvalues of each component as illustrated in Figure 13. In the scree plot, look for the cutoff at which the eigenvalues level off or create an elbow. All components to the right of the elbow should be eliminated, since they add little additional explanatory power. Based on this method, for the results in Figure 13, we would keep components 1 and 2 and eliminate the others. The third method for selecting the appropriate number of components is to establish a threshold for the amount of variability in the original data that is accounted for by the principal components and then select the minimum number of components that meet this threshold. For example, if a researcher wanted to account for 80% of the variation in the original data, and the first four components accounted for 50, 25, 10, and 8% of the variation, respectively, the researcher would use the first three components, which account for $50 + 25 + 10 = 85\%$ of the variation.

Once the principal components are identified and derived, the factor loadings are reported for each component. The $i$th principal component ($C_i$) for a given item is calculated based on the factor loading ($\beta_i$) and value ($X_{ij}$, where $j$ represents the item) of each
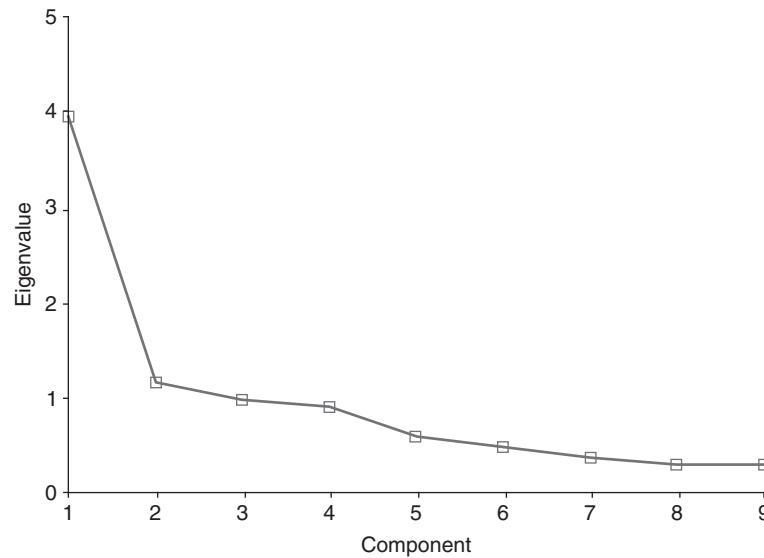
**Figure 13** Scree plot in PCA.

measure that loads on that component: $C_{ij} = \beta_1 X1_j + \beta_2 X_{2j} + \cdots + \beta_i X_{ij}$. The absolute value of the factor loading indicates the weight of that measure in calculating the component. Absolute values close to 1 are very important to the component. The sign of the factor loading indicates the direction of the relationship between the component and original measure; a positive value indicates that an increase in the measure will increase the value of the component score. These factor loadings are used to calculate principal-component scores for each item based on the values of the original measures. After these values are calculated, they can be used for subsequent analysis. Frequently, principal components are used as inputs for cluster analysis, discriminant analysis, and multiple linear regression.

*Drawing Conclusions*     Although many researchers use the results of PCA as input to additional analyses, there are several interesting conclusions that can be made from the results of PCA themselves. These conclusions center around understanding the true dimensionality of the data. For example, if the percentage of variability accounted for by the components selected is sufficiently large, the number of components identified can be interpreted as the number of true dimensions in the construct being measured by the data. By examining the factor loadings for each measure that contributes to a component, we can determine if that component, which represents a dimension of the data, has any interpretable meaning. It is easiest to interpret the components when a measure has a high factor loading for one component but not on any of the others and when the combination of measures that have a high loading are conceptually or logically related in some way. For example, if we completed a PCA on a number of measures that contribute to productivity and one component had high factor

loadings for a worker's education level, number of training certifications, and number of years on the job, we might interpret this component as the worker's knowledge level. Note that in many cases the resulting principal components may not be interpretable. When this is the case, it is often worthwhile to complete a factor analysis on the data set as well, especially since functionality in the current generation of statistical packages makes running either analysis relatively quick and easy.

*References for further details:* Johnson and Wichern (2007), Johnson (1998).

### 4.3.2 Factor Analysis

Conceptually, factor analysis (FA) and PCA are quite similar. Both methods examine the true dimensionality for a large set of correlated measures. However, unlike PCA, which does not depend on an underlying model, FA depends on a reasonable statistical model. Also, FA is focused on explaining the covariance and correlation among the measures, which is in contrast to PCA, which focuses on explaining the variability of the measures. These are subtle distinctions from a conceptual standpoint, but they affect the mathematics used to develop the components, or factors, that result from each method. In PCA, the components identified are linear combinations of the measures. In contrast, in FA, the measures are linear combinations of the factors identified. In most cases, the results of the two methods are highly similar, but sometimes FA results are more interpretable as a result of factor rotation, as discussed next. As with PCA, the output of FA is a new set of uncorrelated measures or underlying factors.

*Methods and Results*     Similar to PCA, FA analyzes the variance or covariance of the measures being examined to identify factors (components) that account

for most of the variability in those measures. However, in FA, several different methods can be used to estimate the factors. Guidelines as to which method is best are relatively vague, but two frequently used methods are principal factor, which is similar to PCA, and maximum likelihood, which should only be used to analyze multivariate normal data. As in PCA, the researcher must determine the number of factors to include in the analysis. The criteria used in PCA, which are used to examine the eigenvalues, are typically used as a starting point, although additional guidelines can be used. For example, Johnson (1998) identifies minimizing Akaike's information criterion (AIC) or Schwarz's Bayesian criterion (SBC) as objective rules that can be used to determine the ideal number of factors.

In factor analysis, the set of factors and their factor loadings generated through the initial analysis are not unique. By multiplying the resulting factor loadings by an orthogonal matrix, the factor loadings can be rotated in space. What exactly does this mean? We won't dwell on the mathematics, but conceptually, rotating the factor loadings obtains a mathematically equivalent set of factor loadings that may be easier to interpret. In other words, we as researchers can look for the rotated set of factor loadings that can be interpreted most meaningfully. Note that some statisticians and researchers view this as a criticism of FA, since exactly which rotation is "meaningful" is open to interpretation; however, many people view this as an advantage since it enables the resulting factors to be more easily interpreted and applied (Johnson, 1998).

Several methods are available for rotating the initial factor loadings. Factor rotation methods have two goals. The primary goal is to rotate the factor loadings for each factor so that all measures either load heavily on that factor (i.e., have an absolute value near 1) or load very little on it (i.e., have an absolute value near zero). The secondary goal is for each measure to load heavily on only one factor. Meeting these goals typically simplifies interpretation since the measures that weigh heavily on a factor are isolated and each measure (ideally) contributes significantly to only one factor. There are two classes of rotation methods: orthogonal rotation methods, which maintain the independence between the identified factors, and oblique rotation methods, which are appropriate only when the factors are not assumed to be independent. In general, orthogonal methods are recommended over oblique rotation methods unless there are theoretical grounds for assuming that the factors are not independent. Of the orthogonal rotation methods available, varimax is used most frequently

*Drawing Conclusions* The conclusions drawn from FA are almost identical to those drawn in PCA: Researchers may use the resulting factors from FA as input to additional analyses and/or draw conclusions about the true dimensionality of the data. The number of components identified can be interpreted as the number of true dimensions in the construct being measured by the data, and by examining the factor loadings for each factor, we can determine if that factor has any interpretable meaning. Because FA uses factor rotation to

ensure that factors have high factor loadings for a subset of measures and low factor loadings for the remaining measures, meaningful interpretation of the factors is usually easier than in PCA. (For example, see the results highlighted in the case study in Figure 14.) However, because of the variety of methods that can be used to determine the number of factors, estimate the factors, and rotate the factors to find meaningful factor loadings, factor analysis is fairly subjective. As such, be prepared to support your conclusions with tests to ensure validity and reliability. When presenting your conclusions, provide sound theoretical support from related work in the literature. Furthermore, validity can be examined by comparing the resulting factors to related outcome measures not used in the FA. Reliability can be examined by completing the same analysis on a similar data set and comparing the number of factors and factor loadings that result from each data set. See the case study in Figure 14 for an example of how the researchers established the validity and reliability of their results.

*Effects of Measurement Errors on Factor Analysis* Measurement errors in correlations can result in the relationship between $X$ and $Y$ to be underestimated or attenuated (Liu and Salvendy, 2009). Because factor analysis is examining intercorrelations, it has similar results from errors in measurement. If measurement errors occur with variable $X_i$, then the correlations between $X_i$ and other variables will be deflated, causing the contribution of the first few important factors to decrease (Liu and Salvendy, 2009). For further information regarding measurement errors in factor analysis, see Liu and Salvendy (2009).

*References for further details:* Johnson and Wichern (2007), Johnson (1998).

# 5  ANALYSIS OF UNSTRUCTURED OUTCOME DATA

It is apparent that a wide variety of methods are available to analyze structured outcomes. But what can be done with all of the unstructured data that are generated? Unstructured outcomes are generated from a number of research methods. For example, unstructured outcomes are the results of open-ended interview or survey questions, observations from the field, observations during an experimental task, documents, and transcripts from video recording and think-aloud and other verbal protocols. These data are rich in detail, and unlocking the themes and trends from these data can provide invaluable insight into the topic being researched.

Unfortunately, unlocking these secrets can be quite challenging, especially when dealing with large amounts of data. Researchers usually begin by pulling out the data that are relevant, where relevance is determined by the nature of the data collected and the research questions to be answered (Wixon, 1995). Once the relevant data are identified, several techniques can be used to develop and communicate findings from unstructured data. In the following sections we discuss methods that can be used to accomplish these analysis objectives for unstructured data: (1) impose structure on unstructured data using coding and classification, (2) provide

*Research Objectives*: Given the proliferation of decision aids used to help individuals make complex, value-laden decisions, there is a need for reliable, objective methods for evaluating the effectiveness of these decision aids. In this study, Sainfort and Booske (2000) develop the decision-attitude scale, which is used to assess the decision maker's satisfaction after making a decision.

*Data Collection Method:* Participants examined information on health insurance plans using a computer-based decision aid. Participants selected a plan at two points in time: first after viewing a minimal amount of information about the plans, and again after having the opportunity to view extensive information about each plan. At each decision point, participants completed the decision attitude survey consisting of ten statements related to various aspects of decision satisfaction. Participants rated whether or not they agreed with each statement using a scale from 1 (strongly disagree) to 5 (strongly agree).

*Analysis 1: Examine relationships among decision satisfaction answers*
Test: Pearson's Correlation
Results: Correlation coef cients

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | I had no problem using the information | 1.00 | | | | | | | | |
| 2 | I am comfortable with my decision | 0.50 | 1.00 | | | | | | | |
| 3 | The information was easy to understand | 0.64 | 0.39 | 1.00 | | | | | | |
| 4 | I wish someone else had made the decision for me | 0.10 | 0.07 | 0.09 | 1.00 | | | | | |
| 5 | It was diffcult to make a choice | 0.40 | 0.52 | 0.43 | 0.13 | 1.00 | | | | |
| 6 | I am satisfied with my decision | 0.54 | 0.70 | 0.46 | 0.07 | 0.50 | 1.00 | | | |
| 7 | My decision is sound | 0.41 | 0.57 | 0.40 | 0.14 | 0.49 | 0.61 | 1.00 | | |
| 8 | More information would help | 0.18 | 0.38 | 0.31 | 0.05 | 0.33 | 0.37 | 0.37 | 1.00 | |
| 9 | My decision is the right one for my situation | 0.29 | 0.46 | 0.35 | 0.12 | 0.40 | 0.49 | 0.57 | 0.40 | 1.00 |
| 10 | Consulting someone else would have been helpful | 0.20 | 0.30 | 0.37 | 0.04 | 0.31 | 0.33 | 0.22 | 0.49 | 0.32 |

Conclusions: Examining the correlation coef cients among the responses to each statement, most of the responses are correlated, with the exception of statement #4, which appears to be relatively independent of the other responses. Because of the correlations among the response, the researchers' next step was to use factor analysis in order to gain insight into the underlying factors driving the variation in these responses.

*Analysis 2: Exploring underlying factors of the correlated responses*
Test: Factor Analysis. Response #4 was excluded from analysis since the correlation indicated it is independent. To ensure the reliability of the scale, the factor analysis was performed on the responses obtained at the rst decision point. A second factor analysis was completed using the responses obtained at the second decision point to ensure that comparable factors and factor loadings resulted from both data sets.
Results: Three factors, accounting for 71% of the total variation, were identi ed. As an example, the factor loadings for factor 1 are provided below, sorted on the factor loadings. Factor 1 had an Eigenvalue of 4.37 and accounted for 48.6% of the variation.

| *Response* | *Loadings on Factor 1* |
|---|---|
| 7. My decision is sound | 0.828 |
| 2. I am comfortable with my decision | 0.741 |
| 9. My decision is the right one for my situation | 0.729 |
| 6. I am satisfed with my decision | 0.721 |
| 5. It was diffcult to make a choice | 0.574 |
| 8. More information would help | 0.364 |

| *Response* | *Loadings on Factor 1* |
|---|---|
| 1.  I had no problem using the information | 0.312 |
| 3.  The information was easy to understand | 0.197 |
| 10.  Consulting someone else would have been useful | 0.082 |

Conclusions: Based on the results, the first five responses (items 7, 2, 9, 6, & 5) factor most heavily on factor 1(loading > 0.55). These items also had much lower loadings on the other two factors. The remaining four factors, all had low loadings on this factor (loading < 0.4) and high loadings on one of the two other factors (loading > 0.75 for one of the remaining factors and loading < 0.3 on the remaining factor). Considering the items that loaded heavily on factor 1, the researchers concluded that factor 1 represents 'Satisfaction with Choice'. The remaining two factors had similar interpretability, representing 'Usability of Information' and 'Adequacy of Information'.

Note: In this study, a number of tests were completed to ensure the validity of the identified factors in the decision satisfaction scale. In the interest of brevity, only the results from one test are highlighted here.
Test: Paired t-test
Results: Paired t-tests examined difference on each of the factors between the time of the first decision point and the second:

| *Factor* | *t-statistic* | *p-value* |
|---|---|---|
| Satisfaction with Choice | − 3.461 | < 0.01 |
| Adequacy of Information | − 9.619 | < 0.01 |
| Usability of Information | − 1.608 | 0.11 |

Conclusions: Because additional information was made available to participants after the first decision point, intuitively, one would expect participants to be more satis ed with their choice and the adequacy of the information at the second decision point Results indicate that participants were indeed more satis ed with their choice (p< 0.01) and the adequacy of the information (p< 0.01) at the second decision point. This is indicates that the scale is sensitive enough to detect the differences in these two aspects of decision satisfaction. Usability of information had only a marginal difference (p=0.11), which, again, makes intuitive sense because though more information was presented between decision points one and two, a similar presentation format was used, so information usability would not necessarily improve. These results led the researchers to conclude that the scale had adequate discriminant ability to detect changes over time attributed to the presentation of additional information.

**Figure 14**   Case study: measuring postdecision satisfaction.

an aggregate view of the topic using figures and tables, and (3) provide a detailed record of the topic through documentation.

## 5.1   Content Analysis and Coding

Content analysis, or coding, is used by researchers to understand and summarize unstructured data. In content analysis, qualitative unstructured data are examined systematically to identify key themes and, subsequently, to classify events, observations, and answers into one or more of those themes or categories. Although there are subtle distinctions between coding and content analysis, we avoid those semantic discussions for now and focus on methods of completing these types of analysis, which are, for the most part, the same. Content analysis and coding are used in a variety of fields, especially in market research and the social sciences, where more descriptive research methods are frequently used. In human factors, coding has been used, for example, in incident and accident analysis and to provide structure to outcomes from verbal protocols such as think-aloud and other unstructured outcomes gathered during a research study.

When completing content analysis, the goal is to apply the coding scheme to the data objectively and systematically. The coding scheme is, in effect, the set of categories or themes to which an event/observation/ answer may belong and the rules used to classify a

given item into the appropriate category or theme. If the coding scheme is not applied consistently, the reliability of the coded data, and as a result the validity of the research results, is questionable. Therefore, close attention must be paid to the process used to code the data to ensure the reliability of the results.

### 5.1.1   Coding Process

The following steps comprise the coding process. For purposes of simplicity, an example of coding responses to an open-ended survey question is used. However, the same method could be applied to coding verbal protocols, observation notes, or other forms of unstructured data. Throughout this process, be on the lookout to avoid systematic biases that can result from inconsistent use of the coding scheme.

1.   Develop the coding scheme.
   a.   If there is an appropriate existing coding scheme that has been validated and used in the literature for the content being analyzed, use that coding scheme to improve the external validity and comparability of the results. For example, coding and classification schemes are available for analyzing certain types of communication. If

an appropriate existing scheme is available, skip to step 2.

b. To develop a specific reliable coding scheme, first examine the responses from a representative sample of the surveys. If there are only a small number of surveys, examine all of the responses.

c. Based on knowledge of the actual responses and research objectives, identify a set of categories for classifying responses. Design the categories such that minimal information is lost in the coding. Also, design the categories to be as distinct from each other as possible to reduce confusion in classification.

d. Establish definite rules and criteria for assigning a response to a category. Provide examples to clarify complex rules, especially when coding requires a judgment on the part of the coder. (For example, assigning a rating of "level of understanding" based on a participant's response requires more coder judgment than that needed for coding a fact-based response such as assigning a job category based on the participant's job title.) Defining clear rules for assigning responses to categories is *crucial* to ensuring intercoder reliability.

2. Establish a coding protocol.

a. Determine the number of coders. At least two coders should be used so that intercoder reliability can be assessed. For large data sets, more than two coders may be needed.

b. Recruit the coders. If possible, people with experience in content analysis and with relevant research content knowledge should be used.

c. Define the coding protocol, which should include how many responses each coder will code, how questions regarding coding of questionable responses will be resolved, how coding will be validated, and how differences in coding will be resolved.

3. Test and revise the coding scheme.

a. Using a small sample of responses, preferably not the same sample as that used to establish the initial coding scheme, test the coding scheme. If any categories appear too broad or narrow, adjust the coding scheme as needed. Also, if any of the classification rules are unclear, revise them as needed to ensure that they will be applied consistently.

4. Train the coders on the coding scheme and coding protocol. Allow them to practice with examples and to ensure that they have an adequate understanding of the coding scheme.

5. Code the responses.

a. Assign the appropriate number of responses to each coder. For small samples, have each coder code all responses. For larger samples, make sure that a representative sample of the responses is coded by multiple coders so that intercoder reliability can be assessed.

b. The primary researcher or a coding supervisor should resolve any questions that arise during the coding. Any clarifications should be communicated to *all* coders, so that coding will be consistent among all coders.

c. The primary researcher or a coding supervisor should spot-check a sample of the responses to ensure that they are coded properly. Any problems identified in the coding should be resolved as quickly as possible to prevent the need to recode responses.

6. Assess intercoder reliability using one of the methods described in Section 5.2.3.

## 5.1.2 Improving Intercoder Reliability

In their study of coder variability, Kalton and Stowell (1979) make several observations and recommendations intended to reduce the amount of variability in coding among coders. Considering and applying these findings when designing and executing coding schemes can improve intercoder reliability:

- When constructing coding frames, take into account the actual responses so that those responses can be more readily mapped to the codes.

- When constructing coding frames, limit the use of catch-all codes (e.g., "other"), since these codes tend to be applied unreliably. Instead, try to use more clear-cut codes whenever possible.

- Be aware that fact-based codings, which are more objective, tend to have higher reliability than judgmental codings, which require more interpretation on the coder's part.

- Use training and strict supervision to ensure that coding frames are applied uniformly.

- Especially when judgmental codings are required, the researcher should stay closely involved with the coding operations (training, spot-checking coded data, etc.) to ensure that the codes are being interpreted and used as intended to support the research objective.

## 5.1.3 Assessing Intercoder Reliability

Intercoder reliability is the degree of match between the assigned codes of two different coders who apply the coding scheme independently to the same set of responses. Even when researchers take great care to ensure that there is adequate intercoder reliability, they should always assess and report intercoder reliability to ensure the credibility of their results. Unfortunately, many researchers omit this step, probably because there is relatively little agreement on the best way to assess

intercoder reliability. Frequently, researchers report the percentage of agreement among coders. However, the one thing that most content analysis researchers agree on is that this is not the best approach. Instead, estimates of intercoder reliability should include adjustments that correct for chance agreement among coders (Hughes and Garrett, 1990; Grayson and Rust, 2001). Consider, for example, a case in which two coders are coding data that can fall into one of two categories. By random chance, those two coders will agree 50% of the time. Therefore, the assessment of reliability needs to indicate that agreement between coders is better than agreement that would be achieved by random chance.

A number of methods for assessing intercoder reliability have been presented in the content analysis literature. All of these measures range from 0 to 1, with values closer to 1 indicating high intercoder reliability. Cronbach's alpha is an intraclass correlation measure frequently used to measure agreement using the ratio of the true score variance to the sum of the true score variance and the error variance. Other methods used frequently include Scott's pi, Krippendorff's alpha, and Cohen's kappa. These three measures are conceptually similar: They are calculated by taking the difference between the agreement observed between coders and the agreement expected using random chance and adjusted based on the agreement expected using random chance (the chance correction). They differ in how they calculate the chance correction. Cohen's kappa is sometimes criticized because in certain cases the maximum possible value of kappa is less than 1. Due to the assumptions made in Scott's pi and Krippendorff's alpha, they are appropriate only in cases where intercoder bias is assumed to be negligible.

Once the appropriate method of calculating intercoder reliability is selected, there are several aspects of intercoder reliability that should be examined, even though the overall reliability is the only thing that is usually reported. First, if there are more than two coders, examine the reliability of each coder. If one coder's reliability is significantly less than the others, it indicates that they may have been using the coding scheme inconsistently with the other coders. Next, examine the reliability of each code. This will highlight any codes that are used less consistently than the others. This may indicate that the rules for assigning responses to this category need to be clarified or the category itself needs to be redefined. For an even deeper assurance of coder reliability, examine the percentages of coders assigning responses to each code. This helps pinpoint where the discrepancies are that reduce the reliability of specific coders or codes. For example, is coder A assigning a disproportionate amount of responses to the "other" category? Is coder B underutilizing category 2? By identifying these discrepancies, appropriate steps can be taken to revise the coding scheme or provide additional training to the coders to improve the overall reliability of the results.

### 5.1.4   Using the Results

Once the coding is completed, the results are structured data, usually in the form of frequencies of occurrence for each category. These results in and of themselves may provide the answers to the research questions of interest. Otherwise, these structured data derived from the unstructured data may be used as input for one or more of the structured analysis methods described previously. In either case, when reporting results based on the coded data, be sure to report an overview of the coding method used and the intercoder reliability achieved so that other researchers can be assured of the validity of the results and conclusions.

*References for further details:* Krippendorff (2003), Hughes and Garrett (1990), Oppenheim (1992).

### 5.2   Figures and Tables

Summarizing unstructured data in an informative figure or table can also be an invaluable tool for communicating findings gleaned from unstructured data. In some cases, figures and tables are more effective than text for communicating certain types of information. They are most effective when combined with written documentation that complements (but does not duplicate) the content in the figure or table. Take, for example, the decision model for selecting the appropriate structured analysis method presented in Figure 2. This figure does a much better job of conveying in a clear and concise manner the methods presented in the chapter, their relationships, and the factors that go into selecting the appropriate method than a written description of that information could. The text in the chapter supplements the figure by providing additional details on each of the decision points and methods identified in the figure. The figure, in effect, provides readers with a map of the content presented, helping them understand the big picture up front before delving into the details of a specific method. Using a figure instead of a text description to present the decision factors and relationships between goals and methods shortens the chapter. Also, presenting the big picture of the content up front makes it easier for readers to understand the detailed material as they read it. Both of these factors combined (hopefully!) make the chapter easier to read.

The appropriate type of figure or table to use will depend on the type of unstructured data being summarized and the purpose and audience of the document (or presentation) in which the figure or table will be included. A wide variety of figures and tables can be observed in the human factors and other literature. Examples of the several useful types are:

- *Maps.* Maps of physical space are obviously effective for conveying information relative to existing and proposed workspaces. They can also be effective for conveying paths that worker and/or materials must follow in the course of a given task.
- *Concept Maps.* Concept maps are a physical representation of relationships between ideas or concepts in a knowledge space. In a concept map, concepts that are closely related conceptually are placed close to each other on the content map, and vice versa. Concept maps, in effect, present

the relationship between abstract concepts in a physical manner.

- *Flowcharts*. Flowcharts have many obvious uses, including conveying processes or a series of tasks, communicating information/material flows, and others.
- *Decision Models*. Decision models can be used to map out the steps in a decision process. They include decision points, decision factors, and related activities (e.g., data gathering) that are part of the overall decision process. In situations where multiple people are involved in a decision, the decision model may also indicate who is responsible at each point in the process.
- *Organization Charts*. Organization charts graphically illustrate members of an organization (people or suborganizations) and how they are related. These can be quite useful for depicting divisions of responsibility, chains of command, and communication channels.
- *Hierarchies*. Hierarchies are useful for presenting the breakdown of high-level concepts or groupings into lower level concepts or grouping. For example, task hierarchies are used to break high-level tasks into the steps required to complete that task.
- *Time Lines*. Time lines are useful for presenting a series of events that took place during research. They can also be effective at presenting a historical perspective on a particular domain of research or results of related research over time.
- *Literature Tables*. Literature tables are useful to summarize concisely the relevant literature related to a particular research study or meta-analysis.
- *Matrices*. Matrices are useful for summarizing data along a small number of dimensions. For example, a matrix could present communication patterns and topics by presenting people along two axes and using a symbol in the row−column square to indicate frequency or topic of communication between the two people.

It is left to the reader to determine which type of figure or table will best convey the information they have to present. However, when developing these tables and figures, think like a human factors researcher and design them to be easy to perceive and use. Use the following guidelines (adapted from Gillan et al., 1998) to present quantitative data in papers to improve the readability and usability of figures and tables:

- Design your figure or table to support your readers' cognitive tasks.
- Make sure that all labels and text are readable. Use a sufficiently large font size and high contrast between the background and text (e.g., white background, black text).
- Use clear, concise wording. Long words and phrases can clutter the display and make it difficult to read. Also, more concise wording enables the use of larger font sizes.

- Make all colors and symbols used to convey information clearly discernible from each other and convey their meaning clearly through a legend or key (if there are many symbols) or labeling (if there are only two or three symbols).
- Use symbols, colors, and other meaningful features consistently within a figure or table and, if possible, across all figures and tables.
- Consider how readers will perceive the figure or table.
  - Make the main point visible at first glance.
  - Attract the readers' attention to the most important features.
  - Use graphical techniques (line thickness, color, etc.) for emphasis.
- Eliminate clutter to improve visual searching of the figure or table.
- Place related items close to each other so that relationships between them are more readily apparent.

Whenever possible, use structures and presentation methods that are familiar to the user. For example, in flowcharts, adhere to the generally accepted meanings associated with different shapes.

## 5.3 Documentation

Research results are presented in document(s) that detail a study's findings. When large volumes of unstructured data are involved, documentation may be the only way to communicate and present those findings. However, especially in the case of unstructured data, it can be difficult to determine the appropriate scope, level of detail, and organization for documenting the findings. When writing, keep in mind that the goal is not to write the largest document possible, including every detail encountered in the unstructured research outcomes. Instead, as my fifth-grade English teacher instructed, the goal is to make it long enough to cover the subject but short enough to be interesting. The last thing you want to do is spend weeks writing a detailed work analysis only to have the 4-in.-thick monster end up as someone's doorstop. "Brevity is the soul of wit," and it saves you and your readers time and increases the readability of your document. So the primary question becomes how you can clearly and concisely present results based on unstructured outcomes.

There are a number of technical writing guides that help provide the answer to that question. Many of these guides (e.g., Gerson and Gerson, 2007; Reep, 2010) provide guidelines for specific types of documents, so we will not delve into those details here. However, it is useful to review more general principles and methods for good technical writing. These principles are helpful in determining which unstructured outcomes to focus on in the report and how to present those outcomes:

1. *Know the audience.* To communicate effectively with the audience through the document, you must first identify and understand the audience. A variety of audience characteristics

should affect how your present the material. For example, it is important to consider the audience's subject knowledge or level of expertise (Gerson and Gerson, 2007; Reep, 2010). Readers who are experts will probably need less detailed explanations of concepts and definitions than novice or lay readers. Also consider the reader's motivation or purpose for reading the document. This will have a strong influence on the appropriate scope, level of detail, and structure of the document.

2. *Define the objective(s) of the document.* Before writing the first word, define the objective(s) of your document. Beginning with these objectives in mind, it will be easier to organize and present the information needed to achieve those objectives. Without a well-defined audience and objectives, you may find yourself mired in a series of extensive reorganizations and revisions. In the case where there are multiple objectives, consider carefully if all of the objectives identified can be met in one document. If any of the objectives require presenting significantly different information or target drastically different audiences, it may be more effective to address those objectives in separate documents. Although writing two documents may take more time, it will be time well spent if it ensures that the documents will be read and your objectives achieved.

3. *Keep in mind general objectives in technical writing.* Gerson and Gerson (2007) recommend that writers strive for clarity, conciseness, accuracy, and organization in their writing. Writing for *clarity* helps ensure that the audience understands what they have read. To achieve clarity, provide answers to anticipated questions, provide specific details, and use terms that are easily understood by the audience. Writing *concisely* is beneficial, as it saves time for both the author and reader, but it can also improve comprehension of the material. Conciseness can be achieved by limiting sentence length, omitting redundancies, and avoiding wordy phrases. *Accuracy* entails ensuring that the document is grammatically, factually, and textually correct. Accuracy can usually be achieved by thorough proofreading. Be especially careful to check figures, equations, and references. The *organization* of the document is crucial to communicating effectively. Although the appropriate organization will depend on the purpose and content of the document, be sure that the information is organized and presented logically. One way to achieve this is to start by constructing an outline.

4. *Construct an outline.* An outline is a useful tool for improving the clarity, conciseness, and organization of the document. Whether informal or formal, outlines usually begin as a list of major topics and subtopics that will be addressed in the document. The list is transformed into an outline as the author(s) reorders the topics until they are presented in a logical order that facilitates understanding the material and achieving the stated document objectives. As Reep (2010) points out, developing an outline enables the author to see the document structure before he or she begins writing. It also enables the author to focus on presenting and explaining information as they write the rough draft, as opposed to organizing and writing at the same time, which often results in poor organization and redundant content presentation. When multiple authors are working on the same document, creating an outline has added benefits. By creating an outline, the authors discuss and agree on the overall content and structure before they begin writing. Then each author can draft his or her sections of the document in parallel, knowing how they fit into the overall content and structure. This greatly simplifies merging those sections later in the writing process.

5. *Use figures and tables.* When appropriate, use figures and tables, as discussed earlier, to convey information. In many cases, figures and tables can convey complex information or provide a big picture overview more clearly and concisely than can a written description. Gerson and Gerson (2007) provide useful criteria for using figures and tables in documents. Good figures and tables:

   - Are integrated with the text (i.e., text explains graphic, and vice versa)
   - Add to the material conveyed in the text but are not redundant with the text
   - Communicate important information that would be difficult to obtain easily in longer text
   - Do not include details that would detract from the information conveyed
   - Are located close to the text that refers to them (preferably on the same page)
   - Are appropriately sized
   - Are readable
   - Are labeled correctly with legends, headings, and titles
   - Use a style consistent with other figures and tables in the document
   - Are well conceived and executed

6. *Use appendixes to convey supplementary information.* Use appendixes to provide useful additional detail that only some readers will need or that is too detailed for inclusion in the main body of the document. For example, an appendix might include highly technical information or examples of surveys. Also, when extensive statistical analysis has been completed, it can be useful to highlight the significant results in the main document and include the full statistical results in an appendix in order to improve the clarity and conciseness of the main document.

## 6 ANALYZING SURVEYS

The results of surveys are a special case in analyzing both structured and unstructured outcome data for human factors research. As shown in Chapter 11, surveys may vary greatly in their purpose, content, and structure, and great care must be taken in their development to ensure that the survey gathers data that are reliable and valid. In this chapter we do not revisit those concepts. Instead, our focus is on applying the analysis techniques discussed previously to analyze survey results in a way which ensures that the results and conclusions are valid.

In terms of format, survey questions can be either structured or unstructured (i.e., open ended). In structured questions, answers may be nominal/categorical, ordinal, or interval/ratio. As demonstrated earlier, the measurement type of the answer has a strong influence on which analysis methods can be used to examine the data and find answers to the study's research questions. Therefore, we recommend considering the research objectives and the data analysis methods that support meeting those objectives when designing survey questions. This ensures that the format of the question results in answers that are of the appropriate measurement type to support the analysis methods that will result in achieving your research objectives.

### 6.1 Validating the Data

Before analyzing survey results, it is important to review the responses to ensure that there are no gaps in the data. One of the steps in validating the data is to check the overall response rate to ensure that it is sufficiently high. This is not as much of an issue for surveys administered by researchers in the course of lab-based research where response rates should be 100%. However, in mail or telephone surveys, response rates are very important. If a large number of potential participants do not respond or there is a pattern to the participants that do not respond, it can inadvertently bias the survey results. If this is the case, do not even bother analyzing the results. Unfortunately, this means that you should reexamine your survey and data collection procedures and modify them as needed to increase your response rate; then begin collecting data again.

If the overall response rate is adequate, review the individual questions to determine if there are missing answers. If a participant did not answer a large portion of the questions, it could indicate that he or she did not understand the survey or questions. If this is the case, it may make sense to remove that person's entire survey from the results. In general, if a participant is missing data for only one or two questions, it is all right to leave the data in for analysis that does not involve that question. Of course, if the missing data are crucial to the analysis, the participant's entire survey may need to be excluded. For example, if a survey is examining gender differences in communication and a participant omits his or her gender, those data are useless. If a large number of participants omit answers to the same question, it could indicate that the question was confusing or that participants were uncomfortable answering the question.

In this case, again, the research must judge whether on not it is valid to include this question in the analysis. Whatever the cause for the missing data, it is crucial for the researcher to be aware that the data are missing and of the implications the missing data have or can have on the results. It is up to the researcher to determine the appropriate course of action to ensure that the missing data do not compromise the results of the survey analysis.

### 6.2 Analysis of Unstructured Answers

Although unstructured questions are typically less time consuming to write, their answers are more time consuming to analyze. The appropriate analysis of unstructured answers, of course, depends on the purpose of the question and research. Some researchers, interested in general information gathering only, may be able to get away with simply reading and summarizing the answers. Unfortunately, a more rigorous method is often required to summarize the answers in a structured manner so that statistical methods can be used to summarize and analyze the results. For this more rigorous analysis of answers, coding is required.

Note that coding for survey answers can be more difficult than coding a researcher's field notes and other unstructured data captured by the researchers involved in a study. One reason for this increased difficulty is that answers from different participants are often not directly comparable (Alreck and Settle, 2003). When unstructured data are gathered from researchers, they (hopefully) share a common understanding of the research objectives and intent and a relatively similar vocabulary. However, survey participants frequently do (and should) vary widely. Participants may use different vocabularies, have different meanings for similar words, or have different understanding of the question's intent. Additionally, their answers may be vague and the researcher is often unable follow up with a participant to get clarification on an answer. (Also, given the amount of time that has elapsed between when the survey is administered and when a clarification is requested, the validity of the clarification may be called into question.) Therefore, when coding unstructured answers to survey questions, it is vital to take special care in developing a coding scheme and assessing intercoder reliability as recommended in Section 5.1.

For unstructured questions, the purpose of coding the answers is often to capture common themes. For example, if the question was "What did you like best about this interface?" the researcher might want to identify common concepts or themes regarding what participants liked best about the interface. Coding the data based on these concepts results in nominal/categorical measure(s). Using these nominal measures, the researcher could count the frequency with which each concept was mentioned and examine relationships between these frequencies and demographic or other measures collected through the research study. In other words, coding the unstructured answers imposes structure on those data so that the researcher can use analysis methods for structured data to gain further insights from the data.

**Table 4  Example Question Formats**

| Question Format | Example | Measurement Type |
|---|---|---|
| Multiple choice | What is your job category?<br>(a)  Sales<br>(b)  Administrative<br>(c)  Engineering<br>(d)  Management<br>(e)  Other | Nominal/categorical |
| Single response item | Have you completed XYZ certification? | Nominal/categorical |
| Order/rank items | Rank the three tools based on which is easiest to use (1 = easiest, 3 = hardest) | Ordinal |
| Likert scale | 1              2         3          4           5<br>Strongly agree    agree    Neutral   Disagree   Strongly disagree | Interval |
| Frequency scale | 1      2        3          4         5<br>Always  Often  Sometimes  Seldom   Never | Interval |

## 6.3  Analysis of Structured Answers

When analyzing structured answers from surveys, a number of the analysis methods described previously in this chapter may be used. However, there are some special considerations that must be taken into account in analyzing survey answers. In this section we address some of those considerations as well as map some of the general measurement and statistical concepts discussed previously as to their use for survey analysis.

### 6.3.1  Measurement Types

At the beginning of this chapter, several measurement types were defined. The answers to structured survey questions can be classified using these measurement types, which determine the appropriate analysis methods. Table 4 identifies the measurement type for several common survey question formats. Note that in multiple-choice questions, where the participant can select multiple items, for data analysis purposes it is sometimes easier to treat each possible answer as a separate single response (yes/no) item and/or to create a measure that indicates the number of responses. For example, if a question asks users to mark all the types of computer applications they use, code each application as a separate answer [e.g., word processing (Y/N), spreadsheet (Y/N)]. This enables comparing use among application types and comparing application use to other measures collected in the study (e.g., task performance, participant preferences). Additionally, comparing the number of application types used (i.e., a count of the application types used) to other participant responses and/or task performance can also provide interesting insights. In this example, the number of application types used can measure the breadth of computer experience, which can be a crucial covariate in human–computer interaction research.

### 6.3.2  Exploring and Describing the Data

As in any analysis of structured outcomes, the first step is exploring and describing the data (answers). The methods described in this chapter apply here. Begin by graphing the answers individually through histograms and box plots to examine the distribution, range, and variability of answers. Next, describe the data using the mean, median, percentiles, minimum, and maximum. Note that the appropriate descriptors will depend on the research objectives (e.g., measuring central tendency vs. range and distribution of values) and the type of data. For example, with ranking data, it may make more sense to examine percentages of responses (e.g., 60% ranked this item 1, 10% ranked it 2, and 30% ranked it 3) instead of mean ranks (what does a mean rank of 1.7 mean?).

After examining each answer individually, the next step is to explore possible relationships between answers. Here, again, scatterplots are useful graphical means of examining possible relationships. Also use correlation to examine and describe relationships among ordinal and interval/ratio answers. Use contingency tables (crosstabs) and the chi-squared test to examine relationships among nominal/categorical answers. When correlations exist, they often provide interesting insight into the relationship between two answers, but keep in mind that correlations do not necessarily imply causality. These correlations are, however, crucial to determining the appropriate way to analyze and make inferences from the data.

### 6.3.3  Making Inferences from the Data

Once we have gotten familiar with the structured data, we can use the more advanced statistical methods presented in this chapter to make inferences about the results. However, because of the nature of survey data, special care must be taken when using these methods. Survey data require special considerations for two reasons: (1) since surveys generally include many questions, statistical tests may be used many times to examine the results for different questions/sets of questions; (2) the answers in many surveys, especially surveys dealing with user preferences and opinions, are frequently correlated.

Recall our earlier discussion on experimentwise error. Because we are performing statistical tests on a

number of potentially correlated answers, there is a danger that the experimentwise error in the survey analysis will be inflated. Although many researchers overlook experimentwise error, researchers are encouraged to consider this and, whenever possible, use statistical methods and analysis approaches that manage this error appropriately. For example, in cases where answers to questions are logically and statically correlated, MANOVA tests should be used to examine group differences instead of only doing separate ANOVA analyses for each question. MANOVA tests are designed to control experimentwise error and provide additional insights into group difference on combinations of answers.

Aside from considerations on experimentwise error, the structured data analysis methods presented earlier can be used as described. Refer once more to Figure 2 for assistance in selecting the appropriate methods based on the analysis objectives and data characteristics of the answers. Although the appropriate analyses will depend on the content of the survey collected and the context of the research design, the following examples are presented to generate ideas on how these methods could be used:

- *Compare demographic groups on survey answers*. For example, use gender or age group to group participants; then compare each group's answers using the appropriate method [e.g., *t*-test for gender (two groups) or ANOVA for age group (more than two groups)].
- *Use demographic and other data to model a preference-related response*. For example, in an experiment examining two interfaces, demographic characteristics, previous experience measures, and performance measures could be used to create a logistic regression model to model the likelihood that a participant will prefer one interface to the other.
- *Use correlated preference or opinion responses to develop a new measurement for a complex concept*. For example, the research presented in the case study in Figure 14 used factor analysis to examine questions that addressed various aspects of decision satisfaction and to decompose those answers into factors that capture four underlying dimensions of decision satisfaction.

## 7 CONCLUSIONS

As demonstrated in this chapter, a number of factors influence which outcome data to gather in human factors research and how to analyze and draw conclusions from those outcomes. We first described the characteristics of outcome data and how those characteristics affect how the outcomes may be analyzed and the types of conclusions that can be drawn from the analysis and outcomes. Next, we examined a number of methods for analyzing both structured outcomes, often using statistical methods, and unstructured outcomes, which require less concrete analysis methods. The discussion of both the outcome characteristics and analysis methods has stressed how both of these relate to the research

objectives through the types of conclusions they support. By applying this information, human factors researchers can ensure that the outcome data they collect and the analysis methods they use produce reliable, valid results that support their research goals.

## REFERENCES

Alreck, P. L., and Settle, R. B. (2003), *The Survey Research Handbook*, 3rd ed., McGraw-Hill/Irwin, Boston.

Argyrous, G. (2000), *Statistics for Social and Health Research with a Guide to SPSS*, Sage, London.

Bollen, K. A. (1989), *Structural Equations with Latent Variables*, Wiley, New York.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978), *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, Wiley, New York.

Carroll, R. J., and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, Chapman and Hall, London.

Chapman, G. B., and Elstein, A. S. (2000), "Cognitive Processes and Biases in Medical Decision Making," in *Decision Making in Health Care*, G. B. Chapman and F. A. Sonnenberg, Eds., Cambridge University Press, Cambridge, pp. 183–210.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Lawrence Erlbaum Associates, Mahwah, NJ.

Devellis, R. F. (2003), *Scale Development: Theory and Applications*, 2nd ed., Sage, Thousand Oaks, CA.

Edwards, W., and Newman, J. R. (1982), *Multiattribute Evaluation*, Sage, Beverly Hills, CA.

Ericsson, K. A. (2006), "Protocol Analysis and Expert Thought: Concurrent Verbalizations of Thinking during Experts' Performance on Representative Tasks," in The Cambridge Handbook of Expertise and Expert Performance, K. A. Ericcson, N. Charness, P. J. Feltovich, and R. R. Hoffman, Eds., Cambridge University Press, Camridge, New York, pp. 223–241.

Fava, J. L., and Velicer, W. F. (1992), "An Empirical Comparison of Factor, Image, Component, and Scale Score," *Multivariate Behavioral Research*, Vol. 27, No. 3, pp. 301–322.

Field, A. (2005), *Discovering Statistics Using SPSS,* 2nd ed., Sage, London.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1988), "Knowing What You Want: Measuring Labile Values," in *Decision Making: Descriptive, Normative and Prescriptive Interactions*, D. E. Bell, H. Raiffa, and A. Tversky, Eds., Cambridge University Press, Cambridge, pp. 398–421.

Fryback, D. G. (1998), "Methodological Issues in Measuring Health Status and Health-Related Quality of Life for Population Health Measures: A Brief Overview of the 'HALY' Family of Measures," in *Summarizing Population Health: Directions for the Development and Application of Population Metrics*, M. J. Field and M. R. Gold, Eds., National Academy Press, Washington, DC.

Fuller, W. A. (1987), *Measurement Error Models*, Wiley, New York.

Gerson, S. J., and Gerson, S. M. (2007), *Technical Communication: Process and Product,* 6th ed., Prentice Hall, Upper Saddle River, NJ.

Ghiselli, E. E., Campbell, J. P., and Zedeck, S. (1981), *Measurement Theory for the Behavioral Sciences*, W. H. Freeman, San Francisco.

Gillan, D. J., Wickens, C. D., Hollands, J. G., and Carswell, C. M. (1998), "Guidelines for Presenting Quantitative Data in HFES Publications," *Human Factors,* Vol. 40, No. 1, pp. 28–41.

Grayson, K., and Rust, R. (2001), "Interrater Reliability," *Journal of Consumer Psychology*, Vol. 10, No. 1–2, pp. 71–73.

Guadagnoli, E., and Velicer, W. F. (1988), "Relation of Sample Size to the Stability of Component Patterns," *Psychological Bulletin*, Vol. 103, No. 2, pp. 265–275.

Hughes, M. A., and Garrett, D. E. (1990), "Intercoder Reliability Estimation Approaches in Marketing: A Generalizability Theory Framework for Quantitative Data," *Journal of Marketing Research*, Vol. 27, No.2, pp. 185–195.

Johnson, D. E. (1998), *Applied Multivariate Methods for Data Analysts*, Duxbury, Pacific Grove, CA.

Johnson, R. A., and Wichern, D. W. (2007), *Applied Multivariate Statistical Analysis*, 6th ed., Prentice Hall, Englewood Cliffs, NJ.

Kalton, G., and Stowell, R. (1979), "A Study of Coder Variability," *Applied Statistics*, Vol. 28, No. 3, pp. 276–289.

Keeney, R. L., and Raiffa, H. (1976), *Decisions with Multiple Objectives*, Wiley, New York.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971), *Foundations of Measurement*, Vol. 1, *Additive and Polynomial Representations*, Academic, San Diego, CA.

Krippendorff, K. (2003), *Content Analysis*, 2nd ed., Sage, London.

Kujala, S. (2003), "User Involvement: A Review of the Benefits and Challenges," *Behaviour & Information Technology*, Vol. 22, No. 1, pp. 1–16.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2004), *Applied Linear Statistical Models*, 5th ed., McGraw-Hill/Irwin, Boston.

Liu, Y., and Salvendy, G. (2009), "Effects of Measurement Errors on Psychometric Measurements in Ergonomics Studies: Implications for Correlations, ANOVA, Linear Regression, Factor Analysis, and Linear Discriminant Analysis," *Ergonomics*, Vol. 52, No. 5, pp. 499–511.

Llewellyn-Thomas, H. A., McGreal, M. J., and Thiel, E. C. (1995), "Cancer Patients' Decision Making and Trial-Entry Preferences: The Effects of 'Framing' Information about Short-Term Toxicity and Long-Term Survival," *Medical Decision Making*, Vol. 15, No. 1, pp. 4–12.

Lomax, R. G. (2007), *An Introduction to Statistical Concepts,* 2nd ed., Routledge Academic, New York.

Mazur, D. J., and Hickam, D. H. (1994), "The Effect of Physicians' Explanations on Patients' Treatment Preferences," *Medical Decision Making*, Vol. 14, No. 3, pp. 255–258.

Meister, D. (2004), *Conceptual Foundations of Human Factors Measurement*, Lawrence Erlbaum Associates, Mahwah, NJ.

Merriam-Webster (2010), "Merriam-Webster Online," available: http://www.merriam-webster.com/dictionary/, accessed September 29, 2010.

Newton, R. R., and Rudestam, K. E. (1999), *Your Statistical Consultant*, Sage, Thousand Oaks, CA.

Oppenheim, A. N. (1992), *Questionnaire Design, Interviewing, and Attitude Measurement*, Pinter, London.

Reep, D. C. (2010), *Technical Writing: Principles, Strategies, and Readings*, 8th ed., Longman, White Plains, NY.

Sainfort, F. C., and Booske, B. C. (2000), "Measuring Post-Decision Satisfaction," *Medical Decision Making,* Vol. 20, No. 1, pp. 51–61.

Scott, A. (2002), "Identifying and Analysing Dominant Preferences in Discrete Choice Experiments: An Application in Health Care," *Journal of Economic Psychology,* Vol. *23*, No. 3, pp. 383–398.

Sheskin, D. J. (2007), *Handbook of Parametric and Nonparametric Statistical Procedures*, 4th ed., Chapman and Hall/CRC Press, Boca Raton, FL.

Slovic, P. (1995), "The Construction of Preference," *American Psychologist*, Vol. 50, No. 5, pp. 364–371.

Sprent, P., and Smeeton, N. C. (2007), *Applied Nonparametric Statistical Methods*, 4th ed., Chapman & Hall/CRC Press, Boca Raton, FL.

von Winterfeldt, D., and Edwards, W. (1986), *Decision Analysis and Behavioral Research*, Cambridge University Press, Cambridge.

Wickens, C. D., Lee, J., Liu, Y. D., and Gordon-Becker, S. (2003), *Introduction to Human Factors Engineering*, 2nd ed., Prentice Hall, Upper Saddle River, NJ.

Wixon, D. (1995, October), "Qualitative Research Methods in Design and Development," *ACM Interactions*, pp. 19–24.

Wu, C. F. J., and Hamada, M. (2009), *Experiments: Planning, Analysis, and Parameter Design Optimization*, 2nd ed., Wiley, New York.

# PART 8
# HUMAN−COMPUTER INTERACTION

**Kevin B. Bennett, Allen L. Nagy, and John M. Flach**
**Wright State University**
**Dayton Ohio**

## 1 INTRODUCTION

Advances in computer science and artificial intelligence currently provide new forms of computational power with the potential to support human problem solving. One use of this computational power is to provide an expert system or an automatic assistant that provides advice to the human operator at the appropriate times. For example, there has been some progress in the use of production systems and neural networks as the drivers for decision support. An alternative, complementary use is to integrate information graphically (or more generally, "perceptibly"). Here computational power is used to create and manipulate representations of the target world rather than to create autonomous machine problem solvers (e.g., representation aiding: Zachary, 1986; Woods and Roth, 1988; Woods, 1991).

More generally, this alternative use of computational resources constitutes interface design or human computer interaction. Effective interfaces provide a very real potential to improve overall performance of human–machine systems. The technologies needed to produce interfaces are mature, and when designed properly, they will maintain the flexibility of the human in the loop and improve the capability of the overall system to respond to unforeseen circumstances. The challenge in providing effective interfaces centers around how best to use these technological capabilities to support human decision making and problem solving. Although this chapter focuses on concepts and principles of design to meet this challenge, we see interface design and machine problem solvers as complementary tools in the designer's tool chest and we expect that for very complex systems both approaches will be necessary.

The design of effective interfaces turns out to be a surprisingly difficult challenge, a fact that is attested to by the often-frustrating experiences we all encounter in today's digital world. Effective interface design requires a deep appreciation of human capabilities with regard to both perception (i.e., displays) and action (i.e., controls). It also requires a deep appreciation of the work domain itself: interface design strategies that are appropriate for one category of domains may very well not be appropriate for another. Elsewhere we have provided a comprehensive treatment of interface design (Bennett and Flach, 2011); here we focus on one facet of interface design (perceptual displays) for one category of domains (law-driven, tightly coupled work domains).

In contrast to most other treatments of display design, we did not provide a "cookbook" of detailed guidelines and recommendations (primarily because they tend to be conflicting and difficult to apply). Instead, we chose to describe a set of general heuristics for display design. Because these heuristics are necessarily abstract, we have made the discussion more concrete by illustrating

them within the context of a simple domain. We describe how the heuristics apply to that domain and annotate our written descriptions with concrete graphic examples. Our goal is to transfer functional knowledge of display design to practitioners.

We begin our discussion with a description of basic physiological, perceptual, and technological considerations in display design. These considerations are the foundation for display design and represent the baseline conditions that must be met for a display to be effective. We next consider four alternative approaches to display design. Each approach emphasizes a different conceptual aspect of the display design puzzle, and each approach has both strengths and weaknesses. A fifth approach is outlined; this approach draws from the strengths of the earlier approaches and incorporates new considerations that are particularly relevant to the design of displays for complex, dynamic domains. We discuss some analytical tools of the approach and illustrate their use in determining the various types of information that are required for a simple domain. We describe alternative displays and discuss how each display provides a specific mapping that emphasizes certain aspects of the domain but de-emphasizes or even eliminates other aspects. We end the chapter by considering the limitations of our discussion and examples and additional challenges for display design.

## 2 PHYSIOLOGICAL, PERCEPTUAL, AND TECHNOLOGICAL CONSIDERATIONS

In this section we consider fundamental aspects of the visual system and visual perception that are relevant for display design. Information on the surface of a display is most often represented by a difference in perceived brightness or a difference in perceived color between the information-carrying stimuli and the background of the display field. This section is concerned primarily with the detection and perceived appearance of these differences. Although this chapter is focused primarily on emissive displays, it is useful to begin by discussing some of the differences between reflective and emissive displays and the implications of these differences for visual perception. Emissive displays, such as the cathode ray tube (CRT), generate the light that is used to produce text, symbols, or pictures that carry information. Reflective displays such as road signs, pages in a textbook, and the speedometer in an automobile do not produce any light but reflect some portion of the light that falls on them. Although emissive displays are much more versatile and flexible in some respects, it is probably safe to say that the use of reflective displays to present information was, and still is, far more common. With regard to the visual system and visual perception, there are some fundamental differences between reflective and emissive displays. We begin by examining properties of achromatic, or colorless, displays that illustrate these differences and later in this section take up chromatic displays.

### 2.1 Reflective Displays

The surface of a reflective display reflects some portion of the light energy that falls on it in many different directions. The percentage of light reflected, known as the *reflectance* of the surface, and the dependence of this percentage on the wavelength of the light, known as the *spectral reflectance function* of the surface, are determined by the physical properties of the surface (Nassau, 1983). We begin by discussing surfaces with flat spectral reflectance functions that reflect approximately the same percentage of light for all wavelengths. Images are placed on the surface by changing the properties of the surface in local regions. For example, suppose that a printer for a personal computer deposits black ink on a gray page so as to form text. The gray page reflects a percentage, perhaps 50%, of the light energy at each wavelength falling on it. The ink deposited on the page appears very dark because it reflects only a small percentage, for example, 5%, of the light energy falling on it. Suppose that an observer views this page tacked to the wall painted uniformly white so that the surface of the wall has a reflectance of 90%.

The reflectance of surfaces varies with the angle of incidence of the illumination and the angle at which the reflectance is measured. Reflectances of surfaces can be described with two components, a specular component and a diffuse component (Shafer, 1985; Hunter and Herold, 1987). The specular component is mirrorlike, in that a large proportion of the light is reflected off at an angle equal to the angle of incidence. The diffuse component is characterized by light reflected off in all directions. Shiny surfaces such as mirrors have a large specular component and a small diffuse component, whereas matte surfaces such as a velvet cloth have a large diffuse component and a small specular component. For simplicity we ignore these complexities here. Figure 1*a* illustrates idealized spectral reflectance curves for the page, the ink, and the wall. Real spectral reflectance curves would only approximate flat curves. Surfaces with flat curves are neutral in the sense that they do not change the spectral quality of the light that falls on them.

To characterize the light reflected back from the surface, we need to know something about the light falling on the surface. A typical spectrum for sunlight is shown in Figure 1*b*, where the relative energy is plotted as a function of wavelength. This spectrum is referred to as typical because the spectrum for sunlight varies with time of day, time of year, latitude, and atmospheric conditions. Not all of the energy in sunlight is effective in generating a visual response. Some wavelengths of light are more likely than others to be absorbed by the receptors in the eye, the rods and cones. A function describing the relative effectiveness of different wavelengths for photopic or cone vision (Figure 2) was standardized by the International Commission on Illumination (CIE) in 1924 (Wyszecki and Stiles, 1982). This function, known as the *photopic luminosity function*, has served as a standard in science and industry ever since.

A similar function for scotopic or rod vision was standardized in 1951 (Wyszecki and Stiles, 1982). Since most displays are viewed under photopic conditions, we concentrate on cone vision here. To get a measure of the visual effectiveness of the light energy from the sun, we multiply the energies at each wavelength in Figure 1*b*

**Figure 2** CIE 1924 photopic luminosity curve.
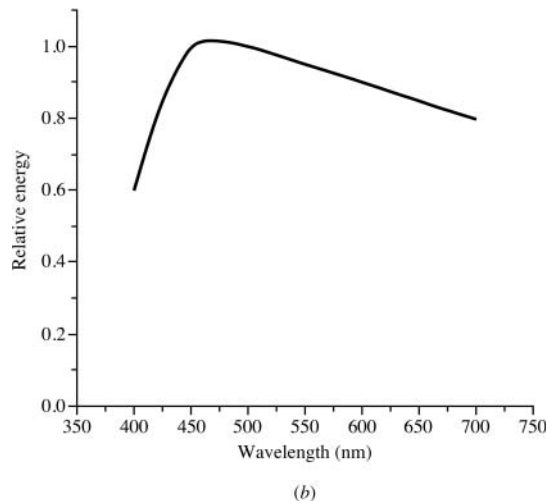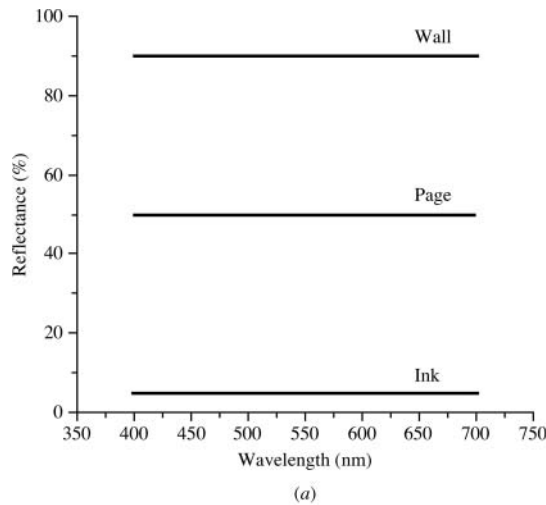
**Figure 1** (*a*) Idealized spectral reflectance curves for the ink, the page, and the wall in the example described in the text; (*b*) relative energy at each wavelength in sunlight.

by the value of the photopic luminosity function at that wavelength. The sum or integral of these weighted energies, multiplied by a constant to convert the energy units to a convenient unit of visual effectiveness, is known as the *luminance* of the source. A commonly used unit for luminance today is the candela per square meter ($cd/m^2$).

For our purposes, the more important measure is the amount of light that actually falls on the wall, the page, and the ink. This quantity is known as *illuminance*, the amount of visually effective light that actually falls on a surface in space. We assume that the wall is evenly illuminated so that this measure is the same across the wall, the text, and the page. A common unit of illuminance is the lux. The measurement of luminance, and the related quantity illuminance, is itself a complex

topic, and many different types of units are used in measuring light. [For discussions of light measurement, see Grum and Bartleson (1980) and Wyszecki and Stiles (1982).] To find the amount of visually effective light reflected from the surface, we multiply the reflectance at each wavelength by the illuminance provided by the sunlight at each wavelength. Alternatively, we could measure directly the amount of visually effective light reflected in a particular direction using a device called a photometer. [For a discussion of devices for measuring light, see Post (1992).]

An important property of reflective displays, such as our page of printed text mounted on the wall, is that the physical contrast between the text and the page, or the page and the wall, does not vary with the amount of light falling on them as long as all of the surfaces are illuminated at the same level. The term *physical contrast* is used to refer to the difference in the light reflected from two regions of a scene. The physical contrast of a stimulus on a background is often defined as the *contrast ratio*, $\Delta L/L$, the difference between the light reflected from the stimulus and the background divided by the background level. In our example the physical contrast between the text and the page could be specified as the difference in the amounts of light reflected by the ink and by the page divided by the amount of light reflected by the page. Note that as the amount of light falling on the wall is changed, the physical contrast ratios calculated for the text and the page, the text and the wall, and the page and the wall will remain constant (Figure 3). The reader can demonstrate this by setting up the contrast ratios and demonstrating that the light level, which appears in both the numerator and the denominator of the contrast ratio, will cancel out, and the contrast ratios are determined by the reflectances alone.

The human visual system appears to have evolved to take advantage of the reflective properties of surfaces. One of the earliest relationships established in the study of visual perception is that the intensity difference between a stimulus and a background necessary for

Contrast ratios:

$$\text{Text / page} = \frac{0.5\,I - 0.05\,I}{0.5\,I} = \frac{0.45}{0.50} = 0.90$$

$$\text{Page / wall} = \frac{0.9\,I - 0.5\,I}{0.9\,I} = \frac{0.40}{0.90} = 0.44$$

$$\text{Text / wall} = \frac{0.9\,I - 0.05\,I}{0.9\,I} = \frac{0.85}{0.90} = 0.94$$
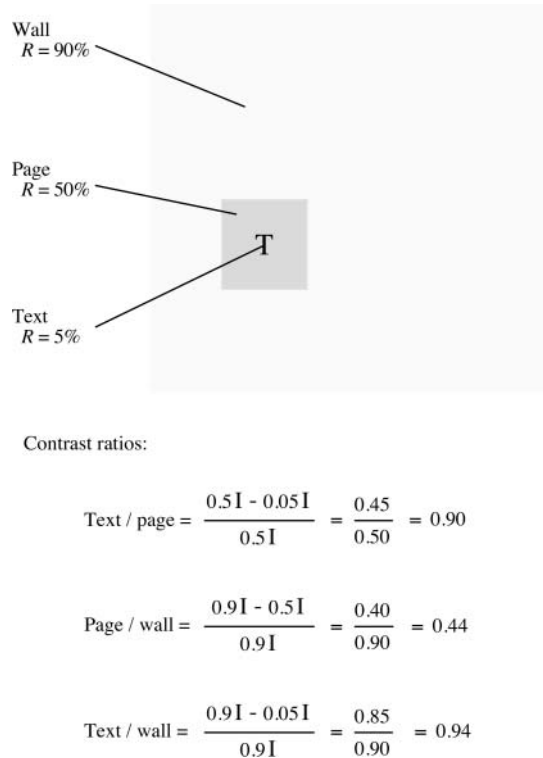
**Figure 3**   Calculation of contrast ratios for the page, the text, and the wall. The values of *R* indicate the reflectances of the three surfaces in the figure. The illumination level *I* is identical for all three surfaces in the figure and therefore cancels out of the equations.

detection of the stimulus is a constant proportion of the intensity of the background field. This rule, known as *Weber's law*, is often written in equation form as $\Delta I = kI$. Here $\Delta I$ refers to the difference between the intensity of the stimulus and the intensity of the background, $k$ is the proportionality constant or the Weber fraction, and $I$ is the intensity of the background field. Weber's law indicates that the visual system becomes less sensitive to differences between the stimulus and the background as the intensity of the background field increases. That is, in order to keep the stimulus detectable, the difference between the stimulus and the background is increased. Notice, however, that if we rearrange Weber's law by dividing both sides of the equation by $I$, we get $(\Delta I/I) = k$.

At threshold, the difference between the intensities of the stimulus and the background ($\Delta I$) divided by the background intensity ($I$) is constant. This is exactly the situation for the reflective displays described above. It means that if the text on a page is detectable at any light level, it will remain detectable as the light level is changed. A somewhat different form of Weber's law also applies to the discrimination of two stimuli presented on a background. In this case, at

threshold the difference in the contrasts between the two stimuli relative to the background must be a constant proportion of one of the contrasts (Whittle, 1986, 1992; Webster et al., 2005; Nagy and Kamholz, 1995). Thus, for reflective displays, if two stimuli at different contrast levels on a background are discriminable from each other, they will remain discriminable as the illumination level is changed. It is well known that Weber's law is only approximately true and that it breaks down under many conditions, perhaps most important, when the light levels involved are low and approach absolute threshold. However, the change in the sensitivity implied by Weber's law is an important property of the visual system. It is a component of another property of the visual system known as *lightness constancy*, which refers to the fact that the visual system operates in such a manner as to keep the perceived appearance of reflective objects approximately constant under changing illumination levels. That is, the wall, the page, and the text in our example appear white, gray, and black, respectively, whether they are viewed outdoors under intense sunlight or indoors under dim illumination. Lightness constancy depends on many other factors in addition to the change in sensitivity indicated by Weber's law and has been a topic of intense interest in the last couple of decades (Gilchrist et al., 1983; Adelson, 1993).

## 2.2   Emissive Displays

We will use a CRT as an example of an emissive display. CRTs generate light by shooting beams of electrons at substances called *phosphors* which are painted on the screen of the CRT. When the electrons hit a point on the screen, light energy is given off by the phosphor at that point. The intensity of the light given off can be changed by varying the strength of the beam of electrons directed at the point. Images are created on the screen by varying the intensity of the electron beam hitting different points on the screen. The physical contrast between different regions of the screen can be defined in the same manner as for reflective displays.

Suppose that we mount the CRT on the white wall and use it to generate a page of dark text on a gray page. Suppose also that we adjust the CRT so that the page gives off 50 units of light and the text gives off 5 units of light. The physical contrast ratio between the text and the page would be 0.90, as it was for the reflective display (see Figure 3). Suppose that the white wall is illuminated initially so that 90 units of light are reflected from it. Also suppose for the moment that the surface of the CRT reflects none of this light. In this case the contrast ratios between the three surfaces would be the same as in our first example with the reflective page of text, and we might expect that the CRT display would look very similar to the reflective display.

Note what happens as the illumination falling on the wall is increased, however. The intensity of the light reflected from it increases, but the intensities of the lights from the text and the page on the CRT do not change. The contrast ratio between the text and the page on the CRT remains constant, but the contrast ratios between the page and the wall and the text and the wall

increase. Thus, we might expect the appearances of the text and the page to change considerably as the light level falling on the wall is changed. If we regard the text and the page as individual incremental stimuli against the large background provided by the wall, Weber's law suggests that their discriminability will decrease as the light reflected from the wall increases. The decrease in discriminability occurs because the difference in contrast ratios decreases with increasing light level. In this case the decrease in the sensitivity of the visual system with increasing background light level reduces the ability to detect the difference between the text and the page, which remains constant.

Any light that is reflected from the glass face of the CRT will reduce the discriminability of the text on the page even further, because it will be reflected from both the region containing the dark text and the region containing the page. The reflected light actually reduces the physical contrast between the text and the page and makes them even less discriminable. Thus, emissive displays behave quite differently than reflective displays in natural environments. These differences do not present much of a problem when emissive displays are placed in a constant environment such as an office illuminated by a fixed light source. However, when emissive displays are placed in natural environments in which the illumination level may vary by a factor of a million or more, the problems caused by the varying contrast ratios are evident. For example, this problem occurs when emissive displays are used in aircraft. The detectability and the appearance of elements within the display may vary dramatically. To keep the appearance of the text and the page constant, the light levels given off by the CRT must be adjusted in accord with the change in the illumination of the wall.

## 2.3 Factors Affecting Perceived Contrast

Besides the physical contrast, there are many other factors, such as adaptive state, location in the visual field, eye movements, and the interpretation of the perceived illuminant, which affect the perceived contrast of a stimulus against a background. One of the most important of these factors is stimulus size. In the last few decades this problem has been investigated very successfully with an approach based on Fourier analysis. [For extensive reviews and applications, see Ginsburg (1986), Olzack and Thomas (1986), Graham (1989), DeValois and DeValois (1990), Pavel and Ahumada (1997), and Makous (2003).] Fourier analysis suggests that any pattern of light and dark on the retina can be described as a sum of sinusoidal components of different frequency and amplitude. The application of this idea to visual perception involves measuring an observer's sensitivity to a number of sinusoidal patterns of different spatial frequency (Figure 4). These repetitive spatial patterns of light and dark are known as *gratings*.

Spatial frequency is essentially a measure of the size of the bars in the pattern. The spatial frequency of the pattern is defined as the number of cycles that occur in 1° of visual angle. As spatial frequency increases, there are more cycles per degree of visual angle and the bars become smaller. Visual angle is used as the unit of
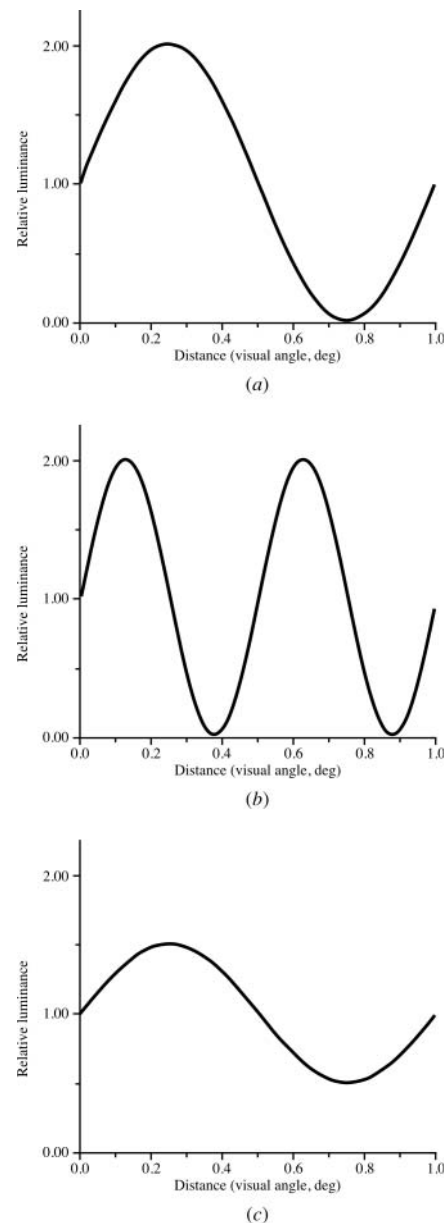


**Figure 4** Variation in luminance for sinusoidal patterns: (*a*) spatial frequency of 1 cycle/degree at a contrast of 100%; (*b*) spatial frequency of 2 cycles/degree at a contrast of 100%; (*c*) spatial frequency of 1 cycle/degree at a contrast of 50%.

size because it gives a measure of the size of the image on the retina (e.g., a book 12 in. long makes a larger image on the retina when it is held up close to the eye than when it is held far away). To get a measure of the size of an image on the retina, the distance between an object and the observer's eye must be considered. Thus, the visual angle subtended by an object is defined as
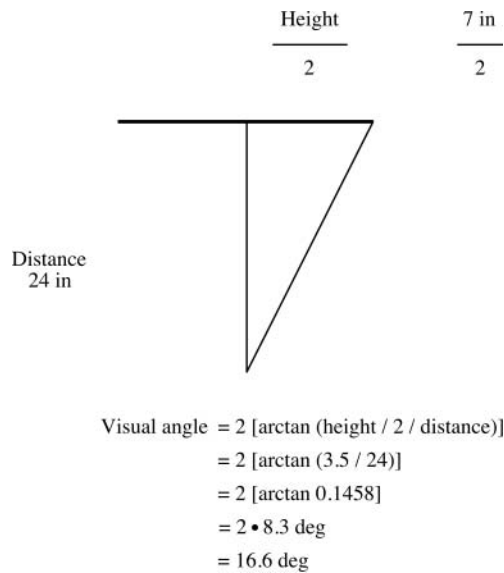
**Figure 5**   Calculation of the visual angle.



**Figure 6**   Typical plot of a contrast sensitivity function for a human observer. (Based on data from DeValois and DeValois, 1990.)

twice the arctangent of one-half the height divided by the distance (Figure 5).

Sensitivity is measured by finding the physical contrast level at which a given pattern of light and dark is just detectable. To give a measure of sensitivity, the reciprocal of the threshold is calculated by dividing 1 by the threshold contrast. The measure of physical contrast typically used in this approach is slightly different from the contrast ratio described above and is called the *Michelson contrast*. It is defined as $L_{max} - L_{min}$ divided by $L_{max} + L_{min}$, where $L_{max}$ is defined as the maximum luminance level in the pattern and $L_{min}$ is defined as the minimum luminance in the pattern. The curve described by plotting contrast sensitivity against the spatial frequency of the grating pattern is called the *contrast sensitivity function*.

A typical contrast sensitivity function for photopic or cone vision obtained from a human observer is shown in Figure 6. The curve shows that when spatial frequency is low (i.e., the bars are large), the sensitivity to contrast is low. As spatial frequency is increased, the sensitivity increases up to spatial frequencies of about 5–10 cycles per degree. With further increases in spatial frequency (i.e., smaller and smaller bars), sensitivity falls off rapidly until at a spatial frequency of approximately 50 cycles per degree a grating of 100% contrast (the highest physical contrast obtainable) is not visible. Spatial patterns of even greater frequency also are not visible. Thus, very fine patterns are visible only if the spatial frequency is below 50 cycles per degree and they are very high contrast.

Over the last few decades many physical factors, such as overall light level, number of cycles present in the pattern, and location of the pattern in the visual field, have been shown to affect the contrast sensitivity function. The shape of the curve as well as the overall
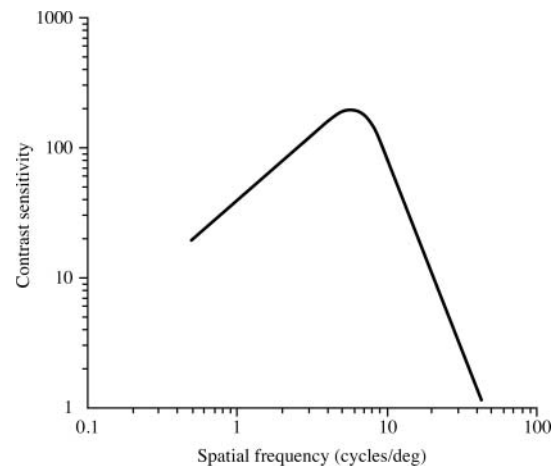
sensitivity can vary considerably. The shape and height of the curve are affected by several components within the visual system that play a role in determining the contrast sensitivity function. For example, the optics of the eye, the lens and cornea, which form an image of the pattern on the retina, influence the contrast sensitivity function, because they do not form a perfect image of the external pattern on the retina. A good introductory treatment of the optics of the eye is given by Millodot (1982). The distribution of rods and cones on the retina also plays a role in determining the contrast sensitivity function. The rods and cones absorb light and initiate neural signals in the visual system. Thus, their size and the distances between them have some effect on the contrast sensitivity function. A good introduction to the sampling properties of rods and cones is given by Wandell (1995). The way the rods and cones are connected to the neurons that carry signals out of the eye also plays a role in determining the contrast sensitivity function, because many receptors are connected to each neuron. Psychophysical evidence suggests that the visual system may be organized into approximately five to seven neural channels, each sensitive to a different band of spatial frequencies (Olzack and Thomas, 1986). Thus, the contrast sensitivity function is the result of many factors which have been studied intensely over the last few decades. Nevertheless, it is a very useful and fundamental description of the ability of a human observer to detect contrast in patterns of different size. For example, recent studies suggest that the recognition of text may be mediated by the same mechanisms that mediate the contrast sensitivity function (Alexander et al., 1994; Solomon and Pelli, 1994).

The perceived contrast of patterns that are well above threshold is not simply related to the contrast sensitivity function (Cannon and Fullenkamp, 1991). That is, if we measure the threshold contrast for sinusoidal patterns at a number of different spatial frequencies and then increase the physical contrast of all of these patterns so

that the contrast for each one is five times the threshold contrast, the patterns will not appear to have equal contrasts. This is similar to the situation in audition where equal-loudness curves for tones of different frequencies do not have the same shape as the audibility curve, a plot of threshold as a function of frequency, and change shape as the loudness level is raised. Thus, the contrast sensitivity function can be used to predict whether a pattern of a given spatial frequency is visible, but it cannot be used to predict accurately the perceived contrast of patterns that are well above threshold. For example, if a display designer wants to equate the perceived contrast of patterns of different size that are well above threshold, the contrast sensitivity function cannot be used to do this accurately.

The notions of visual angle, spatial frequency, and contrast sensitivity that were introduced briefly above are very useful in thinking about both reflective and emissive displays. Here we concentrate on emissive displays. Consider a standard CRT display that is 9.5 in. wide and 7 in. high. Assume that this CRT has 640 columns of pixels, each containing 480 rows (standard $640 \times 480$ resolution). If the observer views this display from a distance of 2 ft, the screen subtends about $22.4°$ horizontally and $16.6°$ vertically (see Figure 5), and each pixel subtends about $0.035°$. If we want to make patterns of light and dark bars on the screen, we might want to know the highest spatial frequency that can be represented. If we make alternate pixels black and white, we need two pixels to make one cycle, which will subtend $0.07°$. Thus, the highest spatial frequency that can be represented accurately will be 1/0.07, or slightly over 14 cycles per degree.

Looking back at our representative contrast sensitivity function, we see that this frequency is well below the upper limit of approximately 50 cycles per degree. Looking at the vertical axis, we find that the sensitivity at 14 cycles per degree is approximately 30. For an observer to detect this pattern on the screen, we can determine that the Michelson contrast will have to be approximately 1/30, or 3.3%. These calculations also tell us something else. Patterns with spatial frequencies higher than 14 cycles per degree just cannot be represented accurately on the monitor. Thus, if we want to view an image with a lot of fine details at high spatial frequencies, such as a digitized photograph that subtends $9.5 \times 7$ in., spatial frequencies greater than 14 cycles per degree that were visible when the original photograph was viewed from a distance of 2 ft will not be represented accurately on the monitor if they are composed of spatial frequencies above 14 cycles per degree.

One solution to this problem is to use a monitor with higher resolution or smaller pixels. For example, if we could pack $1280 \times 960$ pixels into the same $9.5 \times 7$-in. screen, patterns with spatial frequencies up to nearly 29 cycles per degree could be represented. To make a display that matches the upper limit on the resolution of the visual system, we would need to pack about $2240 \times 1660$ pixels into the display. A $9.5 \times 7$-in. CRT with this resolution would permit the presentation of patterns with spatial frequencies up to 50 cycles per degree at a viewing distance of 2 ft. This would be very difficult

to accomplish with the present technology, making the display and the computer hardware that drives it very expensive.

It is also possible to portray patterns with spatial frequencies greater than 14 cycles per degree on the original CRT by moving the observer farther away so that each pixel subtends a smaller visual angle. The drawback to this approach is that the entire display field now subtends a smaller portion of the field of view. For example, if we move the observer back to a distance of about 4 ft, patterns with spatial frequencies up to nearly 29 cycles per degree could be portrayed on the screen. This example helps to illustrate a fundamental trade-off in emissive displays, the trade-off between field of view and resolution. With a fixed number of pixels, this trade-off is always present in an emissive display. If the pixels are spread over a larger viewing area, the resolution will be poor. If they are packed into a smaller viewing area, the resolution will improve but the field of view will decrease.

The resolution of an emissive display may be limited either by the display itself or by the hardware that drives it: that is, the video card in a computer or the signals generated on a television cable. The detail in an image, or the spatial frequencies that can be portrayed, and the field of view that is visible will be limited by this resolution and the size of the screen.

## 2.4  Color

Although black-and-white pictures carry much of the information in the real world, they do not carry information about color. Color in images is certainly important for aesthetic reasons, but in addition to the aesthetic qualities it brings to an image, color serves two important basic functions (Boynton, 1990). First, chromatic contrast between two regions in an image can add to the luminance contrast between these regions to make the difference between the regions much more noticeable, especially when the luminance contrast is small. Second, since color is perceived to be a property of an object (although, in fact, it also depends on illumination, as we will see), it is useful in identifying objects, searching for them, or grouping them. Boynton (1990) regards the second function of color, which he describes as related to categorical perception, as the more important one. It is probably because of these categorical properties that color is often used as a coding device and as a means of segregating information in visual displays (see Widdel and Post, 1992).

Several excellent treatments of the basics of human color vision and the science of specifying colors for applications are available (e.g., Wyszecki and Stiles, 1982; Pokorny and Smith, 1986; Travis, 1991; Post, 1992, 1997; Kaiser and Boynton, 1997; Gegenfurtner and Sharpe, 1999; Nagy, 2003; Hansen and Gegenfurtner, 2006; Eskew, 2009), so a very brief review will be given here. Normal human color vision depends on the presence of three types of cone receptors in the retina. These cones differ in the type of light-absorbing pigment contained in them. One of these pigments absorbs best, meaning the greatest percentage of the light falling on it, in the short-wavelength region of the spectrum;

hence the cone containing it is referred to as the *S cone*. The second pigment absorbs best in the middle of the spectrum, and the cone containing it is referred to as the *M cone*. The third pigment absorbs best at slightly longer wavelengths than the M pigment and the cone containing it is referred to as the *L cone*.

The differences in the signals generated in these cones by a given light provide some information about the spectral content of the light. For example, a light source that gives off more energy in the long-wavelength portion of the spectrum than in the middle- or short-wavelength regions would tend to stimulate the L cones more than the other two cone types. On the other hand, a light source that gives off more energy in the short-wavelength region would tend to stimulate the S cones more than the other two types. The differences in the stimulation of the cone types serve as a means for discriminating between the lights and result in the perception of color.

Since there are only three types of cones, normal human color vision is said to be *three-dimensional* or *trichromatic*. Furthermore, since there are only three signals from different types of cones in the visual system, it follows that only three numbers are needed to specify the perceptual quality of a color. Much effort has gone into developing systems of specifying colors with three numbers such that they represent the perceptual qualities of the stimulus in useful ways. The fact that only three numbers are needed to specify the chromatic quality of a stimulus also means that there are many physically different stimuli that stimulate the three cones in the same way and thus appear to be the same color. Stimuli that are physically different but appear to be the same are called *metamers*.

Consider the reflective display example given above. Suppose that we print the text on our gray page using red ink rather than black ink. The ink appears red because it tends to absorb short- and middle-wavelength light that falls on it while reflecting long-wavelength light. A spectral reflectance curve showing the percentage of light reflected as a function of wavelength for red ink might look like the curve shown in Figure 7. To get the light reflected back from the ink, we multiply the reflectance at each wavelength by the energy at each wavelength. To calculate the luminance of this light, we would weight the reflected energy at each wavelength by the photopic luminosity function and integrate or sum over the entire curve as we did for achromatic stimuli above. However, the text appears to differ from the gray page and the white wall in color as well as in lightness. To characterize this difference, we would like some means of measuring the colors of the text, the page, and the wall. The most widely used system for doing this is based on the CIE 1931 chromaticity diagram. This diagram is based on color matches of normal human observers. A good introduction to the color-matching experiment and the development of chromaticity diagrams can be found in Kaiser and Boynton (1997). In the color-matching task, observers were asked to adjust the intensities of three primary lights that were mixed together in a single stimulus field so as to match the colors of a wide variety of other lights presented in another
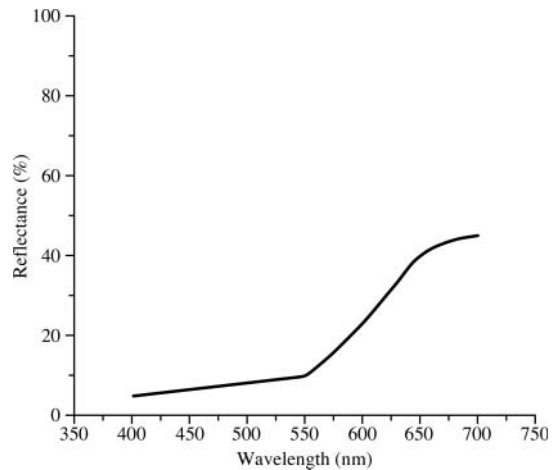


**Figure 7**    Spectral reflectance curve for red ink.

stimulus field. The CIE chromaticity diagram uses three numbers (related to the intensities of the primaries needed to make a match in the color-matching experiment) to represent the color or, more specifically, the chromaticity of a stimulus. These numbers are called the *chromaticity coordinates* of the color and are referred to as $x$, $y$, and $z$. The color-matching data were normalized so that the values of these three chromaticity coordinates add up to 1 for any real color. As a result, only two of the chromaticity coordinates need to be given to specify a color, because the third can always be obtained by subtracting the sum of the other two from 1. Therefore, all colors can be represented in a two-dimensional diagram such as the CIE 1931 diagram shown in Figure 8, where only $x$ and $y$ are plotted. Many measuring instruments
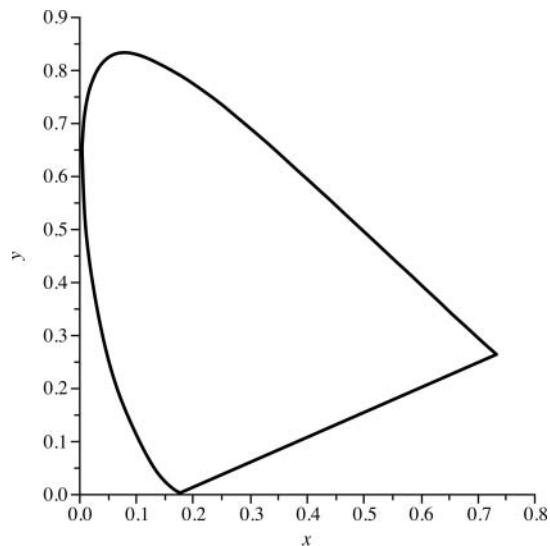


**Figure 8**    CIE 1931 chromaticity diagram. (Based on data from Wyszecki and Stiles, 1982.)

have been developed and are commercially available for measuring the CIE coordinates of a color. [See Post (1992) for some discussion of these.]

The chromaticity coordinates specify the chromatic properties of a color but do not specify its appearance, because the appearance of the color can change with many viewing conditions that do not change its chromaticity coordinates. For example, the size of the stimulus, in terms of visual angle, can affect the color appearance even though the chromaticity coordinates of the ink used to make it do not change (Poirson and Wandell, 1993). This is a severe limitation on the meaning and usefulness of the CIE chromaticity diagram. One would like to have a system in which the appearance of the color is specified, but this is a very difficult problem that has not yet been solved. Nevertheless, the specification of colors in the chromaticity diagram is still very useful, because any two stimuli with the same chromaticity coordinates will appear to be identical in color when viewed under the same conditions. What the chromaticity coordinates specify is how to make a color that will appear the same as a given sample under the same viewing conditions.

The chromaticity coordinates of a reflective display change with the chromaticity of the light used to illuminate it. The change occurs because the amount of light reflected back from an object at each wavelength depends in part on the amount of light falling on it. Therefore, when the chromaticities of objects, or dyes, or paints are specified, they are usually given with reference to a standard light source. [For a discussion of standardized light sources, see Wyszecki and Stiles (1982).] One might expect that the change in the chromaticity coordinates accompanying a change in the light source would change the color appearance of a reflective display. Such changes in light source are actually quite common. As noted above, the spectral quality of daylight changes with time of day, atmospheric conditions, season, and location on Earth. A large variety of artificial light sources are commercially available, and these can differ considerably in the spectral quality of the light given off. However, these changes do not generally result in large changes in the appearances of objects, because mechanisms within the visual system act to maintain a constant color appearance despite these changes in illumination. Color constancy has generally been shown to be less than perfect (Arend and Reeves, 1986; Brainard and Wandell, 1992). However, it appears to work well enough to prevent confusing changes in the appearance of reflective objects. The visual mechanisms mediating color constancy have been of intense interest over the past few decades (D'Zmura and Lennie, 1986; Maloney and Wandell, 1986; Lennie and Movshon, 2005). Selective adaptation within the three cone mechanisms is thought to be one of the major mechanisms mediating color constancy (Worthy and Brill, 1986) much as the change in sensitivity described by Weber's law plays a role in lightness constancy.

Although mechanisms of color constancy work to maintain a constant appearance in reflective displays, they actually work against the maintenance of a constant appearance in emissive displays, much as mechanisms of lightness constancy worked against the constant appearance of black-and-white emissive displays. Color CRTs take advantage of the fact that human color vision is trichromatic by using only three different phosphors. Each phosphor emits light of a different color when it is stimulated. The light from the three phosphors is mixed together in different proportions to give all other colors, including white.

The chromaticity of a color produced on an emissive display does not change with changes in the illumination of the surroundings. Thus, the mechanisms of color constancy, activated by changes in the illumination of the surroundings, introduce changes in the appearance of these chromaticities, which may be quite noticeable to the observer. Under some conditions these changes in appearance may be large enough to cause some confusion in identifying objects on the basis of color.

### 2.4.1 Factors Affecting Perceived Color Contrast

Much as the perception of achromatic contrast is affected by many factors, color contrast is affected by many factors, such as light level, adaptive state, location in the visual field, and stimulus size. The spatial frequency approach has also been applied to the detection of color contrast. It is possible to produce grating patterns which vary sinusoidally in color, with little or no variation in luminance. The color contrast between the bars of the grating required for detection of the pattern can be measured as a function of the spatial frequency (Kelly, 1974; Noorlander and Koenderink, 1983; Mullen, 1985; Sekiguchi et al., 1993). Typical results for red/green and yellow/blue gratings are shown in Figure 9. Comparison of the results for chromatic patterns with those shown for luminance patterns reveals clear differences. Sensitivity to color contrast is high at low spatial frequencies but begins to fall off dramatically
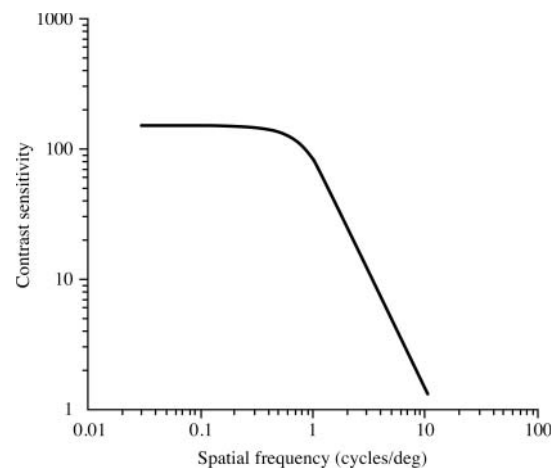


**Figure 9** Typical plot of contrast sensitivity for isoluminant chromatic gratings. (Based on data from Mullen, 1985.)

at rather low spatial frequencies compared to luminance contrast. Above spatial frequencies of approximately 12 cycles per degree color contrast is not detectable even at the highest color contrasts producible. Thus, chromatic contrast information is limited to fairly low spatial frequencies, or large patterns, as compared to luminance contrast information. Within this range of spatial frequencies, the color appearance of the bars of a pattern that is well above threshold is also affected by spatial frequency (Poirson and Wandell, 1993). As the spatial frequency of the pattern is increased, the apparent color contrast between the bars is reduced. Thus, the detectability of color contrast and the color appearance of stimuli are affected dramatically by stimulus size.

## 3  FOUR ALTERNATIVE APPROACHES TO DISPLAY DESIGN

Earlier we discussed physiological, perceptual, and technological considerations in designing visual displays. This has been the traditional focus for human factors research: to design displays that are legible. For example, the knowledge that a user will be seated a particular distance from a particular type of display under a particular set of ambient lighting conditions can be used to determine the appropriate size and luminance contrast that will be necessary for the characters to be seen. Thus, the previous considerations provide us with an understanding of the baseline conditions of display design that must be met (are necessary) for a person to use a display.

Although these considerations are necessary for the design of effective displays, they are not sufficient. Compliance with these considerations will make the *data* required to complete domain tasks available but may not provide the *information* necessary to support an observer in decision making and action. Woods (1991) makes an important distinction between design for data availability and design for information extraction. Designs that consider only data availability often impose unnecessary burdens on the user: to collect relevant data, to maintain these data in memory, and to integrate these data mentally to arrive at a decision. These mental activities require extensive knowledge and tax-limited cognitive resources (attention, short-term memory) and therefore increase the probability of poor decision making and errors.

Our discussion of display design will begin with a consideration of four broadly defined approaches. Each approach is complementary in the sense that it approaches the display design problem from a different conceptual perspective (i.e., graphical arts, psychophysical, attention-based, and problem solving/decision making).

### 3.1  Aesthetic Approach

Tufte (1983, 1990) reviews the design of displays from an aesthetic, graphic arts perspective. Tufte (1983) describes principles of design for data graphics or statistical graphics which are designed expressly to present quantitative data. One principle is the *data–ink ratio*, a measurement of the relative salience of data versus nondata elements in a graph. It is computed by determining the amount of ink that is used to convey the data and dividing this number by the total amount of ink that is used in the graphic. A higher data–ink ratio (a maximum of 1.0) represents the more effective presentation of information. A second measure of graphical efficiency is *data density*. Data density is computed by determining the number of data points represented in the graphic and dividing this number by the total area of the graphic. The higher the data density, the more effective the graphic. Other principles include eliminating graphical elements that interact (e.g., moire vibration) and eliminating irrelevant graphical structures (e.g., containers and decorations) and aesthetics (e.g., effective labels, proportion, and scale).

The two versions of a statistical graphic that are shown in Figure 10 illustrate several of Tufte's principles. The version in Figure 10*a* is poorly designed; the version in Figure 10*b* is more effectively designed. In Figure 10*b* the irrelevant data container (the box) that surrounds the graph in Figure 10*a* has been eliminated. In addition, several other nondata graphical structures have been removed (grid lines). In fact, these grid lines are made conspicuous by their absence in Figure 10*b*. Together, these manipulations produce both a higher data–ink ratio and a higher data density for the version in Figure 10*b*. In Figure 10*a* the striped patterns on the bar graphs produce an unsettling moire vibration and have been replaced in Figure 10*b* with gray-scale patterns. In addition, the bar graphs in Figure 10*b* have been visually segregated by spatial separation. Finally, the three-dimensional perspective in Figure 10*a* complicates visual comparisons and has been removed in Figure 10*b*.

Tufte (1990) broadens the scope of these principles and techniques by considering nonquantitative displays as well. Topics that are discussed include micro/macro designs (the integration of global and local visual information), layering and separation (the visual stratification of different categories of information), small multiples (repetitive graphs that show the relationship between variables across time or across a series of variables), color (appropriate and inappropriate use of), and narratives of space and time (graphics that preserve or illustrate spatial relations or relationships over time). The following quotations (Tufte, 1990) summarize many of the key principles:

- "It is not how much information there is, but rather, how effectively it is arranged." (p. 50)
- "Clutter and confusion are failures of design, not attributes of information." (p. 51)
- "Detail cumulates into larger coherent structures ... Simplicity of reading derives from the context of detailed and complex information, properly arranged. A most unconventional design strategy is revealed: to clarify, add detail." (p. 37)
- "Micro/macro designs enforce both local and global comparisons and, at the same time, avoid the disruption of context switching. All told, exactly what is needed for reasoning about information." (p. 50)
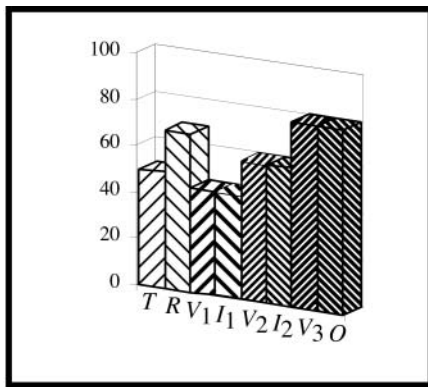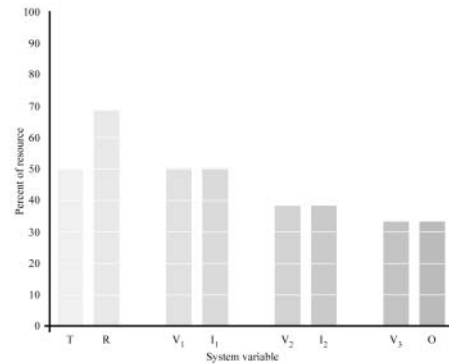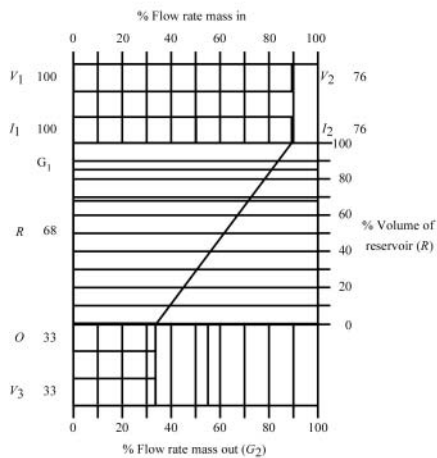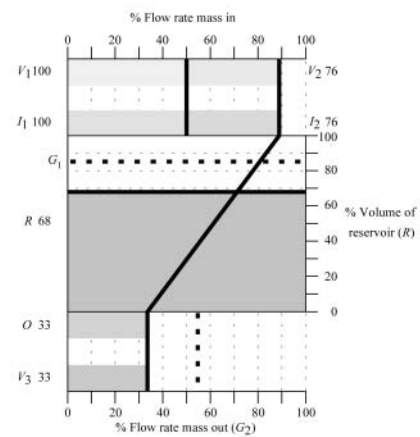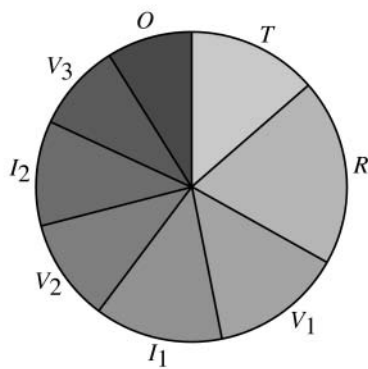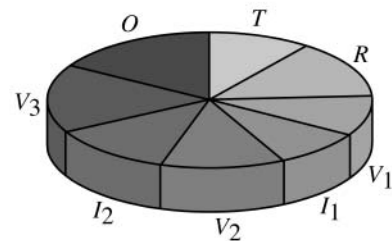
**Figure 10** Six alternative mappings. Parts (*a*) and (*b*) represent alternative versions of a separable (bar graph) graphical format that provide a less effective (*a*) and a more effective (*b*) mapping. Parts (*c*) and (*d*) represent alternative versions of a configural display format that provide a less (*c*) and a more (*d*) effective mapping, due primarily to layering and separation. Parts (*e*) and (*f*) represent the least effective mappings.

- "Among the most powerful devices for reducing noise and enriching the content of displays is the technique of layering and separation, visually stratifying various aspects of the data.... What matters—inevitably, unrelentingly—is the proper relationship among information layers. These visual relationships must be in relevant proportion and in harmony to the substance of the ideas, evidence, and data displayed." (pp. 53–54)

This final principle, layering and separation, is graphically illustrated in Figures 10c and d. These two versions of the same display vary widely in terms of the visual stratification of the information that they contain. In Figure 10c all of the graphical elements are at the same level of visual prominence; in Figure 10d there are at least three levels of visual prominence. The lowest layer of visual prominence is associated with the nondata elements of the display. The various display grids have thinner, dashed lines, and their labels have also been reduced in size and made thinner. The medium layer of perceptual salience is associated with the individual variables. The graphical forms that represent each variable have been gray-scale coded, which contributes to separating these data from the nondata elements. Similarly, the lines representing the system goals ($G_1$ and $G_2$) have been made bolder and dashed. In addition, the labels and digital values that correspond to the individual variables are larger and bolder than their nondata counterparts. Finally, the highest level of visual prominence has been reserved for those graphical elements which represent higher level system properties (e.g., the bold lines that connect the bar graphs). The visual stratification could have been enhanced further through the use of color. The techniques of layering and separation will facilitate an observer's ability to locate and extract information.

To summarize, Tufte (1983, 1990) addresses the problem of presenting three-dimensional, multivariate data on flat, two-dimensional surfaces (focusing primarily on static, printed material) very admirably. He attacks the problem from a largely aesthetic perspective and provides numerous examples of both good and bad display designs that clearly illustrate the associated design principles. Although there are critical aspects of dynamic display design for complex domains that are not considered, the principles can be generalized.

## 3.2  Psychophysical Approach

Cleveland and his colleagues have also developed principles for the design of statistical graphics. However, in contrast to the aesthetic conceptual perspective of Tufte, Cleveland has used a psychophysical approach. As an introduction, consider the following quotation (Cleveland, 1985, p. 229):

When a graph is constructed, quantitative and categorical information is encoded by symbols, geometry, and color. Graphical perception is the visual decoding of this encoded information. Graphical perception is the vital link, the raison d'etre, of the graph. No matter how intelligent the choice of information, no matter how ingenious the encoding of the information, and no matter how technologically impressive the production, a graph is a failure if the visual decoding fails. To have a scientific basis for graphing data, graphical perception must be understood. Informed decisions about how to encode data must be based on knowledge of the visual decoding process.

In their efforts to understand graphical perception Cleveland and his colleagues have considered how psychophysical laws (e.g., Weber's law, Stevens's law) are relevant to the design of graphic displays. For example, psychophysical studies using magnitude estimation have found that judgments of length are less biased than judgments of area or volume. Therefore, visual decoding should be more effective if data have been encoded into a format that requires length discriminations as opposed to area or volume discriminations. Cleveland and his colleagues have tested this and similar intuitions empirically. Their experimental approach was to take the same quantitative information, to provide alternative encodings of this quantitative information (graphs that required different "elementary graphical–perception tasks"), and to test observers' ability to extract the information.

The results of these experiments provided a rank ordering of performance on basic graphical perception tasks: position along a common scale, position along identical, nonaligned scales, length, angle/slope, area, volume, and color hue/color saturation/density (ordered from best to worst performance) (Cleveland, 1985, p. 254). Guidelines for display design were developed based on these rankings. Specifically, graphical encodings should be chosen that require the highest ranking graphical perceptual task of the observer during the visual decoding process. For example, consider the three graphs illustrated in Figures 10b, e, and f. For decoding information contained in Figure 10b, an observer is required to judge position along a common scale (in this case, the vertical extent of the various bar graphs). For Figure 10e the observer is required to judge angles and/or area. Finally, to decode the information in Figure 10f, the observer is required to judge volume (note that because of the three-dimensional representation, angles and area are no longer valid cues). According to the rankings, Cleveland and his colleagues would therefore predict that performance would be best with the bar chart, intermediate with the pie chart, and worst with the three-dimensional pie chart.

## 3.3  Attention-Based Approach

A third perspective on display design is to consider the problem in terms of visual attention and form perception. From this perspective the basic issues in display design include the following. What are the fundamental units of perception? What are the basic types of visual information that are available; what are the relationships between these types of information? How do parts group into wholes? Is the perception of the "parts" of a form secondary to the perception of the

"whole," or vice-versa? What constitutes visual attention and how can it be distributed over parts and wholes? The answers to these questions are relevant to principles of display design since visual displays need to provide different types of information and to support a range of activities.

Specifically, there is a continuum of attention demands that operators might face in complex, dynamic domains. At one end of the continuum are "focused" attention tasks that require selective responses to specific elements in the display. Here a response is contingent on the value of an individual variable (e.g., adjusting vehicular speed in a driving task). At the opposite end of this continuum are "divided" attention tasks that require the distribution of attention across many features that must be considered together to choose an appropriate response. Here a response is contingent on the relationship between a number of variables (e.g., braking or changing lanes when the car in front of you slows down). Thus, tasks can be characterized in terms of the relative demands for selective attention to respond to specific features with specific actions and distributed or divided attention in which multiple display elements must be considered together to choose the appropriate actions.

Attention-based approaches to display design have examined how the design of visual representations can help to meet the cognitive load posed by this continuum of attention demands. Garner (Garner, 1970, 1974; Garner and Felfoldy, 1970) and Pomerantz (Pomerantz et al., 1977; Pomerantz, 1986; Pomerantz and Pristach, 1989) have used the speeded classification task to examine the dimensional structure of stimuli. Carswell and Wickens (1990) have generalized these results by investigating perceptual dimensions that are representative of those found in visual displays. Three qualitatively different relationships between stimulus dimensions have been proposed: separable, integral, and configural (Pomerantz, 1986).

**Separable Dimensions**   A separable relationship is defined by a lack of interaction among stimulus dimensions. Each dimension retains its unique perceptual identity within the context of the other dimension. Observers can attend selectively to an individual dimension and ignore variations in the irrelevant dimension. On the other hand, no new properties emerge as a result of the interaction among dimensions. Thus, performance suffers when both dimensions must be considered to make a discrimination. This pattern of results suggests that separable dimensions are processed independently. An example of separable dimensions are color and shape: The perception of color does not influence the perception of shape, and vice versa.

**Integral Dimensions**   An integral relationship is defined by a strong interaction among dimensions such that the unique perceptual identities of individual dimensions are lost. Integral stimulus dimensions are processed in a highly interdependent fashion: A change in one dimension necessarily produces changes in the second dimension. In their discussion of two integral stimulus dimensions, Garner and Felfoldy (1970, p. 237) state that "in order for one dimension to exist, a level on the other must be specified." As a result of this highly interdependent processing, a redundancy gain occurs. However, focusing attention on the individual stimulus dimensions becomes very difficult, and performance suffers when attention to one (selective attention) or both (divided attention) dimensions are required. An example of an integral stimulus is perceived color: It is a function of both hue and brightness.

**Configural Dimensions**   A configural relationship refers to an intermediate level of interaction between perceptual dimensions. Each dimension maintains its unique perceptual identity, but new properties are also created as a consequence of the interaction between them. These properties have been referred to as *emergent features*. Pomerantz and Pristach (1989, p. 636) describe emergent features in the following fashion: "Basically, emergent features are relations between more elementary line segments, relations that can be more salient to human perception than are the line segments themselves." Using parentheses as our graphic elements will allow us to demonstrate several examples of emergent features. Depending on the orientation, a pair of parentheses can have the emergent features of vertical symmetry, () and)(, or parallelism,)) and ((. The definition offered by Pomerantz and Pristach is overly restrictive, however, since graphical elements other than line segments can produce emergent features. A sampling of other emergent features that can be produced by configural dimensions include colinearity, equality, closure, area, angle, horizontal extent, vertical extent, and good form.

There are two significant aspects of performance with configural dimensions. First, relative to integral and separable stimulus dimensions, there is a smaller divided attention cost, suggesting that performance can be enhanced when both dimensions must be considered to make a discrimination. The second noteworthy aspect of this pattern of results is that there is an *apparent* failure of selective attention. Bennett and Flach (1992) discuss why this failure may be apparent and not inherent; Bennett and Walters (2001) investigate design strategies to overcome potential costs.

### 3.3.1   Attention-Based Principles of Display Design

The "proximity compatibility principle" (PCP) offered perhaps the first set of design guidelines derived from this perspective. The original version of PCP (Barnett and Wickens, 1988; Carswell and Wickens, 1987; Wickens and Andre, 1990) emphasized the role of integral and separable stimulus dimensions, along with the notion of perceptual "objects." PCP predicted an inherent trade-off between displays and tasks. PCP maintained that presenting multiple variables in a single perceptual object (i.e., an object display; see Figure 10*d*) would facilitate performance at divided tasks. This was based on the belief that object displays were composed of integral stimulus dimensions: The global perceptual properties produced by interactions between dimensions

improved divided attention performance. However, the associated loss of unique perceptual identities of the individual dimensions degraded focused performance. The opposite pattern of results was predicted for displays that incorporated separable stimulus dimensions (e.g., a bar graph display with separate and unique representations for each variable; see Figure 10*b*): Focused tasks benefit from the lack of interactions between dimensions; divided tasks suffer from the lack of higher order visual properties that arise from the interactions between dimensions.

In the early 1990s we performed a literature review of the empirical laboratory studies that had been conducted on these issues in display design (Bennett and Flach, 1992). The review indicated that the overall fit between the predictions of PCP and the obtained results was not particularly good. Slightly more than one of three empirical findings (19/54, 35%) were found to support the predictions for divided-attention tasks; less than one of four (7/30, 23%) were consistent with the predictions for focused-attention tasks.

Bennett and Flach (1992) proposed the design principle of "semantic mapping" based on configural stimulus dimensions [as opposed to integral dimensions and perceptual objects; see also Sanderson et al. (1989) and Buttigieg and Sanderson (1991)]. This principle maintains that most representational choices will involve configural stimulus dimensions that produce a hierarchically nested set of emergent features. Because of the unique properties of configural dimensions (see above), an inherent trade-off between display and task is not predicted.

If these emergent features are salient (i.e., they can be picked up easily by the human observer) and if they reflect critical aspects of the task (i.e., the constraints of the work domain), then performance at divided-attention tasks will be enhanced. On the other hand, if the emergent features are not salient or if they are not well mapped to domain constraints, then performance will be degraded. This is true whether the representational format is a geometric form, a collection of bar graphs, a point in space, or any other representational form that could be devised.

Designing displays to support focused-attention tasks involves isomorphic considerations. Specifically, performance depends upon the quality of very specific mappings between the task, the visual properties of the display, and the perceptual abilities of the agent. For example, Bennett et al. (2000) and Bennett and Walters (2001) investigated four design techniques (i.e., bar graphs/extenders, scale markers/scale grids, color coding/layering/separation, and digital values) aimed at improving performance at focused tasks. These techniques were applied alone and in combination to a configural display. The results indicate that three of these design techniques were successful because they provided display constraints (either additional analog visual structure or precise digital information) that matched the constraints of the task (i.e., provide a quantitative estimate of an individual variable).

In summary, the semantic mapping principle of display design does not predict inherent trade-offs between displays and tasks; a carefully designed configural display can support performance across the continuum of attention demands that operators might face in complex, dynamic domains. The issues in design and potential solutions that have been described here are necessarily brief. We encourage the interested reader to visit Bennett and Flach (2011) for much more detailed descriptions and analyses of the foundations of display design from the visual attention perspective, including critiques of a revised version of PCP (e.g., Wickens and Carswell, 1995).

### 3.4 Problem-Solving and Decision-Making Approach

The fourth perspective on display design to be discussed is problem solving and decision making. Recently, there has been an increased appreciation for the creativity and insight that experts bring to human–machine systems. Under normal operating conditions, a person is perhaps best characterized as a decision maker. Depending on the perceived outcomes associated with different courses of action, the amount of evidence that a decision maker requires to choose a particular option will vary. In models of decision making, this is called a *decision criterion*. Under abnormal or unanticipated operating conditions, a person is characterized most appropriately as a creative problem solver. The cause of the abnormality must be diagnosed, and steps must be taken to correct the abnormality (i.e., an appropriate course of action must be determined). This involves monitoring and controlling system resources, selecting between alternatives, revising diagnoses and goals, determining the validity of data, overriding automatic processes, and coordinating the activities of other people. Thus, the literature on reasoning, problem solving, and decision making has important insights for display design.

There is a vast literature on problem solving, ranging from the seminal work of the Gestaltists (e.g., Wertheimer, 1959) to the paradigmatic contributions of Newell and Simon (1972) to contemporary approaches. For the Gestalt psychologists, perception and cognition (more specifically, problem solving) were intimately intertwined. The key to successful problem solving was viewed as the formation of an appropriate gestalt, or representation, that revealed the "structural truths" of a problem. For example, Wertheimer (1959, p. 235) states that "thinking consists in envisaging, realizing structural features and structural requirements . . . ." The importance of a representation is still a key consideration today; it is probably not an overstatement to conclude that the primary lesson to be learned from the problem-solving literature is that the representation of a problem has a profound influence on the ease or difficulty of its solution.

Historically, decision research has focused on developing models that describe the generation of multiple alternatives (potentially, all alternatives), evaluation (ranking) of these alternatives, and selection of the most appropriate alternative. By and large, perception was ignored. In contrast, recent developments in decision research, stimulated by research on naturalistic decision making (e.g., Klein et al., 1993), have begun to

give more consideration to the generation of alternatives in the context of dynamic demands for action. Experts are viewed as generating and evaluating a few "good" alternatives. The emphasis is on recognition (e.g., how this problem is similar to or dissimilar from problems encountered before). As a result, perception plays a dominant role. This change in emphasis has increased awareness of perceptual processes and dynamic action constraints in decision making.

These trends have, either directly or indirectly, led researchers in interface design to focus on the representation problem. Perhaps the first explicit realization of the power of graphic displays to facilitate understanding was the STEAMER project (Hollan et al., 1984, 1987), an interactive inspectable training system. STEAMER provided alternative conceptual perspectives: "conceptual fidelity" of a propulsion engineering system through the use of analogical representations. In addition, the current approach to the design of human–computer interfaces (direct manipulation) (Hutchins et al., 1986; Shneiderman, 1986, 1993) can be viewed as an outgrowth of this general approach. More recently, scientific visualization (the role of diagrams and representation in discovery and invention) is being investigated vigorously (Bonneau et al., 2006; Brodie et al., 1992; Earnshaw and Wiseman, 1992). Thus, the challenge for display design from this perspective is to provide appropriate representations that support humans in their problem-solving endeavors.

## 4   MEANING-PROCESSING APPROACH TO DISPLAY DESIGN

It should be noted that in the aesthetic, psychophysical, and most attention-based approaches little consideration is given to a domain behind the display. It was not necessary for us to describe the "problem" behind the displays shown in Figure 10. However, the correspondence between the visual structure in a representation and the constraints in a problem is fundamental to the problem-solving and decision-making approaches. Recently, a number of research groups have recognized that effective interfaces depend on both the mapping from human to display (the coherence problem) and the mapping from display to a work domain or problem space (the correspondence problem). Terms used to articulate this recognition include direct perception (Moray et al., 1994), ecological interface design (Rasmussen and Vicente, 1989; Vicente, 1991, 1999; Burns and Hajdukiewicz, 2004), representational design (Woods, 1991), or semantic mapping (Bennett and Flach, 1992).

Thus, our approach (Bennett and Flach, 2011) is a problem-driven (as opposed to user- or technology-driven) approach to the design and evaluation of displays and interfaces. By this we mean that the primary purpose of an interface is to provide decision-making and problem-solving support for a user who is completing work in a domain. The goal is to design interfaces that are (1) tailored to specific work demands, (2) leverage the powerful perception–action skills of the human, and (3) use powerful interface

technologies wisely. This can be conceptualized as a "triadic" approach (domain/ecology, human/awareness, interface/representation) to human computer interaction that stands in sharp contrast to the traditional "dyadic" (human, interface, information processing) approaches.

Ultimately the success or failure of an interface is determined by the interactions that occur between all three components of the triad. Each component contributes a set of constraints that will influence the effectiveness (and/or the pleasurableness) of the interaction. A particular work domain will introduce a particular set of constraints (e.g., tasks, goals, limits) that will determine the nature of the work to be completed. Another set of constraints are introduced by the cognitive agent (human, machine) that completes the work. For a human agent this will include a specific set of cognition/perception/action capabilities and limitations. The functionality/design of the interface introduces a third set of constraints: Particular characteristics of the interface will introduce cognitive demands that will vary in terms of the nature and amount of cognitive resources that are required. These three sources of constraints are independent but mutually interactive and mutually constraining. The effectiveness of graphical decision support will ultimately depend upon the quality of *very specific sets of mappings* between these constraints. Thus, the focus of our approach is not on information-processing characteristics, graphical forms, events, trajectories, tasks, or procedures per se. Instead, the focus is on the quality of the mappings between the person, the interface, and the domain. Any approach that fails to consider all of these components and their interactions (i.e., dyadic approaches) will be inherently, and severely, limited.

### 4.1   Correspondence Problem: Semantics of Work

*Correspondence* refers to the issue of content: What information should be present in the interface in order to meet the cognitive demands of the work domain? Correspondence is defined neither by the domain itself nor by the interface itself: It is a property that arises from the interaction of the two. Thus, in Figure 11, correspondence is represented by the labeled arrows that connect the domain and the interface. One convenient way to conceptualize correspondence is as the quality of the mapping between the interface and the workspace, where these mappings can vary in terms of the degree of specificity (consistency, invariance, or correspondence). As we will demonstrate, within this mapping there can be a one-to-one correspondence, a many-to-one, a one-to-many, or a many-to-many mapping between the information that exists in the interface and the structure within the workspace.

### 4.1.1   Rasmussen's Abstraction Hierarchy

Addressing the issue of correspondence requires a deep understanding and explicit description of the "semantics" of a work domain. Rasmussen's (1986) abstraction hierarchy is a theoretical framework for describing domain semantics in terms of a nested hierarchy of functional constraints (including goals, physical laws,
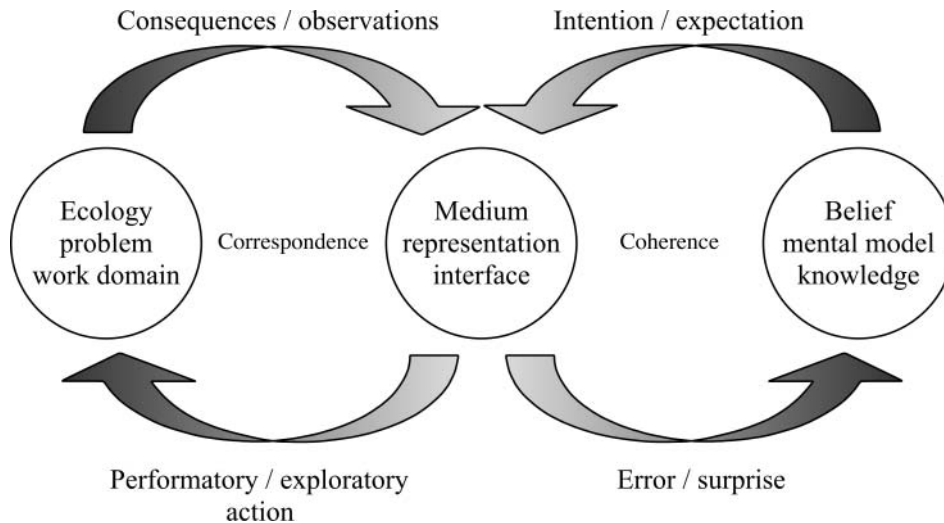
**Figure 11** The dynamics of a meaning-processing approach to interface design. These dynamics involve interactions between a cognitive system and an ecology mediated by an interface (displays and controls). Perception and action are dynamically coupled in parallel so that every interaction has dual implications.(Adapted with permission from Bennett and Flach, 2011. Copyright CRC Press.)

regulations, organizational/structural constraints, equipment constraints, and temporal/spatial constraints). One way to think about the abstraction hierarchy is that it provides structured categories of information (i.e., the alternative conceptual perspectives) that a person must consider in the course of accomplishing system goals. Consider the following passage from Rasmussen (1986, p. 21):

> During emergency and major disturbances, an important control decision is to set up priorities by selecting the level of abstraction at which the task should be initially considered. In general, the highest priority will be related to the highest level of abstraction. First, judge overall consequences of the disturbances for the system function and safety in order to see whether the mode of operation should be switched to a safer state (e.g., standby or emergency shutdown). Next, consider whether the situation can be counteracted by reconfiguration to use alternative functions and resources. This is a judgment at a lower level of function and equipment. Finally, the root cause of the disturbance is sought to determine how it can be corrected. This involves a search at the level of physical functioning of parts and components. Generally, this search for the physical disturbance is of lowest priority (in aviation, keep flying—don't look for the lost light bulb!).

Thus, in complex domains, situation awareness requires the operator to understand the process at different levels of abstraction. Further, the operator must be able to understand constraints at one level of abstraction in terms of constraints at other levels. The correspondence question asks whether the hierarchy of

constraints that define a work domain are reflected in the interface.

### 4.2  Coherence Problem: Syntax of Form

*Coherence* refers to the mapping from the representation to the human perceiver. Here the focus is on the visual properties of the representation. What distinctions within the representation are discriminable to the human operator? How do the graphical elements fit together or coalesce within the representation? Is each element distinct or separable? Are the elements absorbed within an integral whole, thus losing their individual distinctness? Or do the elements combine to produce configural or global properties? Are some elements or properties of the representation more or less salient than other elements or properties?

In general, coherence addresses the question of how the various elements within a representation compete for attentional and cognitive resources. Just as work domains can be characterized in terms of a nested hierarchy of constraints, complex visual representations can be perceived as a hierarchy of nested structures, with local elements combining to produce more global patterns or symmetries. Ultimately coherence refers to the extent to which a human agent can obtain and make sense of information about a work domain that is present in the display.

### 4.3  Mapping Problem

In human–machine systems, a display is a representation of an underlying domain, and the user's tasks are defined by that domain rather than by the visual characteristics of the display itself. Thus, whether or not a display will be effective is determined by both correspondence and coherence. More specifically, the

effectiveness of the display is determined by the quality of the mapping among the agent, interface, and domain.

The constraints that characterize a particular work domain will have a substantial impact on the type of representations that will provide effective support. Rasmussen et al. (1994) have analyzed the various types of constraints that characterize different work domains and have developed a continuum for classification. At one end of the continuum are domains in which the unfolding events arise from the laws of nature (e.g., process control). An example of such a law is the conservation of mass: If more mass is flowing into a reservoir than out of a reservoir, the level of fluid that it contains will rise. In these "law-driven" domains, the user is required to control, monitor, and compensate for the demands that arise from the domain. At the opposite end of the spectrum are domains in which the unfolding events arise from the user's intentions, goals, and needs (e.g., information search and retrieval). In these "intent-driven" domains, the demands are created by the user rather than by the domain. The domain structure is more loosely coupled (e.g., attributes that differentiate among books of fiction) and the process of searching and identifying the appropriate information (e.g., a particular book of fiction to read) is ultimately user dependent. Note that an understanding of the domain structure is still critical to the development of effective decision support (Flach et al., 2011).

The design strategies and techniques that are required to develop effective interfaces for these two categories of domains are quite different. In law-driven domains, the constraints of the system (e.g., physical, functional, and goal-related structure) are the primary consideration. Display design involves the development of abstract geometric forms that reflect these inherent constraints. A simple example is using an axis in a graph to represent time. We will use the general term "analog" to refer to these representations, since a continuous incremental change in a domain variable or property is reflected by a corresponding continuous and incremental change in its graphical representation.

We will also use the more specific term "configural" to refer to these representations, since they use configural perceptual dimensions and produce the emergent features that were discussed in Section 3.3. In configural representations the geometric display constraints will generally take the form of symmetries: equality (e.g., length, angle, area), parallel lines, collinearity, or reflection. In addition, Gestalt properties of closure and good form are useful. Each particular representation that is chosen will produce a different set of display constraints, defined by the spatiotemporal structure (the visual appearance of the display over time).

The core problem in implementing effective configural displays for law-driven work domains is to provide visual representations that are perceived as accurate reflections of the abstract domain constraints: Are the critical domain constraints reflected appropriately in the geometric constraints in the display? Are breaks in the domain constraints (e.g., abnormal or emergency conditions) reflected by breaks in the geometric constraints (e.g., emergent features such as nonequality,

nonparallelism, nonclosure, bad form)? Only when this occurs will the cognitive agent be able to obtain meaning about the underlying domain in an effective fashion.

One source of ideas for configural displays is the graphical representations that engineers use to make design decisions. For example, Beltracchi (1987, 1989) (see also Moray et al., 1994; Rasmussen et al., 1994) has designed a configural display for controlling the process of steam generation in nuclear power plants based on the temperature–entropy graphic used to evaluate thermodynamic engines (Rankine cycle display). Effective interface design for law-driven domains allows trained operators to use high-capacity perceptual and motor skills to monitor and control the system as opposed to limited capacity resources (e.g., working memory).

A very different design strategy is required for intent-driven domains, where the needs and goals of the user are the driving force in the unfolding interaction. Relative to law-driven domains, agents working in intent-driven domains will interact with the system more sporadically, will have far less training and experience, and will possess more diverse sets of skills or knowledge. Under these circumstances the appropriate interface design strategy is to use metaphors and icons.

Metaphorical representations use spatial or symbolic relations from other, more familiar work domains to convey meaning. They are designed to relate the functioning of the system and the requirements for interaction to concepts and activities with which the majority of potential agents will already be familiar. Ultimately, the goal is to enhance the transfer of skills from one domain to another. Perhaps the most obvious example is the "desktop" metaphor that is used in personal computer systems. Another example is the BookHouse metaphor, developed by Pejtersen (1980; 1992) to facilitate library information retrieval. Rasmussen et al. (1994, pp. 289–291) describe the metaphor and its justification:

> The use of the BookHouse metaphor serves to give an invariant structure to the knowledge base.... Since no overall goals or priorities can be embedded in the system, but depend on the particular user, a global structure of the knowledge base reflects subsets relevant to the categories of users having different needs and represented by different rooms in the house.... This gives a structure for the navigation that is easily learned and remembered by the user.... The user "walks" through rooms with different arrangements of books and people.... It gives a familiar context for the identification of tools to use for the operational actions to be taken. It exploits the flexible display capabilities of computers to relate both information in and about the data base, as well as the various means for communicating with the data base to a location in a virtual space.... This approach supports the user's memory of where in the BookHouse the various options and information items are located. It facilitates the navigation of the user so that items can be remembered in given physical locations that one can then retraverse in order to retrieve a given item and/or freely browse in order to gain an overview.

In addition to metaphors and configural displays, there is a third type of representation: propositional. This refers to the use of digital values (i.e., numbers), the alphabet (i.e., words and language), and other forms of alphanumeric labels. Propositional representations are compact and precise. They capitalize on an extensive knowledge base and provide the opportunity for the most detailed and precise representation of an ecology. However, unlike metaphorical and analogical representations, the mapping between symbol and referent is an arbitrary one for propositional representations. This form of representation is therefore the most computationally expensive, in terms of placing demands on knowledge-based processes. Thus, propositional representations can be an important source of information, when configured within metaphorical and analogical forms of representations.

Whether analogical, metaphorical, or combined representations are used, the key to successful design is the quality of the mapping. The visual salience of the information in the display must reflect the relative importance of that information in terms of the work domain. For analogical, configural displays the geometric symmetries must correspond to higher-order constraints on the process. For metaphorical displays, the intuitions and skills elicited by the representational domain must map appropriately to the target domain.

## 5   EXAMPLE-BASED TUTORIAL OF MEANING-PROCESSING APPROACH

The concepts and principles of display design that have been introduced thus far include correspondence, coherence, process constraints, display constraints (e.g., emergent features), and the mappings between process and display constraints. These concepts and principles are necessarily abstract, and for them to be useful for display design they must be presented in a clear and unambiguous fashion. In this section we provide a tutorial that illustrates these concepts and principles through a series of concrete examples.

We begin with an analysis of a law-driven domain: a simple system from the domain of process control. The goal is to provide a description of the associated process constraints. We then consider various types of displays that could be devised for the system. The goal is to consider the alternative mappings between process constraints and geometric (display) constraints that are provided by each representation: in particular, the implications for correspondence and coherence. The representations are chosen to illustrate the continuum of visual forms from separable, through configural, to integral geometries. We then examine one representation in greater detail and discuss the implications of this mapping for normal and abnormal operating conditions. We end the section with a set of practical guidelines for display design.

### 5.1   Simple Domain from Process Control

The process is a generic one that might be found in process control, and it is represented graphically in the lower portion of Figure 12. There is a reservoir (or tank, represented by the large rectangle in the middle of the figure) that is filled with a fluid (e.g., coolant). The volume, or level, of the reservoir ($R$) is represented by the filled portion of the rectangle. Fluid can enter the reservoir through the two pipes and valves located above the reservoir; fluid can leave the reservoir through the pipe and valve located below the reservoir. We categorize the information in this simple process using a simple distinction in which the term *low-level data* refers to local constraints or elemental state variables that might be measured by a specific sensor. The term *higher level properties* will be used to refer to more global constraints that reflect relations or interactions among multiple variables.

***Low-Level Data (Process Variables)***   There are two goals associated with this simple process. First, there is a goal ($G_1$) associated with $R$, the level of the reservoir. The reservoir should be maintained at a relatively high level to ensure that sufficient resources are available to meet long-term increases in demanded output flow rate ($O$). The second goal ($G_2$) refers to the specific rate of output flow that must be maintained to meet an external demand. These goals are achieved and maintained by adjusting three valves ($V_1$, $V_2$, and $V_3$) that regulate flow through the system ($I_1$, $I_2$, and $O$). Thus, this simple process is associated with a number of process variables that can be measured directly: these low-level data are listed in the upper, left-hand portion of Figure 12 ($V_1$, $V_2$, $V_3$, $I_1$, $I_2$, $O$, $G_1$, $G_2$, and $R$).

***High-Level Properties (Process Constraints)***   In addition, there are relationships between these process variables that must be considered when controlling the process (see the upper, right-hand portion of Figure 12). The most important high-level properties are goal related: Does the actual reservoir volume level ($R$) match the goal of the system ($G_1$)$-K_5$? Does the actual system output flow rate ($O$) match the flow rate that is required ($G_2$)$-K_6$? Even for this simple process, some of the constraints or (high-level properties) are fairly complex. For example, an important property of the system is mass balance, which is determined by comparing the mass leaving the reservoir ($O$, the output flow rate) to mass entering the reservoir (the combined input flow rates of $I_1$ and $I_2$). This relationship determines the direction and the rate of change for the volume inside the reservoir ($\Delta R$). For example, if mass in and mass out are equal, the mass is balanced, $\Delta R$ will equal 0.00, and $R$ will remain constant.

Controlling even this simple process will depend on a consideration of both high-level properties and low-level data. As the earlier example indicates, decisions about process goals (e.g., maintaining a sufficient level of reservoir volume) generally require consideration of relationships between variables (is there a net inflow or a net outflow or is mass balanced?) as well as the values of the individual variables themselves (what is the current reservoir volume?).

#### 5.1.1   Abstraction Hierarchy Analysis

The constraints of the simple process in Figure 12 will be characterized in terms of the abstraction

Low-level data
(process variables)

High-level properties
(process constraints)

$T$ = time
$V_1$ = setting for valve 1
$V_2$ = setting for valve 2
$V_3$ = setting for valve 3
$I_1$ = flow rate through valve 1
$I_2$ = flow rate through valve 2
$O$ = flow rate through valve 3
$R$ = volume of reservoir

$K_1$ = $I_1 - V_1$  Relation between comman-
$K_2$ = $I_2 - V_2$  ded flow ($V$) and actual flow
$K_3$ = $O - V_3$  ($I$ or $O$)

$K_4$ = $\Delta R = (I_1 + I_2) - O$
Relation between reservoir
volume ($R$), mass in ($I_1 + I_2$),
and mass out ($O$)

$G_1$ = volume goal
$G_2$ = output goal (demand)

$K_5$ = $R - G_1$  Relation between actual states
$K_6$ = $O - G_2$  ($R, O$) and goal states ($G_1, G_2$)
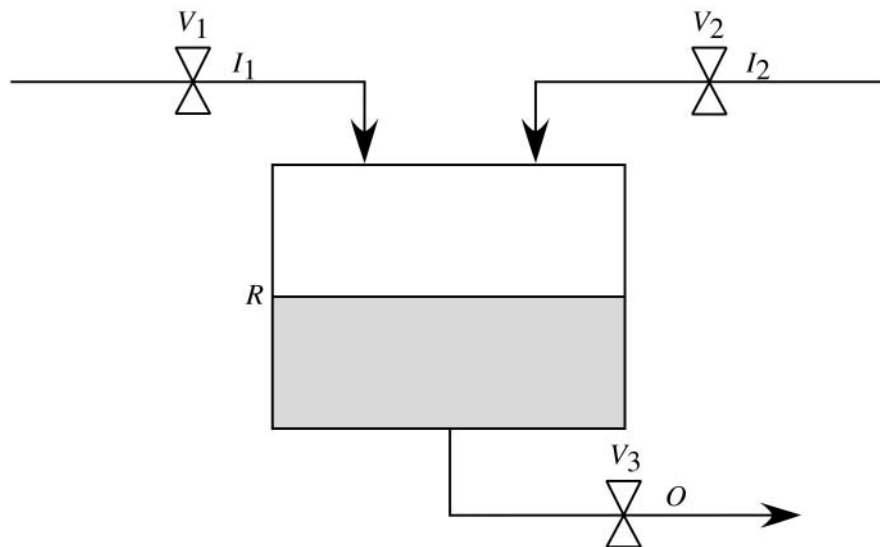


**Figure 12** Simple domain from process control that has a reservoir for storing mass, two input streams that increase the volume of mass in the reservoir, and a single output stream that decreases the volume. The low-level data (the measured domain variables), the high-level properties (constraints that arise from the interaction of these variables and the physical design), and the domain goals (requirements that must be met for the system to be functioning properly) are listed.

hierarchy. Typically, the hierarchy has five separate levels of description, ranging from the physical form of a domain to the higher level purposes it serves. The highest level of constraints refers to the *functional purpose* or design goals for the system. For our simple process these are constraints $K_5$ and $K_6$. For example, consider the relationship between $R$ and $G_1$. When the actual reservoir volume ($R$) equals the goal reservoir volume ($G_1$),

the difference between these two values will assume a constant value (0.00). This process constraint is represented by the equation associated with the higher level property $K_5$ in Figure 12. For an actual work domain, the associated values (costs and benefits) underlying these particular goals might be considered. The *abstract functions* or physical laws that govern system behavior are another important source of constraints. In our

example, $K_4$ reflects the law of conservation of mass. Change of mass in the reservoir ($\Delta R$) should be determined by the difference between the residual mass in ($I_1 + I_2$) and the mass out ($O$). Similar constraints associated with the mass flow are represented as $K_1$, $K_2$, and $K_3$. Flow is proportional to valve setting (this assumes a constant-pressure head). Further constraints arise as a result of the *generalized function* (sources, storage, sink). In our example, there are two sources: a single store and a single sink. The physical processes behind each general function represents another source of constraint, *physical function*. In this case there are two feedwater streams, a single output stream, and a reservoir for storage. Similarly, the moment-to-moment values of each variable ($T$, $V_1$, $V_2$, $V_3$, $I_1$, $I_2$, $O$, and $R$) should be considered at the level of physical function. Finally, the level of *physical form* provides information concerning the physical configuration of the system, including information related to causal connections, length of pipes, position of valves on pipes, and size of the reservoir. All of these constraints will be satisfied if the process is being controlled in a proper fashion.

To summarize, an abstraction hierarchy analysis provides information about the hierarchically nested constraints that constitute the semantics of a domain and therefore defines the information that must be present in the interface for a person to perform successfully. The product of this analysis (interrelated categories of information) provides a structured framework for display development, as we will demonstrate shortly. It should be emphasized that this analysis and description is independent of the interface and therefore differs from traditional task analysis. Although space limitations do not permit a complete discussion, we view abstraction hierarchy analysis and task analysis (traditional or cognitive) as complementary processes that are necessary for the development of effective displays.

## 5.2   Coherence and Correspondence: Alternative Mappings

In this section we provide six examples that illustrate alternative mappings between domain semantics and representations (displays) for our simple process (Figure 13). The discussion is organized in terms of the distinction between integral, configural, and separable dimensions that was outlined in Section 3.3. One goal is to illustrate what these terms, originally coined in the attention literature, mean in the context of display design for complex systems. A second goal is to focus on the quality of the mapping that each display provides, especially with respect to the ability of each display to convey information at various levels of abstraction (see Sections 4.1.1 and 5.1.1). To illustrate the quality of the mapping explicitly, we have provided a summary listing (at the right of each display in Figure 13) that sorts the associated process constraints into two categories (P and D). Process constraints that are represented directly in the display (i.e., which can be "seen") have been placed in the P category (*P*erceived). Process constraints that are not represented directly and must be computed or inferred are placed in the D category (*D*erived). Process constraints that are related to physical structure are

represented by the theta symbol ($\phi$); process constraints related to the functional structure are represented by the integration symbol ($\int$).

***Separable Displays***   Figure 13*a* represents a separable display that contains a single display for each individual process variable present. Each display is represented in the figure by a circle, but no special significance should be attached to the symbology: The circles could represent digital displays, bar graphs, and so on. For example, four instantiations of this display are shown in Figures 10*a*, *b*, *e*, and *f*. In Figures 10*a* and *b* the display constraints are the relative heights of the bars in response to changes in the underlying variables.

In terms of the abstraction hierarchy, the class of displays represented by Figure 13*a* provides information only at the level of physical function: Individual variables are represented directly. Thus, there is not likely to be a focused-attention cost for low-level data. However, there is likely to be a divided-attention cost, because the observer must derive the high-level properties. To do so, the observer must have an internalized model of the functional purpose, the abstract functions, the general functional organization, and the physical process. For example, to determine the direction (and cause) of $\Delta R$ would require detailed internal knowledge about the process, since no information about physical relationships ($\phi$) or functional properties ($\int$) is present in the display.

Simply adding information about high-level properties does not change the separable nature of the display. In Figure 13*b* a second separable display has been illustrated. In this display the high-level properties (constraints) have been calculated and are displayed directly, including information related to functional purpose ($K_5$ and $K_6$) and abstract function ($K_1$, $K_2$, $K_3$, and $K_4$). This does off-load some of the calculational requirements (e.g., $\Delta R$). However, there is still a divided-attention cost. Even though the high-level properties have been calculated and incorporated into the display, the relationships among and between levels of information in the abstraction hierarchy are still not apparent. The underlying cause of a particular system state still must be derived from the separate information that is displayed. Thus, although some low-level integration is accomplished in the display, the burden for understanding the causal structure still rests in the observer's stored knowledge.

***Configural Displays***   The first configural display, illustrated in Figure 13*c*, provides a direct representation of much of the low-level data that are present in the display in Figure 13*a*. However, it also provides additional information that is critical to completing domain tasks: information about the physical structure of the system ($\phi$). This "mimic" display format was introduced in STEAMER (Hollan et al., 1984), and issues in the animation of these formats have been investigated more recently (Bennett, 1993; Bennett and Madigan, 1994; Bennett and Nagy, 1996; Bennett and Malek, 2000).

The mimic display is an excellent format for representing the generalized functions in the process. It
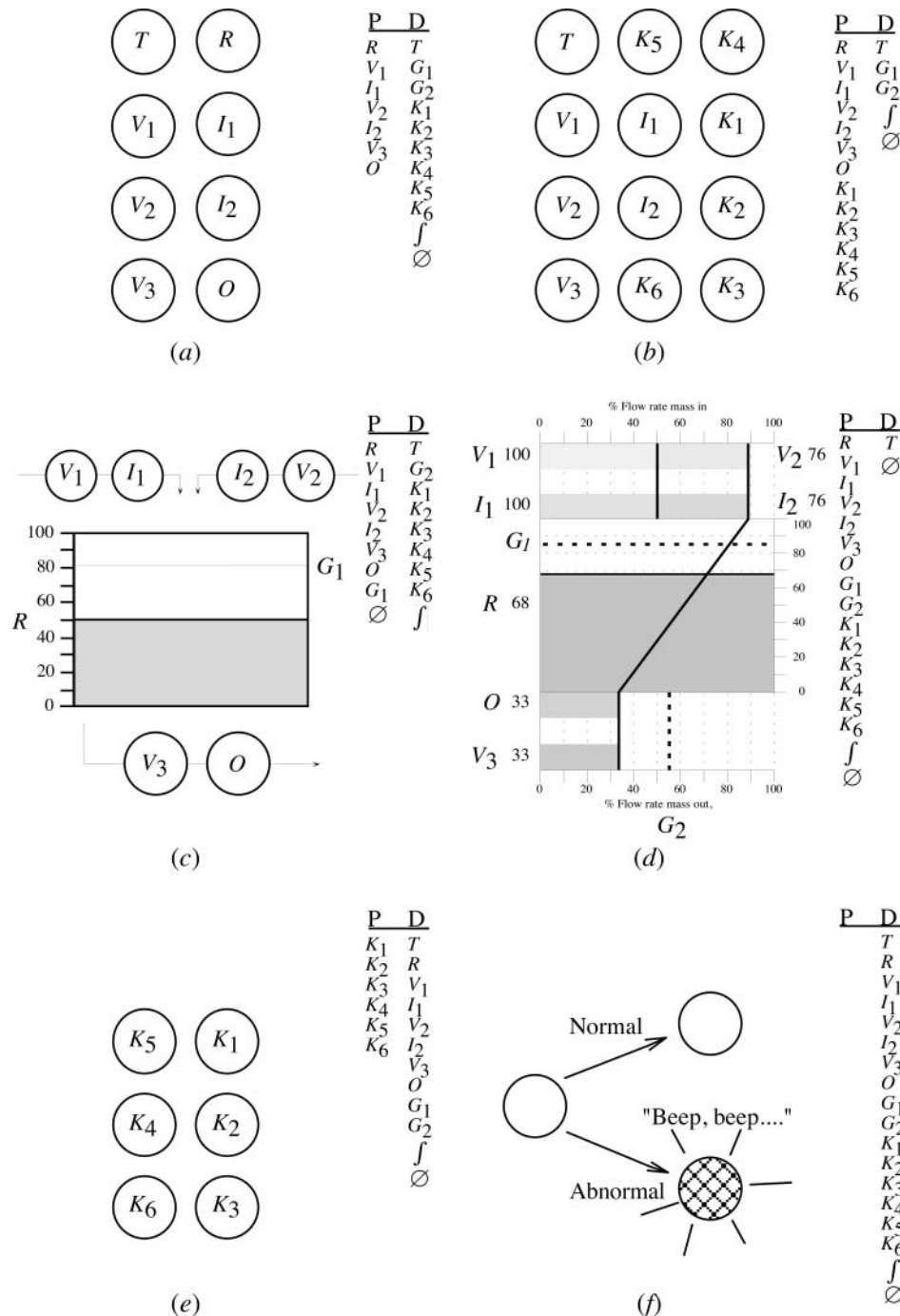
**Figure 13** Six alternative mappings for the domain constraints described in Figure 12. The circles represent generic separable displays, which could be bar graphs, pie charts, or digital displays. The data and properties outlined in Figure 12 have been placed in two categories for each mapping: P for data that can be perceived directly from the display and D for data that must be derived from the display by the observer. Parts (*a*) and (*b*) represent separable mappings, (*c*) and (*d*) represent configural mappings, and (*e*) and (*f*) represent integral mappings. These mappings illustrate how the terms *separable, configural*, and *integral* have different meanings when applied to display design (as opposed to their meaning in the attention literature).

has many of the properties of a functional flow diagram or flowchart. The elements can represent physical processes (e.g., feedwater streams), and by appropriately scaling the diagram, relations at the level of physical form can be represented (e.g., relative positions of valves). Also, the moment-to-moment values of the process variables can easily be integrated within this representation. Not only does this display include information with respect to generalized function, physical function, and physical form, but also the organization provides a visible model illustrating the relations across these levels of abstraction. This visual model allows the observer to "see" some of the logical constraints that link the low-level data. Thus, the current value of $I_2$ can be seen in the context of its physical function (feedwater stream 2) and its generalized function (source of mass); in fact, its relation to the functional purpose in terms of $G_1$ is also readily apparent from the representation.

Just as in the displays listed in Figures 13a and b, there is not likely to be a cost in selective attention with respect to the low-level data. However, although information about physical structure illustrates the causal factors that determine higher level system constraints, the burden of computing these constraints (e.g., determining mass balance) rests with the observer. Thus, what is missing in the mimic display is information about abstract function (information about the physical laws that govern normal operation).

The second configural display, illustrated in Figure 13d, is slightly more complex [the logic is similar to that of Vicente (1991)] and will be described in detail before discussing the quality of the mapping that it provides. The valve settings $V_1$ and $V_2$ are represented as back-to-back horizontal bar graphs that increase or decrease in horizontal extent with changes in settings. The measured flow rates ($I_1$ and $I_2$) have the same configuration of graphical elements and are located below the valve settings in the display. The horizontal bar graphs depicting valve settings and flow rates for a particular pipe (e.g., $V_1$ and $I_1$) are connected with a bold vertical line (in Figure 13d both of the lines are perpendicular because the settings and flow rates are equal in both input streams). The volume of the reservoir ($R$) is represented by a bold horizontal line and as the filled portion of the rectangle inside the reservoir. The value of $R$ can be read from the scale on the right side of the display and the associated digital value on the left; in Figure 13d the value of $R$ is 68. The associated reservoir volume goal ($G_1$) is represented by the bold horizontal dashed line (approximately 85). The flow rate of the mass leaving the reservoir is represented by the horizontal bar graph labeled $O$ at the bottom of the display; the corresponding valve setting is represented by the bar graph labeled $V_3$. These two bar graphs are also connected by a bold vertical line. The mass output goal ($G_2$) is represented by the bold vertical dashed line (approximately 55). The relationship between mass in ($I_1 + I_2$) and mass out ($O$) is highlighted by the bold angled line that connects the corresponding bar graphs.

Unlike the displays discussed previously, this configural display integrates information from all levels of the abstraction hierarchy in a single representation, making extensive use of emergent features that include equality, parallel lines, and collinearity. The general functions are related through a graphical "flow chart" with input (source) at the top, storage in the center, and output (sink) on the bottom. The abstract functions are related using the emergent features of equality and the resulting collinearity across the bar graphs. For example, the constraints on mass flow ($K_1$, $K_2$, $K_3$) are represented by a salient emergent feature (i.e., equality of the horizontal extent of the bars labeled $V_1/I_1$, $V_2/I_2$, and $V_3/O$). In addition, the constraints relating rate of volume change and mass balance ($K_4$) are represented by the horizontal extent of $I_1 + I_2$ relative to the horizontal extent of $O$, and these relationships are highlighted by the bold line connecting these bars. Thus, the mass balance is represented by the symmetry between the input bar graphs and the output bar graphs; the orientation of the line connecting them constitutes an additional emergent feature that should be proportional to rate of change of mass in the reservoir. Constraints at the level of functional purpose are illustrated by the difference between the goal and the relevant variable. For example, the constraint on mass inventory ($K_5$) is shown using the relative position between the hatched area representing volume within the reservoir and the bold horizontal dashed line representing the goal level $G_1$.

Although not a direct physical analog, this configural display preserves important physical relations from the process (e.g., volume and filling). In addition, it uses a variety of emergent features that provide a direct visual representation of the process constraints and that connect these constraints so as to make the functional logic of the process visible within the geometric form. In short, the visual properties of this display, most notably emergent features, provide a set of geometric constraints that specify the domain constraints directly (note that when we use the term geometric constraints we will be primarily referring to these emergent features). As a result, performance for both focused- and divided-attention tasks is likely to be facilitated substantially.

**Integral Displays** Figure 13e shows an integral mapping in which each of the process constraints are shown directly, providing information at the higher levels of abstraction. However, the low-level data must be derived. In addition, there is absolutely no information about the functional processes behind the display, and therefore the display does not aid the observer in relating the higher level constraints to the physical variables. Because there would normally be a many-to-one mapping from physical variables to the higher order constraints, it would be impossible for the observer to recover information from this display at lower levels of abstraction.

Figure 13f shows the logical extreme of this continuum. In this display the process variables and constraints are integrated into a single "bit" of information that indicates whether or not the process is working properly (all constraints are at their designed value). It should be obvious that although these displays may have no divided-attention costs, they do have selective-attention

costs and provide little support for problem solving when the system fails.

**Summary** This section has focused on issues related to the quality of mapping between process constraints and display constraints. Even the simple domain that we chose for illustrative purposes has a nested structure of domain constraints: There are multiple constraints that are organized hierarchically both within and between levels of abstraction. The six alternative displays achieved various degrees of success in mapping these constraints. The principle of correspondence is illustrated by the fact that these formats differ in terms of the amount of information about the underlying domain that is present. The display in Figure 13*f* has the lowest degree of correspondence; the displays in Figures 13*b* and *d* have the highest degree of correspondence. These two displays are roughly equivalent in correspondence, with the exception of the two goals that are present in Figure 13*d* but absent in Figure 13*b*. Although these two displays are roughly equivalent in correspondence, it should be clear from the prior discussion that they are definitely not equivalent in terms of coherence. Figure 13*d* allows a person to perceive information concerning the physical structure, functional structure, and hierarchically nested constraints in the domain directly, a capability that is not supported by the format in Figure 13*b*. The coherence of Figure 13*d* will be explored in greater detail in the following section. This section has also illustrated the duality of meaning for the terms *integral, configural*, and *separable*. In attention, these terms refer to the relationship between perceptual dimensions, as described in Section 3.3; in display design, these terms refer more appropriately to the nature of the mapping between the domain and the representation.

### 5.3 Meaning Processing: Normal and Abnormal Operating Conditions

In Section 10 we outlined differences in correspondence and coherence that resulted from six alternative mappings for our simple domain. In this section we explore issues related to coherence in greater detail, focusing on Figure 13*d* and the implications of the mapping for performance under both normal and abnormal or emergency operating conditions. To begin, we discuss the facilitating role that graphical constraints (i.e., emergent features) representing information in the abstraction hierarchy (in particular, abstract function—the physical laws that govern normal operation) can play under normal conditions. Properly designed configural displays will provide a powerful representation for control: breaks in the domain constraints will generally be seen as breaks in display constraints (e.g., nonsymmetries) and will suggest appropriate control inputs. This information is, perhaps, even more important for detecting faults (e.g., a leak). The possibility that these types of displays can change the fundamental nature of the behavior that is required on the part of the operator will also be entertained. Finally, the implications for the reduction of errors (more likely to occur under abnormal or emergency conditions) will be discussed.

The mapping between domain constraints and geometric constraints that is provided in the configural display shown in Figure 13*d* provides a powerful representation for control under normal operating conditions. In Figure 14*a* the display is shown with values for system variables indicating that all constraints are satisfied. The figure indicates that the flow rate is larger for the first mass input valve ($I_1$, $V_1$) than for the second ($I_2$, $V_2$) but that the two flow rates added together match the flow rate of the mass output valve ($O$, $V_3$). In addition, the two system goals ($G_1$ and $G_2$) are being fulfilled.

In contrast, Figures 14*b–d* illustrate failures to achieve system goals. In these displays not only is the violation of the goal easily seen, but each system variable is seen in the context of the control requirements. Thus, in Figure 14*b* it is apparent that the $K_5$ constraint is not being met (the actual level of the reservoir is higher than the goal). It is also apparent that the $K_4$ constraint is broken. The orientation of the line connecting mass in ($I_1 + I_2$) and the mass out ($O$) specifies that a positive net inflow for mass exists (mass in is greater than mass out). In essence, the deviation in orientation of this line from perpendicular is an emergent feature corresponding to the size of the difference. Under these circumstances control input is required immediately: An adjustment at valve 1 and/or valve 2 will be needed to avoid overflow from the reservoir. The observer can see these valves in the context of the two system goals; the representation makes it clear that these are the appropriate control inputs to make. For example, although adjusting valve 3 from 54 to a value greater than 70 would also cause the reservoir volume to drop, it is an inappropriate control input because goal 2 would then be violated.

In Figure 14*c* the situation is exactly the same, with one exception: There is a negative net inflow for mass, as indicated by the reversed orientation of the connecting line. Under these circumstances the operator can see that no immediate control input is required. Because mass in is less than mass out, the reservoir volume is falling, and this is exactly what is required to meet the $G_1$ reservoir volume goal. Of course, a control input will be required at some point in the future (mass will need to be balanced when the reservoir level approaches the goal). Similarly, in Figure 14*d* the observer can see that the $K_5$ and $K_6$ constraints are broken and that an adjustment to valve 3 (a decrease in output) is needed to meet the output requirements ($G_2$) and the volume goal ($G_1$).

Thus, in complex dynamic domains it is the pattern of relationships between variables, as reflected in the geometric constraints, that determines the significance of the data that are presented. It is this pattern that ultimately provides the basis for action, even when the action hinges on the value of an individual variable. When properly designed, configural displays will directly reflect these critical data relationships via emergent features and suggest the appropriate control input.

A similar logic applies for operational support under abnormal or emergency conditions. As in Figure 14, Figure 15*a* represents a configuration with all system
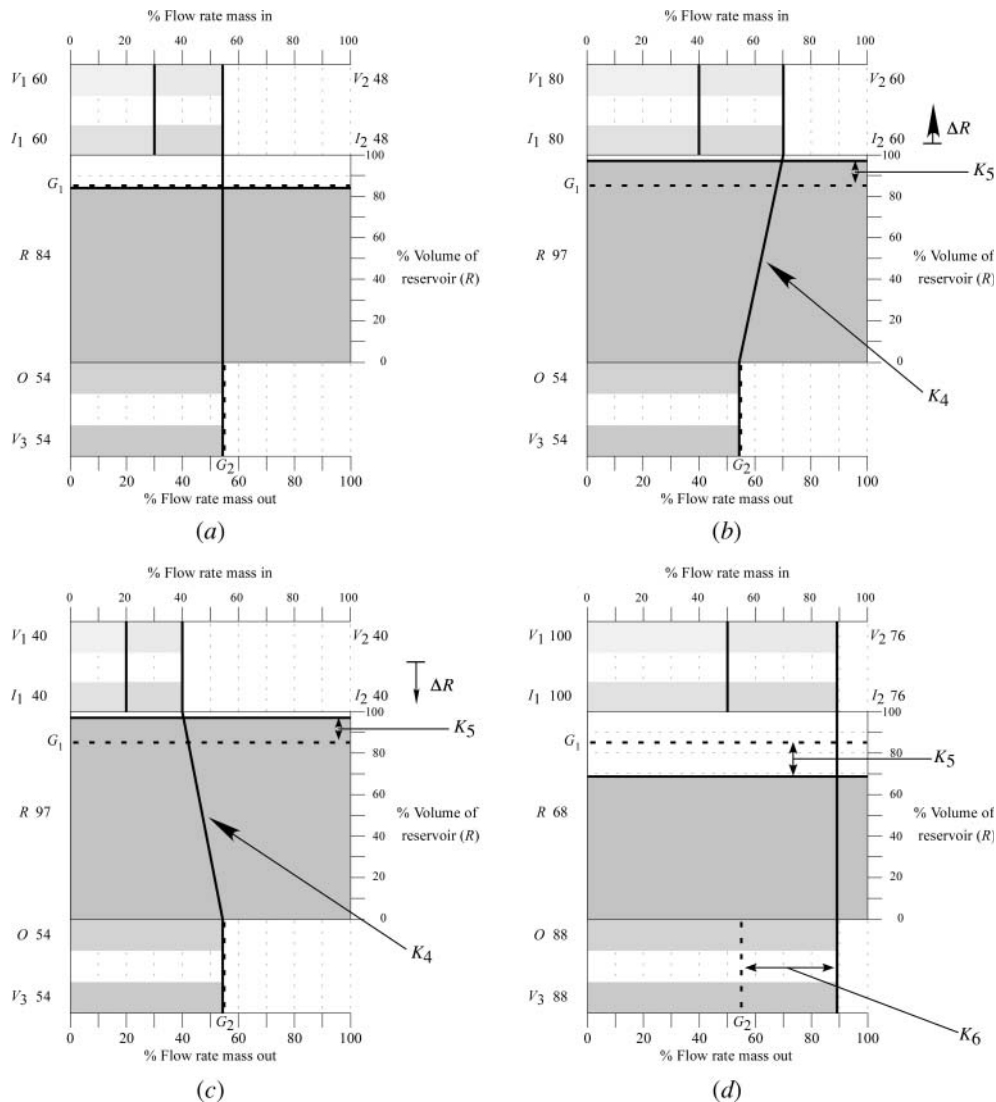
**Figure 14** Mapping between the domain constraints (data, properties, goals) and the geometric constraints (visual properties of the display, including emergent features such as symmetry and parallelism) under relatively normal operating conditions.

constraints being met. In Figure 15b the first constraint ($K_1$) is broken. There are two aspects of the display geometry indicating that the flow rate ($I_1$) does not match the commanded flow or valve setting ($V_1$). First, the horizontal extent of the two bar graphs in the top left portion of the display are not equal, and this relationship is emphasized by the orientation of the bold line connecting the two graphs (similar to the emergent feature for mass balance). There are a number of potential causes for this discrepancy, which include (1) a leak in the valve, (2) a leak in the pipe prior to the point at which the flow rate is measured, or (3) an obstruction in the pipe. In contrast, the fact that the line connecting $V_2$ and $I_2$ is not perpendicular

(but is parallel to the first connector line) does not indicate that the $K_2$ constraint is broken. Instead, this is an indication that the commanded and actual mass flows in the second mass input stream are equal (and therefore that the discrepancy is isolated in the first mass input stream). A similar mapping between geometric constraints and domain constraints represents a fault in the $K_3$ constraint, as illustrated in Figure 15c.

Figure 15d illustrates changes in the visual display (breaks in the geometric constraints) that are associated with a fault in the system (a break in the mass balance constraint, $K_4$). In this example, there is a positive net inflow of mass which is normally associated with an increase in the volume of the reservoir (again, specified
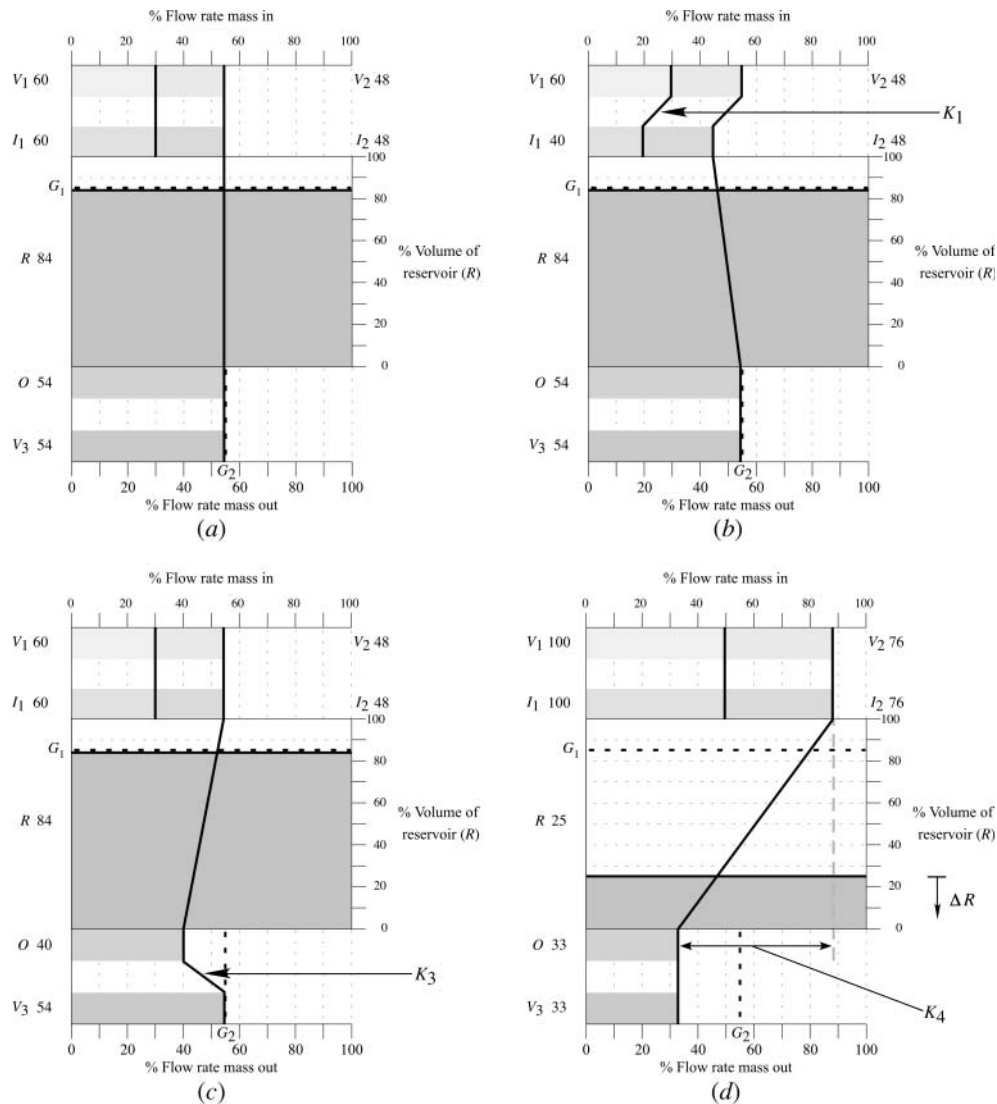
**Figure 15** Mapping between the domain constraints (data, properties, goals) and the geometric constraints (visual properties of the display, including emergent features such as symmetry and parallelism) under abnormal or emergency operating conditions.

by the emergent feature of orientation). However, in this case the mass inventory is falling, as we have indicated in the diagram by the downward-pointing arrow located near the $\Delta R$ symbol (this is difficult to represent in a static diagram but would be seen clearly on a dynamic display). Again, there are several potential explanations for this fault. The most likely explanation is that there is a leak in the reservoir itself; however, there could be a leak in the pipe between the reservoir and the point at which the flow measurement is taken. It should be noted that while the nature of the fault can be seen (e.g., leak or blockage in feedwater line) this representation would not be very helpful in physically locating the leak within the plant (e.g., locating valve 1).

These examples illustrate that properly designed displays can change the fundamental type of behavior that is required of an operator under both normal and abnormal operating conditions. With separable displays (e.g., the separable configurations illustrated in Figure 13) the operators are required to engage in knowledge-based behaviors: They must rely on internal models of system structure and function (and therefore use limited capacity resources—working memory) to detect, diagnose, and correct faults. As a result, the potential for errors is increased dramatically. In contrast, properly designed configural displays present externalized models of system structure and function through geometric constraints. This allows operators to

utilize skill-based behaviors (e.g., visual perception and pattern recognition) that do not require limited capacity resources. As a result, the potential for errors will be decreased dramatically. As Rasmussen and Vicente (1989) have noted, changing the required behavior from knowledge-based behavior to rule- or skill-based behavior is a goal for display design.

Properly designed configural displays will also reduce the possibility of *underspecified action errors* (Rasmussen and Vicente, 1989). In complex, dynamic domains people can form incorrect hypotheses about the nature of the existing problem if they do not consider the relevant subsets of data (Woods, 1988). Observers may focus on these incorrect hypotheses and ignore disconfirming evidence, showing a kind of tunnel vision (Moray, 1981). Observers may also exhibit *cognitive hysteresis* and fail to revise hypotheses as the nature of the problem changes over time (Lewis and Norman, 1986). Configural displays that reflect the semantics of a domain directly can reduce the probability of these types of errors by forcing an observer to consider relevant subsets of data.

One final point needs to be made in closing this section. It is important to note that, even though multiple variables may testify with regard to higher level domain properties, configural displays that use geometric forms are *not* always the appropriate design solution. Ultimately, the design decision hinges on the relationships and interactions between variables. Consider the case of a mobile army commander engaged in tactical battlefield operations (Bennett et al., 2008). There are five combat resources that are the primary determinants of the higher level property of combat power: tanks, personnel carriers, ammunition, fuel, and personnel. The key design criterion is that these resources are essentially independent; there is no physical law or causal relationship that explicitly defines higher order patterns between variables. It is true that the combat resources can be correlated: A unit engaged in an intense offensive battle might suffer equipment and personnel losses while expending substantial amounts of fuel and ammunition. However, any number of factors can change the relationship between them. For example, ammunition is likely to be expended more quickly than fuel in a defensive mission.

The use of a highly configural representation, such as a polar graphic display (e.g., Woods et al., 1981) for the five combat resources, is a tempting, but inappropriate, design choice. This display would produce numerous, salient, and hierarchically nested emergent features that highlighted the relationships and interactions between combat resources. This produces a poor mapping between display and domain: The display constraints (i.e., the particular asymmetries and other distortions of the polar graphic) would not uniquely specify underlying domain constraints (i.e., higher level domain properties arising from the interaction of variables). Many, if not most, of the emergent features produced by the display would be essentially meaningless and would therefore need to be ignored (an extremely difficult, if not impossible, task for actors to accomplish). Bennett et al. (2008) chose a separable format similar to that

portrayed in Figure 10*b*. This representation maintains the independence of the individual combat resources while still providing limited configurality (Sanderson et al., 1989).

## 5.4   Practical Guidelines

In conclusion, we believe that the application of this approach to display design will improve overall human machine performance through the development of configural displays that support normal control as well as fault detection, diagnosis, and repair. The potential for errors will be decreased dramatically, because the critical information for control is represented directly in the interface. This, in turn, dramatically reduces the requirement for knowledge-based reasoning on the basis of internalized models. Recent applications of this approach include aviation displays (e.g., Amelink et al., 2005), anesthesiology (e.g., Drews and Westenskow, 2006), and process control (Lau et al., 2008). To summarize, we offer three general heuristics for analog, configural displays designed for law-driven domains:

1. Each relevant process variable should be represented by a distinct element within the display. If precise information about this variable is desirable, a reference scale and supplemental digital information should be provided.
2. The display elements should be organized so that the emergent properties (symmetries, closure, parallelism) that arise from their interaction correspond to higher order constraints within the process. Thus, when process constraints are broken (i.e., a fault occurs), the corresponding geometric constraints are also broken (the display symmetry is broken).
3. The symmetries within the display should be nested (from global to local) in a way that reflects the hierarchical structure of the process. High-order process constraints (e.g., at the level of functional purpose or abstract function) should be reflected in global display symmetries; lower order process constraints (e.g., functional organization) should be reflected in local display symmetries.

## 6   CHALLENGES OF COMPLEX SYSTEMS

The simple process described above is convenient for a tutorial introduction to some of the important decisions that must be made when designing a graphical representation. However, this example greatly underestimates the complexity seen in many advanced human–technological systems (e.g., nuclear power, air traffic control, advanced tactical aviation, command and control centers for managing military and space operations, minimally invasive and remote surgery). These systems typically have multiple modes of operation (each with different constraints and boundary conditions) and require multiple windows into the process. In these systems, the goal remains the same, to make the real constraints of the work process (at all levels of

abstraction) visible to the human operator. The designer must still address the problems of correspondence (so that all relevant process constraints are represented in the interface) and coherence (so that the representation is comprehensible to the human operator). For these complex systems, however, it will not be possible to achieve both correspondence and coherence with a single graphic display. Thus, the added problem of navigating through multiple views (i.e., windows, screens, pages) must be addressed.

A principal threat to these complex systems is *mode error* (Woods, 1984), which occurs when the operator loses track of dynamic changes in the operating constraints governing a process. The operator responds to one set of constraints (i.e., mode) when a different set of constraints is, in fact, governing the process. The design challenge is to coordinate the multiple windows necessary for complete representation with the changing operational modes. In simple terms, the problem is how the interface can be designed to ensure that the appropriate window is always coupled with the appropriate mode, that is, to ensure that the important information is salient at the appropriate times.

Two classes of solutions might be considered for dealing with the navigation problems that typically lead to mode errors, computational and graphical. Computational solutions or adaptive interfaces include an inference engine that manages the representation automatically. This computational engine adjusts the representation automatically based on inferences about the state of the system and the state of the operator. Projects such as the "pilot's associate" are examples of attempts to design automatic systems to aid operators to navigate through the representations associated with a complex work domain. However, the focus of this chapter is on graphical solutions. For this reason we use the remaining space to consider briefly some graphical approaches to this problem.

Woods (1984) introduced the term *visual momentum* to refer to the cognitive costs associated with switching from one reference frame to another. If visual momentum is high, the cost of switching views is low. In this case, the new display is consistent with expectations created by the prior display. If visual momentum is low, there is a high cost of switching. That is, the new display is not consistent with expectations and the cognitive system must effectively recalibrate before information can be extracted from the new display. To ensure high visual momentum, the design of each graphical display must be considered relative to the other displays that operators may be using. Are the graphical conventions [e.g., coordinates, scales, directions, motions, colors, stimulus–response (S-R) mappings] used in one display consistent with those in another?

A graphical device that Woods (1984) has suggested to increase visual momentum is the use of *landmarks*, graphical elements that provide an orientation point that relates one display to another. Just as a tall building or mountain that is visible from many different parts of the landscape might help a person to orient to the geography, graphical landmarks can be designed with the objective of aiding the operator to orient within the functional landscape of the work domain. For example, Aretz (1991) used a shaded wedge within an electronic map display as a landmark to specify the region within the map that corresponded to the head-up forward view of the pilot.

Another graphical device to help operators navigate across multiple display pages is a map or overview display. This display can be implemented as a separate window or as an embedded landmark in all windows. This overview might use a flow diagram or hierarchical tree structure to show functional links among the multiple display pages.

The concrete examples we have provided involve law-driven work domains and configural displays. We provide one brief description of an intent-driven domain and a metaphorical display. The BookHouse interface designed by Pejtersen (1980, 1992) uses a spatial metaphor in which rooms in a "house" are set up for various categories of users. This spatial metaphor allows the operator to apply natural abilities for navigating in three-dimensional spaces to the task of navigating in the more abstract space of a library database. For a more detailed treatment of intent-driven domains and metaphorical displays, consider Flach et al. (2011), who discuss ecological interface design for the Web. For a more detailed treatment of visual momentum, consider Bennett and Flach (2011), who provide additional discussion and concrete examples from several different work domains.

In the BookHouse, the three-dimensional space is implemented in a two-dimensional display. Virtual reality systems now offer the possibility for effective three-dimensional representations. With these systems, designers have the opportunity to maximize the transfer of natural human ability to orient and navigate in three-dimensional environments to more abstract environments and to combine natural three-dimensional representations with imagery obtained by advanced sensor systems. For example, virtual displays for minimally invasive surgery are being designed that integrate the three-dimensional image of a patient's anatomy with information obtained by magnetic resonance imaging (MRI) scans and other advanced imaging technologies. Thus, virtual three-dimensional spatial metaphors might provide another technique for integrating complex information from distributed sensors into a coherent representation.

The central theme of this chapter is that problem solving can be critically influenced by the nature of visual representations. Building effective representations requires designers to go beyond the simple psychophysical questions of data availability to the more complex questions of information availability, where information refers to the specification of domain constraints and boundary conditions. This specification depends both on the mapping from display to human (i.e., coherence) and that from display to domain (i.e., correspondence).

## 7 ACKNOWLEDGMENTS

# REFERENCES

Adelson, E. H. (1993), "Perceptual Organization and the Judgment of Brightness," *Science*, Vol. 262, pp. 2042–2044.

Alexander, K. R., Xie, W., and Derlacki, D. J. (1994), "Spatial Frequency Characteristics of Letter Identification," *Journal of the Optical Society of America, Series A*, Vol. 11, pp. 2375–2382.

Amelink, M. H. J., Mulder, M., van Paassen, M. M., and Flach, J. M. (2005), "Theoretical Foundations for a Total Energy-Based Perspective Flight-Path Display," *International Journal of Aviation Psychology*, Vol. 15, No. 3, pp. 205–231.

Arend, L. E., and Reeves, A. (1986), "Simultaneous Color Constancy," *Journal of the Optical Society of America, Series A*, Vol. 3, pp. 1743–1751.

Aretz, A. J. (1991), "The Design of Electronic Map Displays," *Human Factors*, Vol. 33, No. 1, pp. 85–101.

Barnett, B. J., and Wickens, C. D. (1988), "Display Proximity in Multicue Information Integration: The Benefits of Boxes," *Human Factors*, Vol. 30, No. 1, pp. 15–24.

Beltracchi, L. (1987), "A Direct Manipulation Interface for Heat Engines Based upon the Rankine Cycle," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 17, No. 3, pp. 478–487.

Beltracchi, L. (1989), "Energy, Mass, Model-Based Displays, and Memory Recall," *IEEE Transactions on Nuclear Science*, Vol. 36, No. 3, pp. 1367–1382.

Bennett, K. B. (1993), "Encoding Apparent Motion in Animated Mimic Displays," *Human Factors*, Vol. 35, No. 4, pp. 673–691.

Bennett, K. B., and Flach, J. M. (1992), "Graphical Displays: Implications for Divided Attention, Focused Attention, and Problem Solving," *Human Factors*, Vol. 34, No. 5, pp. 513–533.

Bennett, K. B. and Flach, J. M. (2011), *Display and Interface Design: Subtle Science, Exact Art*, CRC Press, Boca Raton, FL.

Bennett, K. B., and Madigan, E. (1994), "Contours and Borders in Animated Mimic Displays," *International Journal of Human-Computer Interaction*, Vol. 6, No. 1, pp. 47–64.

Bennett, K. B., and Malek, D. A. (2000), "Evaluation of Alternative Waveforms for Animated Mimic Displays," *Human Factors*, Vol. 42, No. 3, pp. 432–450.

Bennett, K. B., and Nagy, A. L. (1996), "Spatial and Temporal Considerations in Animated Mimic Displays," *Displays*, Vol. 17, No. 1, pp. 1–14.

Bennett, K. B., and Walters, B. (2001), "Configural Display Design Techniques Considered at Multiple Levels of Evaluation," *Human Factors*, Vol. 43, No. 3, pp. 415–434.

Bennett, K. B., Payne, M., Calcaterra, J., and Nittoli, B. (2000), "An Empirical Comparison of Alternative Methodologies for the Evaluation of Configural Displays," *Human Factors*, Vol. 42, No. 2, pp. 287–298.

Bennett, K. B., Posey, S. M., and Shattuck, L. G. (2008), "Ecological Interface Design for Military Command and Control," *Journal of Cognitive Engineering and Decision Making*, Vol. 2, No. 4, pp. 349–385.

Bonneau, G., Ertl, T., and Nielson, G. M., Eds. (2006), *Scientific Visualization: The Visual Extraction of Knowledge from Data*, Springer-Verlag, Berlin.

Boynton, R. M. (1990), "Human Color Perception," in *Science of Vision*, K. N. Leibovic, Ed., Springer-Verlag, New York, pp. 211–253.

Brainard, D. H., and Wandell, B. A. (1992), "Asymmetric Color-Matching: How Color Appearance Depends on the Illuminant," *Journal of the Optical Society of America, Series A*, Vol. 9, pp. 1433–1448.

Brodie, K. W., Carpenter, L. A., Earnshaw, R. A., Gallop J. R., Hubbold, R. J., Mumford, A. M., Osland, C. D., and Quarendon, P., Eds. (1992), *Scientific Visualization: Techniques and Applications*, Springer-Verlag, Berlin.

Burns, C. M., and Hajdukiewicz, J. R. (2004), *Ecological Interface Design*, CRC Press, Boca Raton, FL.

Buttigieg, M. A., and Sanderson, P. M. (1991), "Emergent Features in Visual Display Design for Two Types of Failure Detection Tasks," *Human Factors*, Vol. 33, No. 6, pp. 631–651.

Cannon, M. W., and Fullenkamp, S. C. (1991), "A Transducer Model for Contrast Perception," *Vision Research*, Vol. 31, pp. 983–998.

Carswell, C. M., and Wickens, C. D. (1987), "Information Integration and the Object Display," *Ergonomics*, Vol. 30, No. 3, pp. 511–527.

Carswell, C. M., and Wickens, C. D. (1990), "The Perceptual Interaction of Graphical Attributes: Configurality, Stimulus Homogeneity, and Object Integration," *Perception and Psychophysics*, Vol. 47, pp. 157–168.

Cleveland, W. S. (1985), *The Elements of Graphing Data*, Wadsworth, Belmont, CA.

DeValois, R. L., and DeValois, K. K. (1990), *Spatial Vision*, Oxford University Press, New York.

Drews, F. A., and Westenskow, D. R. (2006), "The Right Picture Is Worth a Thousand Numbers: Data Displays in Anesthesia," *Human Factors*, Vol. 48, No. 1, pp. 59–71.

D'Zmura, M., and Lennie, P. (1986), "Mechanisms of Color Constancy," *Journal of the Optical Society of America, Series A*, Vol. 3, pp. 1662–1672.

Earnshaw, R. A., and Wiseman, N., Eds. (1992), *An Introductory Guide to Scientific Visualization*, Springer-Verlag, Berlin.

Eskew, R. T., Jr. (2009), "Higher Order Color Mechanisms: A Critical Review," *Vision Research*, Vol. 49, pp. 2686–2704.

Flach, J. M., Bennett, K. B., Stappers, P. J., and Saakes, D. P. (2011), "An Ecological Perspective to Meaning Processing: The Dynamics of Abductive Systems," in *Handbook of Human Factors in Web Design*, 2nd ed., R. Proctor and K. L. Vu, Eds., CRC Press, Boca Raton, FL, pp. 509–526.

Garner, W. R. (1970), "The Stimulus in Information Processing," *American Psychologist*, Vol. 25, pp. 350–358.

Garner, W. R. (1974), *The Processing of Information and Structure*, Lawrence Erlbaum Associates, Mahwah, NJ.

Garner, W. R., and Felfoldy, G. L. (1970), "Integrality of Stimulus Dimensions in Various Types of Information Processing," *Cognitive Psychology*, Vol. 1, pp. 225–241.

Gegenfurtner, K. R., and Sharpe, L. T. (1999), *Color Vision: From Genes to Perception*, Cambridge University Press, Cambridge.

Gilchrist, A., Delman, S., and Jacobson, A. (1983), "The Classification and Identification of Edges as Critical to the Perception of Reflectance and Illumination," *Perception and Psychophysiology*, Vol. 33, pp. 425–436.

Ginsburg, A. (1986), "Spatial Filtering and Visual Form Perception," in *Handbook of Human Perception and Performance*, K. R. Boff, L. Kaufmann, and J. P. Thomas, Eds., Wiley, New York, pp. 34/1–34/41.

Graham, N. V. S. (1989), *Visual Pattern Analyzers*, Oxford University Press, New York.

Grum, F., and Bartleson, C. J. (1980), *Optical Radiation Measurements*, Vol. 2, Colorimetry, Academic, New York.

Hansen, T., and Gegenfurtner, K. R. (2006), "Higher Level Chromatic Mechanisms for Image Segmentation," *Journal of Vision*, Vol. 6, pp. 239–259.

Hollan, J. D., Hutchins, E. L., and Weitzman, L. (1984, Summer), "Steamer: An Interactive Inspectable Simulation-Based Training System," *AI*, pp. 15–27.

Hollan, J. D., Hutchins, E. L., McCandless, T. P., Rosenstein, M., and Weitzman, L. (1987), "Graphical Interfaces for Simulation," in *Advances in Man-Machine Systems Research*, Vol. 3, W. B. Rouse, Ed., JAI Press, Greenwich, CT, pp. 129–163.

Hunter, R. S., and Herold, R. W. (1987), *The Measurement of Appearance*, Wiley, New York.

Hutchins, E. L., Hollan, J. D., and Norman, D. A. (1986), "Direct Manipulation Interfaces," in *User Centered System Design*, D. A. Norman, and S. W. Draper, Eds., Lawrence Earlbaum Associates, Mahwah, NJ, pp. 87–124.

Kaiser, P. K., and Boynton, R. M. (1997), *Human Color Vision*, Optical Society of America, Washington, DC.

Kelly, D. H. (1974), "Spatio-Temporal Frequency Characteristics of Color-Vision Mechanisms," *Journal of the Optical Society of American, Series A*, Vol. 64, pp. 983–990.

Klein, G. A., Orasanu, J., and Zsambok, C. E., Eds. (1993), *Decision Making in Action: Models and Methods*, Ablex, Norwood, NJ.

Lau, N., Jamieson, G. A., Skraaning, G., and Burns, C. M. (2008), "Ecological Interface Design in the Nuclear Domain: An Empirical Evaluation of Ecological Displays for the Secondary Subsystems of a Boiling Water Reactor Plant Simulator," *IEEE Transactions on Nuclear Science*, Vol. 55, No. 6, pp. 3597–3610.

Lennie, P., and Movshon, J. A. (2005), "Coding of Color and Form in the Geniculostriate Pathway," *Journal of the Optical Society of America A*, Vol. 22, pp. 2013–2033.

Lewis, C., and Norman, D. A. (1986), "Designing for Error," in *User Centered System Design*, D. A. Norman and S. W. Draper, Eds., Lawrence Earlbaum Associates, Mahwah, NJ.

Makous, W. (2003), "Threshold and Suprathreshold Spatiotemporal Contrast Sensitivity," in *Encyclopedia of Optical Engineering*, Marcel Dekker, New York, pp. 2828–2850.

Maloney, L. T., and Wandell, B. A. (1986), "Color Constancy: A Method for Recovering Surface Reflectance," *Journal of the Optical Society of America, Series A*, Vol. 3, pp. 29–33.

Millodot, M. (1982), "Image Formation in the Eye," in *The Senses*, H. Barlow and J. D. Mollen, Eds., Cambridge University Press, New York, pp. 46–61.

Moray, N. (1981), "The Role of Attention in the Detection of Errors and the Diagnosis of Failures in Man-Machine Systems," in *Human Detection and Diagnosis of System Failures*, J. Rasmussen and W. B. Rouse, Eds., Plenum, New York.

Moray, N., Lee, J., Vicente, K. J., Jones, B. G., and Rasmussen, J. (1994), "A Direct Perception Interface for Nuclear Power Plants," in *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 481–485.

Mullen, K. (1985), "The Contrast Sensitivity of Human Color Vision to Red-Green and Blue-Yellow Chromatic Gratings," *Journal of Physiology*, Vol. 359, pp. 381–400.

Nagy, A. L. (2003), "Visual Search: Color and Display Issues," in *Encyclopedia of Optical Engineering*, R. G. Driggers and C. Hoffman, Eds., Marcel Dekker, New York, pp. 2955–2962.

Nagy, A. L., and Kamholz, D. (1995), "Luminance Discrimination, Color Contrast, and Multiple Mechanisms," *Vision Research*, Vol. 35, pp. 2147–2155.

Nassau, K. (1983), *The Physics and Chemistry of Color*, Wiley, New York.

Newell, A., and Simon, H. A. (1972), *Human Problem Solving*, Prentice Hall, Englewood Cliffs, NJ.

Noorlander, C., and Koenderink, J. J. (1983), "Spatial and Temporal Discrimination Ellipsoids in Colour Space," *Journal of the Optical Society of America, Series A*, Vol. 73, pp. 1533–1543.

Olzack, L., and Thomas, J. (1986), "Seeing Spatial Patterns," in *Handbook of Human Perception and Performance*, K. R. Boff, L. Kaufmann, and J. P. Thomas, Eds., Wiley, New York, pp. 7/1–7/56.

Pavel, M., and Ahumada, A. J. (1997), "Model Based Optimization of Display Systems," in *Handbook of Human-Computer Interaction*, 2nd ed., M. G. Helander, T. K. Landauer, and P. V. Prabhu, Eds., Elsevier, New York, pp. 65–86.

Pejtersen, A. M. (1980), "Design of a Classification Scheme for Fiction Based on an Analysis of Actual User-Librarian Communication, and Use of the Scheme for Control of Librarians' Search Strategies," in *Theory and Application of Information Research*, O. Harbo and L. Kajberg, Eds., Mansell, London, pp. 167–183.

Pejtersen, A. M. (1992), "The BookHouse: An Icon Based Database System for Fiction Retrieval in Public Libraries," in *The Marketing of Library and Information Services*, Vol. 2, B. Cronin, Ed., Aslib, London, pp. 572–591.

Poirson, A. B., and Wandell, B. A. (1993), "The Appearance of Colored Patterns," *Journal of the Optical Society of America, Series A*, Vol. 12, pp. 2458–2471.

Pokorny, J., and Smith, V. (1986), "Colorimetry and Color Discrimination," in *Handbook of Human Perception and Performance*, K. R. Boff, L. Kaufmann, and J. P. Thomas, Eds., Wiley, New York, pp. 8/1–8/51.

Pomerantz, J. R. (1986), "Visual Form Perception: An Overview," in *Pattern Recognition by Humans and Machines*, Vol. 2, *Visual Perception*, H. C. Nusbaum and E. C. Schwab, Eds., Academic, Orlando, FL, pp. 1–30.

Pomerantz, J. R., and Pristach, E. A. (1989), "Emergent Features, Attention, and Perceptual Glue in Visual Form Perception," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 15, No. 4, pp. 635–649.

Pomerantz, J. R., Sager, L. C., and Stoever, R. J. (1977), "Perception of Wholes and of Their Component Parts: Some Configural Superiority Effects," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 3, pp. 422–435.

Post, D. (1992), "Colorimetric Measurement, Calibration, and Characterization of Self-Luminous Displays," in *Color in Electronic Displays*, H. Widdel and D. L. Post, Eds., Plenum, New York, pp. 299–312.

Post, D. (1997), "Color and Human Computer Interaction," in *Handbook of Human-Computer Interaction*, 2nd ed., M. G. Helander, T. K. Landauer, and P. V. Prabhu, Eds., Elsevier, New York, pp. 573–616.

Rasmussen, J. (1986), *Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering*, Elsevier, New York.

Rasmussen, J., and Vicente, K. (1989), "Coping with Human Errors through System Design: Implications for Ecological Interface Design," *International Journal of Man-Machine Studies*, Vol. 31, pp. 517–534.

Rasmussen, J., Pejtersen, A. M., and Goodstein, L. P. (1994), *Cognitive Systems Engineering*, Wiley, New York.

Sanderson, P. M., Flach, J. M., Buttigieg, M. A., and Casey, E. J. (1989), "Object Displays Do Not Always Support Better Integrated Task Performance," *Human Factors*, Vol. 31, No. 2, pp. 183–198.

Sekiguchi, N., Williams, D. R., and Brainard, D. H. (1993), "Aberration Free Measurements of the Visibility of Isoluminant Gratings," *Journal of the Optical Society of America, Series A*, Vol. 10, pp. 2105–2117.

Shafer, S. A. (1985), "Using Color to Separate Reflection Components," *Color Research and Applications*, Vol. 10, pp. 210–218.

Shneiderman, B. (1986), *Designing the User Interface*, Addison-Wesley, Reading, MA.

Shneiderman, B., Ed. (1993), *Sparks of Innovation in Human-Computer Interaction*, Ablex, Norwood, NJ.

Solomon, J. A., and Pelli, D. G. (1994), "The Visual Filter Mediating Letter Identification," *Nature (London)*, Vol. 369, pp. 395–397.

Travis, D. (1991), *Color Displays: Theory and Practice*, Academic, London.

Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Graphics, Chesire, CT.

Tufte, E. R. (1990), *Envisioning Information*, Graphics, Chesire, CT.

Vicente, K. J. (1991), "Supporting Knowledge Based Behavior through Ecological Interface Design," Technical Report EPRL-91-1, Engineering Psychology Research Laboratory and Aviation Research Laboratory, University of Illinois, Urbana-Champaign, IL.

Vicente, K. J. (1999), *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*, Lawrence Erlbaum Associates, Mahwah, NJ.

Wandell, B. A. (1995), *Foundations of Vision*, Sinauer Associates, Sunderland, MA.

Webster, M. A., Werner, J. S., and Field, D. J. (2005), "Adaptation and the Phenomenology of Vision," in *Fitting the Mind to the World: Adaptation and After-effects in High Level Vision*, *Advances in Cognition Series*, Vol. 2, C. Clifford and G. Rhodes, Eds., Oxford University Press, Oxford, pp. 241–277.

Wertheimer, M. (1959), *Productive Thinking*, Harper & Row, New York.

Whittle, P. (1986), "Increments and Decrements: Luminance Discrimination," *Vision Research*, Vol. 26, pp. 1677–1692.

Whittle, P. (1992), "Brightness, Discriminability, and the 'Crispening Effect,'" *Vision Research*, Vol. 32, pp. 1493–1508.

Wickens, C. D., and Andre, A. D, (1990), "Proximity Compatibility and Information Display: Effects of Color, Space, and Objectness on Information Integration," *Human Factors*, Vol. 32, No. 1, pp. 61–78.

Wickens, C. D., and Carswell, C. M. (1995), "The Proximity Compatibility Principle: Its Psychological Foundation and Its Relevance to Display Design," *Human Factors*, Vol. 37, pp. 473–494.

Widdel, H., and Post, D. L. (1992), *Color in Electronic Displays*, Plenum, New York.

Woods, D. D. (1984), "Visual Momentum: A Concept to Improve the Cognitive Coupling of Person and Computer," *International Journal of Man-Machine Studies*, Vol. 21, pp. 229–244.

Woods, D. D. (1988), "Coping with Complexity: The Psychology of Human Behavior in Complex Systems," in *Mental Models, Tasks and Errors: A Collection of Essays to Celebrate Jens Rasmussen's 60th Birthday*, L. P. Goodstein, H. B. Andersen, and S. E. Olsen, Eds., Taylor & Francis, New York.

Woods, D. D. (1991), "The Cognitive Engineering of Problem Representations," in *Human-Computer Interaction and Complex Systems*, G. R. S. Weir and J. L. Alty, Eds., Academic, London, pp. 169–188.

Woods, D. D., and Roth, E. M. (1988), "Cognitive Systems Engineering," in *Handbook of Human-Computer Interaction*, M. Helander (Ed.), Elsevier Science, Amsterdam, pp. 1–41.

Woods, D. D., Wise, J. A., and Hanes, L. F. (1981), "An Evaluation of Nuclear Power Plant Safety Parameter Display Systems," in *Proceedings of the Human Factors Society 25th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 110–114.

Worthy, J. A., and Brill, M. H. (1986), "Heuristic Analysis of Color Constancy," *Journal of the Optical Society of America, Series A*, Vol. 3, pp. 1708–1712.

Wyszecki, G., and Stiles, W. S. (1982), *Color Science*, 2nd ed., Wiley, New York.

Zachary, W. (1986), "A Cognitively Based Functional Taxonomy of Decision Support Techniques," *Human-Computer Interaction*, Vol. 2, pp. 25–63.

# CHAPTER 43

# INFORMATION VISUALIZATION

**Chris North**
**Virginia Polytechnic Institute and State University**
**Blacksburg, Virginia**

## 1 INTRODUCTION

The information revolution is changing the way that many people live and think. Vast quantities and diverse types of information are being generated, stored, and disseminated, raising serious issues about how to make such information usable. The need to understand and extract knowledge from stored information is becoming a ubiquitous task. As examples, in everyday life people must sort through a variety of personal information, such as e-mail communications, schedules, news, finances, and social media. Students can access countless digital libraries of educational materials. Online shoppers must make decisions among dozens of alternative products, models, vendors, and prices. New disciplines such as bioinformatics are leading the revolution in information-intensive science using high-throughput data collection technologies such as microarrays and online data repositories. Government intelligence analysts must sift through massive collections of information gathered on a daily basis from sensor networks and other sources.

*Information visualization* has evolved as an approach to make large quantities of complex information intelligible. An information visualization is a visual user interface to information, with the goal of providing users with information *insight* (Spence, 2001). The basic method is to generate interactive visual representations of the information that exploit the perceptual capabilities of the human visual system and the interactive capabilities of the cognitive problem-solving loop (Ware, 2004).

The goal of this chapter is to highlight the critical high-level design decisions in the information visualization design process. Lower level details of visual display and human perception can be found elsewhere in this book. Other aspects of the design process that apply to the design of user interfaces in general, such as evaluation methods, are also covered in other chapters. While other major references focus on the "what" and "why" of information visualization (Card et al., 1999; Chen, 1999; Wickens and Hollands, 2000; Spence, 2001; Ware, 2004; Shneiderman and Plaisant, 2005), here we emphasize the "how."

### 1.1 Insight

Human vision contains millions of photoreceptors and is capable of rapid parallel processing and pattern recognition (Ware, 2004). The impressive bandwidth of vision as a mode of communication leads to the efficient transfer of data from digital storage to human mind. Yet a more important benefit is the human ability to reason visually about the data and extract higher level

knowledge, or insight, beyond simple data transfer (Card et al., 1999). This enables people to infer new mental models of the real phenomena represented by the data.

For example, Figure 1 demonstrates the mapping of a database of census demographics. From the visual representation in scatterplot form, one can readily recognize the approximate proportional relationship between education and income, various outliers, such as New York, NY, and the predominance of large population counties in high income and education. These insights are not explicitly stored within the data set but are inferred through visual pattern recognition. These insights are not so readily identifiable from the textual spreadsheet representation. Clearly, appropriate design choices in

the visual representation are important to enabling this insight. A poorly designed visualization can hide this insight or even mislead with incorrect insight.

Researchers increasingly recognize that the benefit of visualization goes beyond tapping into human visual abilities to human interactive abilities (Thomas and Cook, 2005; Fekete et al., 2008). Visualizations provide humans with a medium for interacting with information. Cognitive psychology theories of embodied interaction (Wilson, 2002) and distributed cognition (Liu et al., 2008) suggest that insight is gained through the interactive dialogue that takes place between the user and the visualization. In the example in Figure 1, the perceived relationship between income and education might



(a)



(b)



(c)  (d)  (e)  (f)

**Figure 1** Census demographics data set of 3140 U.S. counties shown in (a) spreadsheet form and (b) scatterplot form using Spotfire. The plot shows counties by income per capita vs. percentage of adult population that has college degree, with dots sized by population, and labels for some outliers. The plot is interactive and reveals details of a county when the user clicks on a dot. Dots can be filtered by other county attributes, such as median rental cost of housing, using the dynamic query slider widgets on the right. The plots at bottom show the result when filtering median rent to (c) low, (d) medium–low, (e) medium–high, and (f) high. [From Ahlberg and Wistrand (1995). Courtesy of Spotfire.]

prompt one to investigate how cost of living relates. Interactive filtering reveals an interesting animated trend on the median cost of rental housing. Low rent is limited to areas that are low income and low education. Increasing rent first extends to high-income, low-education areas, then to low-income, high-education areas, then finally to high-income, high-education areas. This additional insight was enabled by the interactive affordances of the visualization. The design of the interaction is important to allow users to selectively pursue many possible lines of investigation, depending on new questions that arise from discoveries made, in accordance with their thought processes.

Visualization can enable a broad range of information insight, and several categories of such insights are listed below (see also Wehrend and Lewis, 1990; Zhou and Feiner, 1998; Wickens and Hollands, 2000; Shneiderman and Plaisant, 2005; Amar et al., 1995). The first two are simplistic and can be supported readily by textual or query-based user interfaces such as spreadsheets or search forms, because they are precisely defined and have solutions consisting of a single data entity. However, the latter are more complex and are well supported by visualization. These involve open-ended questions with complex answers that require seeing the whole. A strength of visualization is the capacity for discovery, the recognition of new insights unexpected by the users and potentially unforeseen by the visualization designers.

Simple insights:

- *Summaries:* minimum, maximum, average, percentages
- *Find:* known item search

Complex insights:

- *Patterns:* distributions, trends, frequencies
- *Relationships:* correlations, multiway interactions
- *Trade-offs:* balance, combined minimum/maximum
- *Comparisons:* choices (1:1), context ($1 : M$), sets ($M : N$)
- *Clusters:* groups, similarities
- *Structures:* breadth, depth, decompositions, order
- *Paths:* distance, multiple routes, connectedness
- *Outliers:* exceptions
- *Anomalies:* data errors

## 1.2 Design

Like any user interface, effective information visualizations are difficult to design. Fundamentally, information visualizations make *abstract information* perceptible. Abstract information has no inherent perceptual form, as in the case of databases or computer directories. Hence, there are no natural constraints on the types of visual representations and interactions that creativity can produce for abstract information, and the possibilities are limitless. As a result, there is significant

challenge, excitement, and opportunity both in creating novel visual representations and in identifying the most effective representations among the endless possibilities. In contrast, *scientific visualization* (Rosenblum et al., 1994) typically emphasizes the visualization of data that represent physical three-dimensional phenomena, which offer some natural constraints on visual representations and focus the challenges on realism.

The two most challenging characteristics of information that make designing effective information visualizations difficult are (1) the *complexity* of diverse abstract information that may have multiple interrelated data types and structures and (2) the *scalability* to very large quantities of information. Because of these characteristics, visual representations alone are not sufficient, and interactive techniques must also be designed. Although the design principles of static graphs and illustrations are fundamental to visualization design (Cleveland, 1993; Gillan et al., 1998; Wilkinson, 1999; Tufte, 2001), new human–computer interaction issues related to these two challenging characteristics come to the forefront in information visualization design.

As an overview, the visualization design process involves iterative requirements analysis, design, and evaluation (e.g., Rosson and Carroll, 2001). In the requirements analysis phase, it is important to identify the two primary inputs to design: (1) the characteristics of the information to be visualized and (2) the types of insights that the visualization should enable. Characteristics of the information include the data schema, underlying information structures, and data quantity. Since the number of data attributes and desired insights may be large, identifying a prioritization of attributes and insights will be helpful in balancing design trade-offs. Other elements of requirements analysis include broader user tasks, users' domain knowledge, data semantics, and computer system requirements. In the design phase, major design decisions (presented in this chapter) include the visual mapping of the information, the representation of information structures, visual overview strategies, navigation strategies, and interaction techniques.

The evaluation phase must be considered continually during the design process (Plaisant, 2004). A claims analysis identifies the positive and negative impacts of a visualization design's features on its insight capability and seeks to overcome or balance these trade-offs through iterative design (Rosson and Carroll, 2001). Begin with analytic evaluations to determine if designs meet requirements, such as scalability to data quantity and appropriateness for producing desired insights. In later iterations, empirical evaluations involving users should be undertaken, such as the "wizard of Oz" technique, usability testing, or controlled experiments (Chen and Yu, 2000; Tory and Möller, 2004a). Desired insights identified in requirements analysis should be implemented as benchmark user tasks in the empirical evaluations. Alternatively, since benchmark tasks often overly constrain the testing to simplistic insights that discount the discovery aspect of visualization, the insight-based methodology (Saraiya et al., 2004) attempts to measure the insight generated by visualizations by using an

open-ended experimental protocol without benchmark tasks. In the following sections we highlight the major design decisions in the information visualization design process.

## 2   VISUALIZATION PIPELINE

The *visualization pipeline* is the computational process of converting information into a visual form with which users can interact (Card et al., 1999) (Figure 2). The first step is to transform raw information into a well-organized canonical data format. The resulting format typically consists of a data set containing a set of data entities each of which has associated data attribute values. Various data-processing steps can be used to manipulate the data as needed. Derived data, such as data mining or clustering results, can be very useful for assisting in insight generation (Fayyad et al., 2001). The second step, the heart of the visualization process, is to map the data set into visual form. The visual form contains visual glyphs that correspond to the data set entities. The third step embeds this visual form into views, which display the visual form on screen and provide various view transformations, such as navigation. The view is then presented to the user through the human visual system. Users interpret the view to (partially) mentally reconstruct the underlying information. Finally, users can interact with any of the steps in the pipeline to alter the resulting visualization and make further interpretations. This entire pipeline comprises an information visualization.

### 2.1   Visual Mapping

The visual mapping at the second step is the heart of visualization and must be designed carefully. The goal is to communicate information from computer to human. The medium of communication is a visual representation of the information. The data set is mapped computationally into visual form by some function $f$, which takes the data set as input and generates the visual representation as output. Then, when the visual representation is communicated to users, they must cognitively reverse the visual mapping by inverting function $f$ to decode the information from the visual representation. It is yet unclear how, when, and to what degree $f^{-1}$ is cognitively applied in the perceptual process, and a variety of models exist (Ware, 2004). Although some cognitive reasoning operates on the visual representation itself, eventually meaning must be decoded. Nonetheless, this visual communication process implies four important characteristics of the visual mapping function $f$:

1. *Computable*. $f$ is a mathematical function that can be computed by some algorithm. Although there is significant room for creativity in the design of these functions, execution of the functions must be algorithmic.

2. *Invertible*. It must be possible to use $f^{-1}$, the inverse of mapping function $f$, to reconstruct the data from the visual representation to a desired degree of accuracy. If this is not possible, the visualization will be ambiguous, misleading, or not interpretable.

3. *Communicable*. $f$ (or preferably $f^{-1}$) must be known by the user to decode the visual representation. It must be communicated with the visualization or already known by the user through prior experience. In usability terms, this is a learnability issue.

4. *Cognizable*. $f^{-1}$ should minimize cognitive load for decoding the visual representation. This is a human perception and performance issue.

The visual mapping step is accomplished by two sub-steps (Card et al., 1999) (Figure 3). First, each data entity is mapped into a visual *glyph*. The vocabulary of possible glyphs consists primarily of points (dots, simple shapes), lines (segments, curves, paths), regions (polygons, areas, volumes), and icons (symbols, pictures). Second, attribute values of each data entity are mapped onto *visual properties* of the entity's glyph. Common visual properties of glyphs include spatial position ($x$, $y$, $z$), size (length, area, volume), color (gray scale, hue, intensity), orientation (angle, slope, unit vector), and shape. Other visual properties include texture, motion, blink frequency, density, and transparency. For example, in Figure 1, U.S. counties are mapped to circular points. Income and education levels of each county are mapped to the horizontal and vertical position of the point, respectively, and population value is mapped to the size of the point.

### 2.2   Visual Properties

In general, data attributes should be prioritized according to the problem requirements and desired insights. The prioritization can then be applied to map the higher priority data attributes to the most effective visual properties. Spatial position properties are the most effective and should be reserved to lay out the data set in the visual representation according to the most important data attributes.
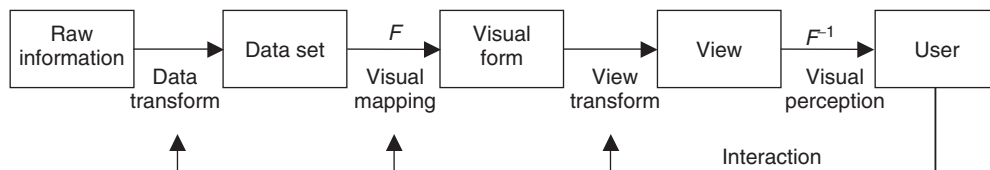


**Figure 2**   Visualization pipeline, converting information into interactive visual representations. (Adapted from Card et al., 1999.)

## Visual Glyphs

Points

Lines

Regions

Icons

## Visual Properties of Glyphs

Position

Size

Color

Orientation

Shape
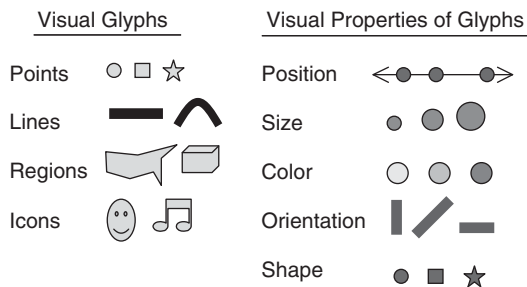
**Figure 3** Vocabulary of glyphs and some visual properties of glyphs. (Adapted from Card et al., 1999.)

The remaining visual properties, called *retinal properties* (Bertin, 1983), can be used next. The effectiveness of these properties are determined by many interdependent factors, including preattentive processing (Healey et al., 1996), perceptual independence and separability (Ware, 2004), data type (quantitative, ordinal, categorical) (Card et al., 1999), polarity (greater than, less than) (Ware, 2004), task (Carswell, 1992; Wickens and Hollands, 2000), and attention (Chewar et al., 2002). Commonly accepted orderings of the effectiveness of these attributes are based on empirical evidence (e.g., Cleveland and McGill, 1984; Nowell et al., 2002) as well as experience (Bertin, 1983; Mackinlay, 1986).

The order shown in Figure 3, most effective on top, is intended for quantitative data. To interpret quantitative data visually, effective visual properties should enable users to visually estimate order and ratios of multiple data values. Spatial properties are most effective because the human visual system can accurately judge spatial ratios. Color maps are less effective and must be carefully designed (Harrower and Brewer, 2003). Rainbow color maps are not effective, because they lack perceptual ordering and linear scale. For categorical data, color and shape become more effective. Categorical data only requires users to be able to distinguish groups. Unique hues assigned to each group are an effective approach. Some visualization design systems attempt to use such rules to generate effective mappings automatically [e.g., Apt (Mackinlay, 1986), Sage (Roth et al., 1994), and Tableau (Tableau Software, 2010)].

Finally, for any remaining data attributes, interaction techniques can be applied. In general, direct visual mapping of information is most effective for rapid insight, but only a limited number of data attributes can be simultaneously mapped, whereas interaction techniques require slower physical actions by the user to reveal insights but are essentially unlimited. When mapping additional data attributes would overly clutter the visual representation and reduce comprehension of more important attributes, interaction techniques can be used instead. Interaction techniques enable users to alter the visual mapping function or other stages of the visualization pipeline, based on additional data attributes. By viewing the resulting changes in the visual representation, users can infer additional information about those attributes over time. For example, the *dynamic queries*

technique provides interactive query widgets for unmapped data attributes and can be used to dynamically filter the entity glyphs (Ahlberg and Wistrand, 1995) (Figure 1).

## 3 INFORMATION STRUCTURE

The visual mapping process provides an initial starting point for visualization design, but more advanced methods are needed as data complexity increases. Identifying underlying structures within the target information helps to further guide the design process. These structures provide high-level organization to a data set and often provide guidance for the design of appropriate visualizations. Since these structures are likely to be very important to users' mental models of the information, they are typically mapped to the spatial position attributes and form the primary layout of the visualization. In general, there are four common classes of information structures [adapted from Shneiderman 1996, Card et al. (1999), and Spence 2001], as discussed in Sections 3.1–3.4. These are not strict or mutually exclusive classifications but useful guides.

### 3.1 Tabular Structure

Data tables consist of rows (*entities*) and columns (*attributes*). This is often referred to as *multidimensional data*, because each attribute defines a dimension of the data space within which each entity identifies a single point. Examples include databases and spreadsheet tables, such as the census data in Figure 1. Visualizations of tables that contain a small number of attributes can be designed relatively easily using the visual mapping process described in Section 2. However, such visualizations lack scalability to many attributes, due to the limited number of nonconflicting visual properties from which to choose. To address this problem, a variety of creative methods have been developed for tables of many attributes. Primarily, these involve the use of more complex glyphs and spatial layouts.

First, heatmaps (e.g., Saraiya et al., 2004) preserve the tabular spreadsheet visual representation but represent each cell of the table as a simple colored square by using a color scale to map the data values. This offers users a familiar visual structure in a highly compact visual form. TableLens (Rao and Card, 1994) (Figure 4*a*) uses a similar approach but converts cells to horizontal bar glyphs with cell values mapped to bar length. This exploits the length property, which is better than color for encoding quantitative data. Also, since the bars are very thin, many values can be packed onto the screen, providing an excellent overview of a long tabular data set. TableLens encodes each data entity (row) with multiple glyphs (bars), one glyph for each of the entity's attribute values (columns). Interactively selecting a set of rows will expand them to reveal the detailed data values in textual form. Users can vertically sort the table by any attribute. By spatially sorting the data according to one attribute, distributions and relationships to other attributes can be seen. However, it is perceptually difficult to relate two unsorted attributes. Hence, users must sort each attribute interactively to explore all potential relationships.

(a)



(b)
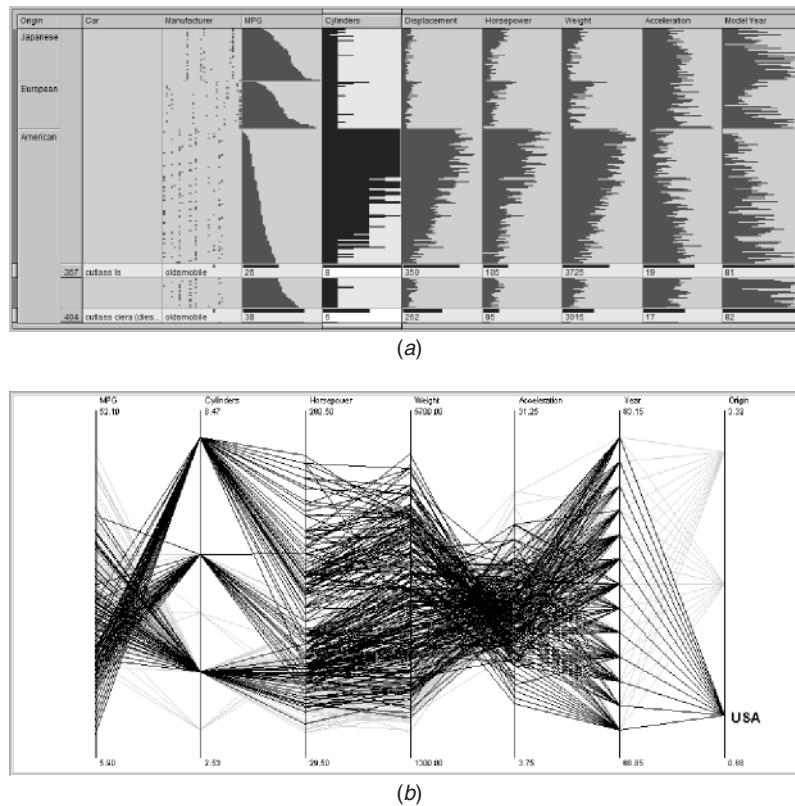
**Figure 4** (a) InXight TableLens; (b) XmdvTool's Parallel Coordinates. Both show the same data set about automobiles. Clearly, the United States dominates the market in large, powerful, gas-guzzling cars. Two outliers, cars with large engines yet high MPG, are highlighted in TableLens, revealing their detailed values. TableLens sorts the data by one attribute, whereas Parallel Coordinates sorts all the attributes simultaneously. [(a) From Rao and Card (1994). Courtesy of InXight). (b) From Ward (1994). Courtesy of Matthew Ward.]

The *proximity compatibility principle* (Wickens and Hollands, 2000) predicts that representations that use a single glyph per data entity, such as in scatterplots (Figure 1), would be better than TableLens for recognizing relationships between attributes. But as indicated previously, such representations are more limited in scalability. Herein is the trade-off: TableLens provides an overview of many attributes with reasonable capability for relationship insights, whereas scatterplots provide excellent insight on relationships but only for the two attributes mapped to *x* and *y* (and potentially a small number of other attributes using color, size, etc.).

To analyze the scalability of heatmaps, consider an approximate screen resolution of $1000 \times 1000$ pixels. If each colored cell requires a $10 \times 10$-pixel area (about the size of one typed character) to be recognizable, then a heatmap could display $100 \times 100 = 10,000$ cells. For TableLens, if each bar glyph is only 1 pixel thick, 1000 data entities (rows) can be shown. Columns will need to be approximately 50–100 pixels wide to enable reasonable visual discretion of quantitative data such as percentages, resulting in 10–20 attributes being visible. Hence, TableLens can display a tabular

data set containing 1000 data entities and 20 attributes. Much larger data sets can be explored in TableLens by using its aggregation and interactive navigation (e.g., scrolling) strategies, but only $1000 \times 20$ values are visible simultaneously. For data sets larger than the screen size, TableLens aggregates adjacent rows by showing averages or minimum and maximum values so as to reduce the data to fit the screen. Because of their scalability, heatmaps are a popular approach for data tables that have a large number of attributes (on the order of 100 columns).

A second approach is to use nonorthogonal axes. The Cartesian coordinate system uses orthogonal axes to visually map two or three attributes of a tabular data set to space (Figure 1). However, orthogonal axes fundamentally limit scalability of the number of attributes. As an alternative, Parallel Coordinates (Inselberg, 1997) (Figure 4b) displays attribute axes as parallel vertical lines. Each data entity is then mapped to a polyline that connects the entity's attribute values on each attribute axis. Hence, data attributes are mapped to the vertical position of the respective vertices of the polyline. Users can recognize clusters of similar

entities and relationships between adjacent attributes. Patterns of crossing lines between adjacent axes indicate an inverse relationship between those two attributes, and noncrossing lines indicate a proportional relationship. To combat occlusion and clutter, interactively selecting entities highlights their polylines across all axes. A scalability analysis of Parallel Coordinates would give a result similar to that of TableLens. However, it can potentially display many more rows because the lines overlap. Other possible arrangements of axes include radial (Kandogan, 2000) and circumferential (Miller, 2004). Parallel Coordinates is a popular approach for data tables that have many rows (over 1000), because it emphasizes viewing patterns between attributes and recognizing clusters of entities in the data rather than individual entities.

A third approach is to use more complex iconic glyphs. For example, star plots (Chambers et al., 1983) map data attributes to the length of the radial needles emanating from a star icon, and Chernoff faces (Chernoff, 1973) attempt to exploit the human ability to rapidly recognize facial features and expressions. Although these iconic methods do not scale up very well, they are frequently used when combining with other information structures (such as networks) because they leave the spatial position properties available for other uses (Ward, 2002).

A fourth approach is to simplify the glyphs and visual representations by splitting the attributes up into *multiple views*. For example, four attributes can be displayed using the *x* and *y* axes of two separate scatterplots. Scatterplot matrices take this approach to the extreme, displaying many plots for all possible combinations of attribute pairs (Cleveland, 1993) as a large matrix of small plots. An interactive technique called *brushing and linking* connects the plots (Becker and Cleveland, 1987). When users select glyphs in one plot (*brushing*), the corresponding glyphs for the same underlying data entities are highlighted in the other plot (*linking*). The multiple views with brushing-and-linking technique is frequently used to view a single data set in several different visual representations at the same time, such as plots and geographic maps, to relate the various contexts (Roth et al., 1996; North et al., 2002) (see Section 6.2).

## 3.2 Spatial and Temporal Structures

Spatial and temporal structures have a strong one-, two-, or three-dimensional component in which navigation is likely to be required. One-dimensional (1D) examples include time lines, music, video streams, lists, linear documents, and slide shows. Two-dimensional (2D) examples are road maps, satellite images, photographs, and blueprints. Three-dimensional (3D) examples are magnetic resonance imaging (MRI) and computed-tomography (CT) medical scans, computer-aided design/manufacturing (CAD/CAM) architectural plans, and virtual environments. Continuous functions, including those with domains greater than three dimensions, also fall in this category (Tory and Möller, 2004b). These spatial and temporal structures are the most natural for mapping onto spatial displays.

For example, the Music Animation Machine (Malinowski, 2004) (Figure 5a) provides a simple time line representation of music that scrolls as the music plays. Notes are represented as horizontal bars, with vertical position representing pitch, horizontal position representing timing, length indicating duration, and color indicating other attributes, such as instrument, timbre, or hand (in the case of piano). Similarly, LifeLines (Plaisant et al., 1996) represents events in a person's medical history, but in a more compact form, with zooming for navigation. For time lines that contain periodic cycles, such as calendars, visual spirals can be used to proximate the cycles while maintaining a continuous line (Carlis and Konstan, 1998). Streaming video can be viewed as a 3D video cube (2D frames + 1D time) (Elliott and Davenport, 1994).

In 3D data spaces, the main challenge is viewing the interior of the 3D structure beyond the exterior surface when occlusion is problematic. Architectural walk-through applications (polygonal data) typically use first-person perspective projection, with six-degree-of-freedom navigation for a lifelike experience (Stoakley et al., 1995). For medical imagery (volumetric data), strategies include slicing and transparency. For example, the Visible Human Explorer presents 2D slices that can be animated through the 3D body (North et al., 1996). In 3D volume rendering, transparency can give users X-ray vision into the space by adjusting the opacity of various contents within the space through interactive control of the visual transfer function (Kniss et al., 2001).

Hyperdimensional continuous spaces must somehow be reduced to three or fewer dimensions for display. Worlds within Worlds (Beshers and Feiner, 1993) (Figure 5b) displays subspaces of hyperdimensional functions by nesting a 3D coordinate frame within another 3D coordinate frame. The location of the origin (0,0,0) of the inner frame within the outer frame determines the values of the outer frame dimensions used to generate the subspace for the inner frame. By interactively sliding the inner frame around the inside of the outer frame, the full space can be explored. Repeated nesting can enable greater numbers of dimensions. Other methods include hierarchical axes (Mihalisin et al., 1991) and slicing (van Wijk and van Liere, 1993).

## 3.3 Tree and Network Structure

Tree and network structures contain specific *connections* between individual entities. In graph theory terms, a network consists of a set of vertices (entities) connected by a set of edges (connections), which can be either directed or undirected. Like data entities, connections can also contain attributes. Examples include social networks (Hansen et al., 2010), literature citations, or hyperlinks between Web pages. Tree structures are a special subset of networks that are distinct and common enough in digital information to warrant separate treatment. Trees have a hierarchical structure that connects data entities by parent−child connections. To be a tree, each child entity should have only one parent. Examples include computer file directories, menu systems, organization charts, and taxonomies such as the Dewey decimal system. Other useful variants of tree structures exist, such as
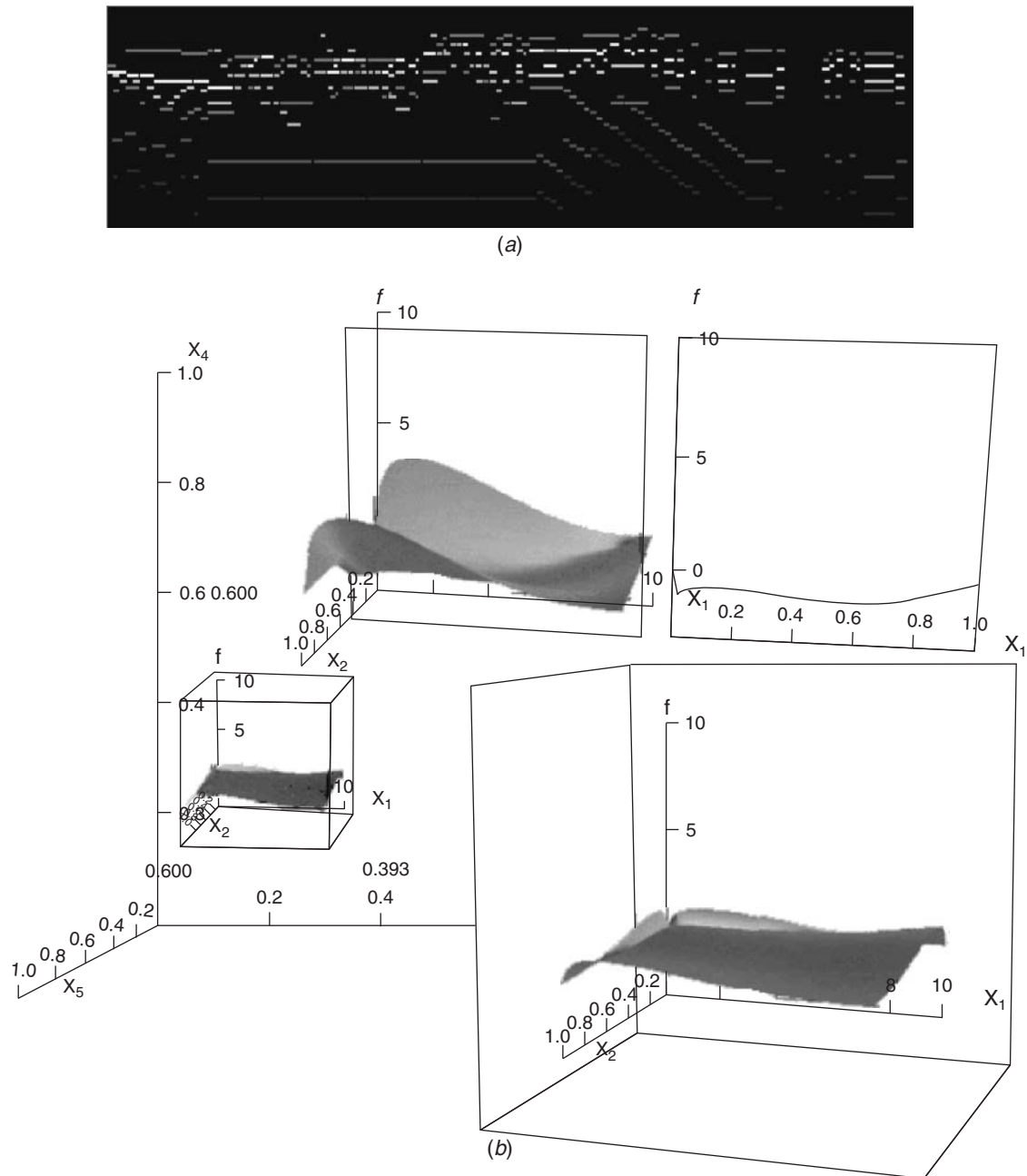
**Figure 5** (*a*) Music Animation Machine showing a portion of Bach's Brandenburg Concerto No. 4, third movement; (*b*) Worlds within Worlds nests inner coordinate frames within outer frames to decompose hyperdimensional spaces. [(*a*) From Malinowski (2004). Courtesy of Stephen Malinowski. (*b*) From Beshers and Feiner (1993). Copyright © 1993 IEEE.]

multitrees (Furnas and Zacks, 1994) and polyarchies (Robertson et al., 2002). New types of insights involve understanding the connection structure, such as the breadth or depth of the tree. The primary challenge for visualization is the spatial layout of the network or tree to reveal the structure of the connections. The

secondary challenge is to visualize data attributes of the entities and connections.

### 3.3.1  Trees

For tree structures, two primary approaches exist for representing parent–child connections visually: *link* and

*containment*. The *link* approach uses node−link diagrams. Entities are mapped to visual nodes, and connections are mapped to visual links between the nodes. Alternative spatial layouts of node−link diagrams include nested−indented, as in Windows Explorer or Mac Finder; top−down or left to right, as in SpaceTree (Grosjean et al., 2002); radial, as in Hyperbolic Tree (Lamping et al., 1995); balloon view (Jeong and Pang, 1998); or 3D ConeTrees, which combines balloon and left−right (Robertson et al., 1993) (Figure 6*a*). These systems emphasize the display of a single data attribute as a text label on the nodes.

Node−link diagrams tend to be space consuming, due to the amount of white space needed within each of these spatial layouts, making it difficult to get beyond 100 or 1000 nodes visible. Since large tree structures cannot be displayed completely on the screen, each layout requires interactive navigation. Navigation should be designed as fluid as possible to reduce tedious operations. The focus + context technique (Section 5.3) is a natural match for tree navigation, enabling users to drill down within an individual branch of focus in the tree while maintaining the context of the path to the root and siblings.

An important design goal is to reveal the size and depth of all tree branches, even when the entire tree cannot be displayed. SpaceTree accomplishes this by displaying collapsed subtrees as triangles, which are size encoded based on the number of nodes in the subtrees. This gives users clues about the overall tree structure. Hyperbolic Tree shrinks the size of nodes near the periphery to pack more nodes on the display. Three-dimensional approaches such as ConeTrees exploit the third dimension for additional space to lay out the tree, but due to occlusion, it is unclear whether that extra virtual space is helpful. The most important factor in three-dimensional designs is the interactive navigation (Wiss and Carr, 1999). Simple six-degree-of-freedom camera movement through a static three-dimensional scene is clearly not effective in these structures. ConeTrees employs a much more efficient interaction technique, cascading rotation of the three-dimensional cones to bring the desired child nodes to the front.

A particularly difficult user task with tree structures is comparing the structure of multiple trees, as in biological taxonomies. Tree Juxtaposer (Munzner et al., 2003) places two trees side by side for comparison using left−to−right layout, synchronized focus+context navigation, and color-coded highlighting of shared branches.

The *containment* approach for tree layout is exemplified by Treemaps (Johnson and Shneiderman, 1991) (Figure 6*b*). Child nodes, represented as rectangles, are contained visually within their parent nodes as in Venn diagrams. Treemaps are space filling, to maximize the use of every available pixel, and scales easily to 10,000 entities. Data attributes are mapped to retinal properties of the node rectangles, such as size and color. Hence, Treemaps emphasizes the visualization of nontextual attributes. In dense Treemaps, not enough space is left for textual node labels. Although clusters of nodes are visible, Treemaps can make it difficult to recognize the structure of the tree.

Nodes can be arranged within their parent node according to a variety of algorithms. The original Treemap used a slice-and-dice algorithm. It was simple but tended to generate rectangles with many different aspect ratios, some square and some long and narrow, which are difficult to visually compare. Newer algorithms generate more perceptually effective squarified Treemaps (Bederson et al., 2002), which attempt to keep node rectangles as close to square as possible (Figure 6*b*). SunBurst (Stasko et al., 2000) offers a radial version of the containment approach based on the stacked pie chart. In comparison to Treemaps, SunBurst can improve learnability for novices but reduces scalability because the number of leaf nodes is limited by one-dimensional circumferential space rather than the full two-dimensional area available to Treemaps.

In general, the link approach is better for gaining insight about the structure of the tree, while the containment approach is better for insights concerning node attributes within the structure and is more scalable to large trees.

### 3.3.2 Networks

For visualizing networks, the *node−link* approach is dominant. Many algorithms have been devised to lay out network diagrams spatially (Herman et al., 2000) and increasingly are tuned to specific types of networks (Figure 7*a*). Designs must consider network features such as number of nodes and links, directedness of links, node degree, any common patterns within the network structure, and attributes of nodes and links that should be visible. In general, the goal is to lay out the network to reveal hidden network patterns and avoid the "hair ball" problem. Graph layout algorithms can seek to optimize aesthetic constraints such as minimizing link crossings, minimizing link lengths, and maximizing symmetries (Purchase et al., 2002; Ware et al., 2002). Links can be drawn as straight lines, arcs, or orthogonal polylines or can be bundled together to reduce link clutter (Holten and van Wijk, 2009).

Some common graph layout algorithms include circular, layering, force directed, and predefined. *Circular* algorithms simply arrange the nodes in a circle and draw the links inside the circle and are useful for small graphs. *Layering* algorithms identify one or more root nodes and lay out all other nodes based on their shortest distance from the root(s) as a series of horizontal or vertical lines or concentric circles. Layering is useful for directed acyclic graphs, because it produces a treelike structure with distinct levels. *Force-directed* algorithms simulate links as physical stretchy springs connecting the nodes. When the simulation settles, it pulls connected nodes near each other, thus producing results similar to the way a person might draw a social network diagram. It also affords natural interaction in which users can directly manipulate node placement while the simulation runs. For this reason, force-directed algorithms are popular but are computationally expensive and lack scalability. *Predefined* layouts exploit node attributes to lay out the graph (Shneiderman and Aris, 2006). For example, SeeNet (Becker et al., 1995) arranges communications nodes according to geographical position and raises links
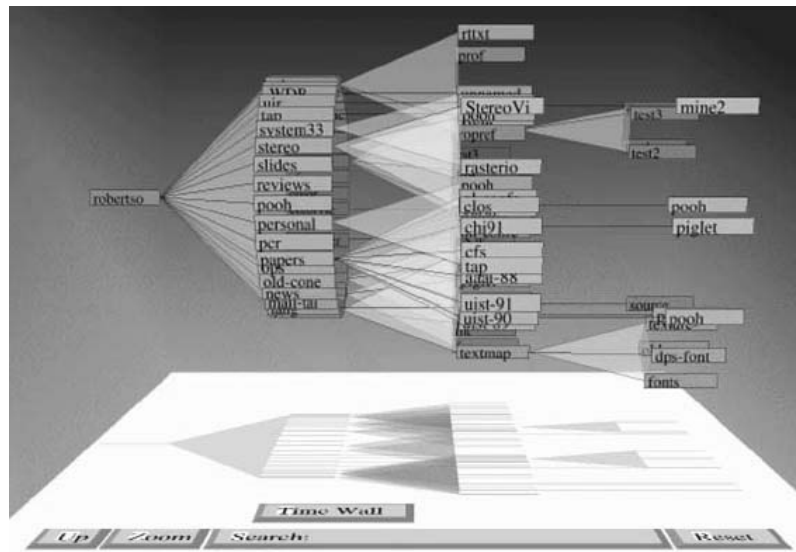
(a)
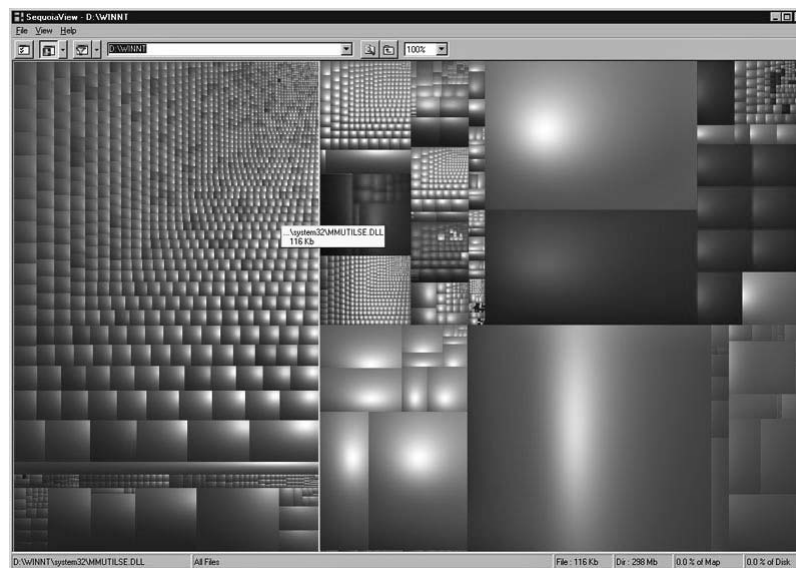


(b)

**Figure 6** (a) ConeTrees; (b) SequoiaView Treemap, showing directory structures on a computer system. ConeTrees emphasizes the tree structure and node labels, whereas Treemap emphasizes node attributes such as file size and type. [(a) With permission from Robertson et al. (1993). Copyright © 1993 ACM, Inc. (b) From van Wijk and van de Wetering (1999). Courtesy of Jarke van Wijk.]

off the surface as three-dimensional arcs. Arc properties such as color and line thickness are used to represent communications type and bandwidth.

To support larger graphs, *hierarchical clustering* algorithms group nodes into clusters based on connectivity. Clusters can then be collapsed into a single visual node to reduce the complexity of large graphs. This technique is often coupled with focus+context navigation to

enable users to explore the contents of the groups (van Ham and van Wijk, 2004). It is also possible to reduce networks to trees using minimum spanning trees or hierarchical aggregation (Feiner, 1988), thereby enabling the use of tree visualization methods. Alternatively, node navigation can be used in conjunction with layering to represent the network from the perspective of one node in focus (Yee et al., 2001).

Burch/Cheswick map of the Internet
showing the major ISPs. Data collected 28 June 1999

(a)



(b)

**Figure 7** (a) Map of the Internet color coded by major Internet service providers (ISPs); (b) NodeTrix visualization of a co-authorship network combines adjacency matrix and node–link representations. [(a) From Cheswick (1998). Courtesy of William Cheswick. (b) From Henry et al. (2007). Courtesy of Nathalie Henry.]

A different approach is to visualize the network as an *adjacency matrix* (Henry et al., 2007) (Figure 7*b*). An adjacency matrix emphasizes the connections instead of the nodes, mapping each potential connection to a cell in the $N \times N$ matrix of all nodes. This approach easily supports very dense graphs that are too cluttered in the node–link approach. It enables insights into individual node connectivity and dense clusters but does not easily support path following and distance insights as the node–link approach does. The NodeTrix system attempts to combine the best of both methods (Henry et al., 2007).

### 3.4 Text and Document Collection Structure

This structure consists of arbitrary collections of documents, often text. Examples include digital libraries, news archives, digital image repositories, and software code. Of the four types of information structure, this type is the least structured and hence can be the most difficult to design visualizations for. Text is particularly challenging to map to visual form, because it is not obvious how text can be the input to a mapping function as described in Section 2.1. Mapping functions must take advantage of the minimal structure and other characteristics of text to generate useful data for computing visual representations. External structures of text or document collections such as tables of contents (tree structure), metadata (tabular structure), or citations (network structure) are categorized as other types of information structures and were discussed earlier. The emphasis of the structure described in this section is on the full text or the documents themselves. Solutions range from the macroscale (overview of large collections) to the microscale (a single document fragment).

A major class of text visualizations focuses on providing semantic maps of large document collections based on document topics. Generally, the goal is to cluster documents spatially in the visualization such that similar documents (documents containing similar content) are near each other and dissimilar documents are distant. This creates a map of the document space based on the metaphor of a physical library in which books are arranged carefully by topic. Similarity between documents can be measured in many ways and is generally the domain of *information retrieval* (Baeza-Yates and Ribiero-Neto, 1999). A common method is to compare the frequency of occurrence of dictionary words or phrases between the two documents. Densities of document clusters can then be analyzed to extract topic keywords for labeling the map. Document Galaxies (Wise et al., 1995) and Kohonen self-organizing maps (Lin, 1992) map individual documents to tiny dots that are clustered by text content. Selecting a dot from the map reveals a document summary or opens the full document. ThemeView (Wise et al., 1995) (Figure 8*a*) emphasizes the documents' topics, creating a three-dimensional terrain landscape representing the themes in the collection. Themes are mapped to terrain landscapes of mountains, with theme strength mapped to mountain height. Mountains that are adjacent or joined indicate the presence of documents that span both themes.

The keyword query approach provides a more focused map based on keywords specified by the user. VIBE (Olsen et al., 1993) visualizes how documents relate to the keywords. It spreads the user's keywords around the periphery of the display. Then, document dots are mapped into the space according to their strength of match to each keyword using a spring-based attraction model. TileBars (Hearst, 1995) inverts the map, showing how the keywords relate to each document. Document hits are listed as in a normal textual search engine, but each document has a tilebar that shows the density of the keywords in each section of the document.

Finally, documents can be arranged by the users themselves or by some default order. Miniature representations of the documents can be displayed to promote browsing by content. Web Book and Forager (Card et al., 1996) collects favorite Web pages in a virtual three-dimensional book that users can flip through quickly and scan visually. Books can be arranged on a virtual bookshelf. With DataMountain (Robertson et al., 1998), users arrange images of favorite Web pages or photos on an inclined plane (Figure 8*b*), taking advantage of spatial memory for recall. At the lowest level of text visualization, SeeSoft (Eick et al., 1992) visualizes the text of software code using a miniaturized representation. It displays each line of text as a tiny colored line segment (more on SeeSoft in Section 4.2).

### 3.5 Combining Multiple Structures

Frequently in real-world applications, information involves complex combinations of multiple information structures. Furthermore, information of one structure type could be computationally massaged into a different structure type to offer new ways to conceptualize the information. For example, an e-commerce website may consist of a text document collection of product pages which also contains a network structure of hyperlinks, is organized by a tree-structured site map, and has tabular metadata about product prices and page accesses. A visualization designer should consider each of these separate structures as a potential visual index into the underlying product information.

A frequent insight goal in these situations is to relate the various structures. However, designing a visual representation that effectively combines multiple structures is difficult. Since a structure typically consumes the primary spatial portion of the visual mapping, combining multiple structures in a single mapping can result in conflict. A primary decision is whether to attempt to combine them or to separate the structures into multiple views (Baldonado et al., 2000). Multiple views simplify the design, since each structure can use its most optimal mapping independently. The structures can then be related by interactive linking between the views (see Section 6.2). Linking is useful for querying one structure with respect to another. However, because interactive linking reveals only a small number of associations at a time, users must mentally integrate the relationships between the structures over time and can easily miss interesting associations. On the other hand, integrating two structures into a single view typically requires that one structure be used as the spatial basis, while the
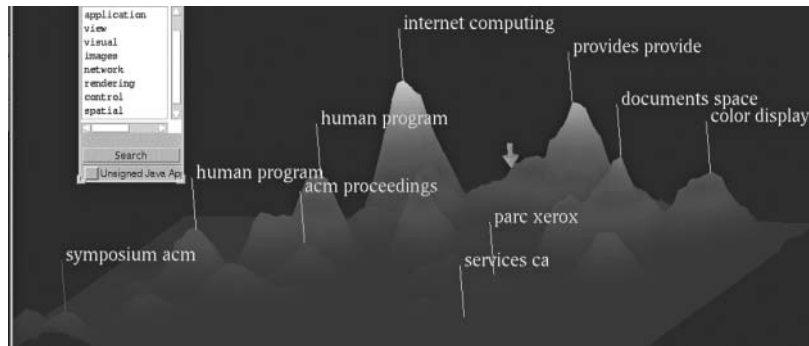
(a)



(b)

**Figure 8** (*a*) ThemeView represents common themes in a large document collection as labeled mountains in a landscape. Probing reveals specific documents. (*b*) DataMountain lets users visually organize Web favorites, documents, or digital photos for later recall. [(*a*) From Wise et al. (1995). Courtesy of Pacific Northwest National Laboratory. (*b*) With permission from Robertson et al. (1998). Copyright © 1998 ACM, Inc.]

other is dismantled and embedded within that space. This enables a clear representation of how the second structure depends on the first, but clarity of the second structure can be lost. The order of nesting of the structures should be designed to match the users' task structure. Another potential solution is animating or morphing between the two structures (Robertson et al., 2002).

An example is PathSim (Polys et al., 2004) (Figure 9), an information-rich virtual environment for biology simulation which combines three-dimensional spatial structure of human anatomy with tabular structure of data collected on viral infection within anatomical components. In this design the tabular data are visually embedded directly within the three-dimensional anatomy as small manipulable visualizations adjacent to their corresponding anatomical components. The tabular

structure is dismantled to associate portions of the data set visually with components in the three-dimensional scene. Although this supports the task of understanding the effects in each anatomical component, it does not enable a single overview of all tabular results. To overcome this, a heads-up display is included according to the multiple-views approach. It shows aggregated tabular information as a summary of what is visible in the entire scene as users navigate in the three-dimensional anatomy.

## 4 OVERVIEW STRATEGIES

Designing methods for the visual representation of very large quantities of information is one of the fundamental problems in visualization research. As information quantity increases, it becomes more difficult to
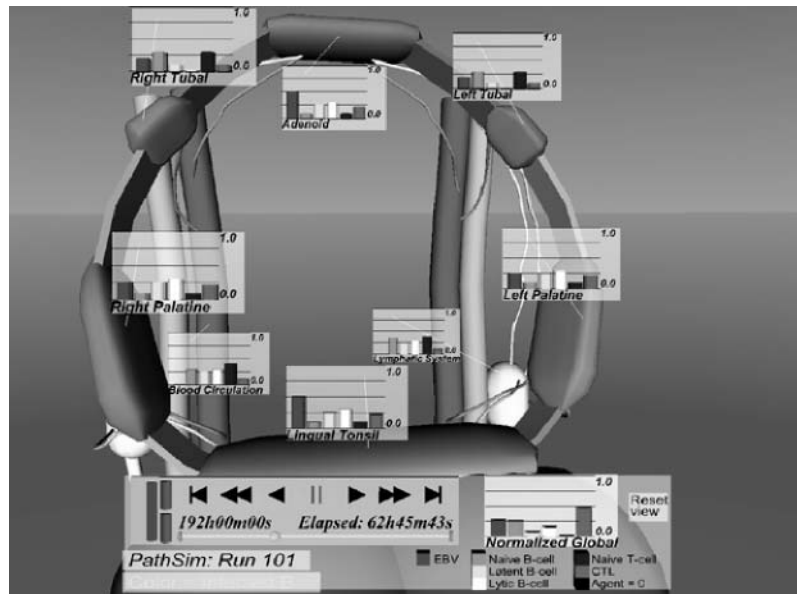
**Figure 9** PathSim reveals simulated viral infection data within the human anatomy, combining tabular structure with three-dimensional spatial structure. Zooming in or out navigates to lower or higher levels of anatomical structure and shows correspondingly lesser or greater levels of aggregation of tabular data. Here, we see the effects of an Epstein–Barr virus infection of the tonsils. (From Polys et al., 2004.)

pack all the information visually on the available screen space. There simply are not enough pixels. Even if there were enough pixels, including all the details in a single display might make it appear visually cluttered. In general, a naive visualization design would be to consume the entire display with the full detail of only a few of the data entities and thereby limit the display to a relatively small portion of the full data set. This is analogous to peering into a vast room through a tiny keyhole and is called the *keyhole problem*. For example, the spreadsheet in Figure 1 shows the detailed numerical data, but the scrolling window reveals only about 40 rows at a time on a typical display.

To support visualization of very large information spaces, Shneiderman suggests the design mantra "*overview first, zoom and filter, then details on demand*" (Shneiderman and Plaisant, 2005). The solution to the keyhole problem is to start users with a broad overview of the full information space, sacrificing information details. Then provide interaction mechanisms that enable users to zoom in on desired information and filter out anything not of interest. Finally, quickly retrieve and display detailed information about individual data entities when selected by the user. There are several advantages to providing an initial visual overview of the information:

- An overview supports the formation of mental models of the information space.
- It reveals what information is present or not present.

- It reveals relationships between the parts of the information, providing broader insights.
- It enables direct access and navigation to parts of the information simply by selecting them from the overview.
- It encourages exploration.

Empirical evidence confirms that the use of visual overviews results in improved user performance in various information-seeking tasks [some studies are listed in Hornbæk et al. (2002)]. In general, visualization designers should seek to pack as much information into the overview as cleanly as possible. A major design decision is choosing which information to percolate up to the overview and which information to bury in the lower detail levels that can only be reached through user interaction. This is somewhat analogous to choosing which products to show in the store window. Ideally, an overview should provide some visual "scent" of all the detailed information hiding beneath it (Pirolli and Card, 1999).

To create overviews that attempt to pack a large data set onto a relatively small screen, there are two possible approaches in the visual mapping process: (1) reducing the quantity of data in the data set before the mapping is applied or (2) reducing the physical size of the visual glyphs created in the mapping.

## 4.1 Reducing Data Quantity

One method for reducing the data quantity while maintaining reasonable representation of the original data is

*aggregation*. Aggregation groups entities within the data set, creating a new data set with fewer total entities. Each aggregate becomes an entity itself, temporarily replacing the need for all entities contained within the aggregate. For example, a histogram applies aggregation to represent data distribution on one attribute (Spence, 2001).

When using aggregation, the first design decision is choosing which entities should be grouped together. Entities can be grouped by common attribute values (Stolte et al., 2002) or by more advanced methods such as clustering algorithms (Yang et al., 2003) or nearest neighbors. The next decision is determining the new attribute values of the aggregates. Ideally, aggregates' values should be representative of the member entities contained. Statistical summaries such as mean, minimum, maximum, and count are commonly used. Aggregation can be iteratively applied to generate tree structures of groups and subgroups (Conklin et al., 2002). The final decision is the visual representation of the aggregates, which, ideally, reveals some hint of their contents.

Aggregate Towers (Rayson, 1999) (Figure 10*a*) groups entities spatially if they overlap on a map. The aggregates are shown as towers whose height represents the number of entities in the aggregate. Zooming out of the map causes further aggregation as needed, and zooming in on the map segregates towers until no towers are needed. XmdvTool (Yang et al., 2003) (Figure 10*b*) clusters tabular data in a parallel coordinates plot. The extent of the contents of each aggregate is revealed by a glowing shadow that emanates from the aggregate's representative polyline.

Aggregation can also be used to group data attributes together. Dimensionality reduction methods reduce the number of data attributes in large multidimensional tabular data sets so that they can be visualized more easily (Rencher, 2002). The reduced set of attributes should approximately capture the main trends found
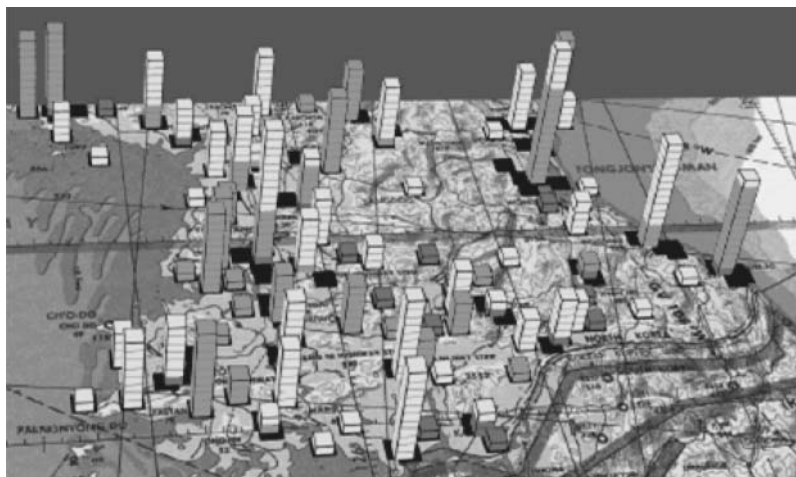


(*a*)
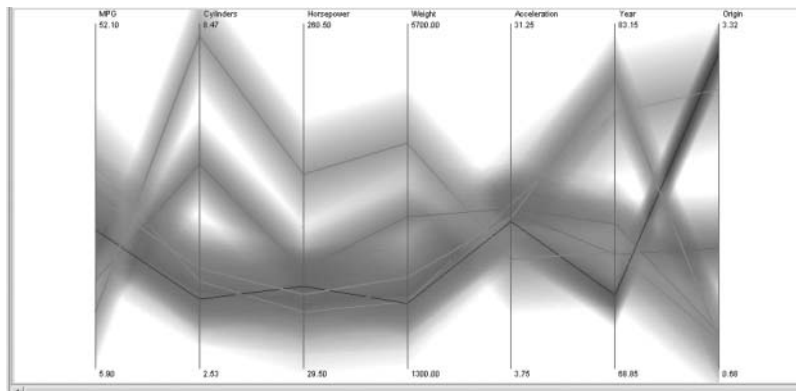


(*b*)

**Figure 10** (*a*) Aggregate Towers stacks military units that spatially overlap on a map. The black footprints show the spatial coverage of each tower. Zooming in segregates the units. (*b*) XmdvTool's Parallel Coordinates clusters the data from Figure 4*b* into six entities to reduce clutter. Translucent shading reveals the approximate spread of each cluster. [(*a*) From Rayson (1999). Copyright © 1999 IEEE. (*b*) From Yang et al., (2003). Courtesy of Matthew Ward.]

in the full set of attributes. For example, *principal-components analysis* projects the data entities onto a subspace of the original data space that best preserves the variance in the data. *Multidimensional scaling* uses measures of similarity between entities, based on their many attribute values, to compute a one-, two-, or three-dimensional map that groups similar entities spatially.

*Filtering* can also be used to reduce data quantity. VIDA (Woodruff et al., 1998) selects a representative subset of data entities based on data density and entity importance. Spotfire (Ahlberg and Wistrand, 1995) relegates less important data attributes to interactive methods such as dynamic queries, eliminating them from the input to the initial visual mapping function. Tree-structured information is easily reduced simply by filtering deeper levels of a tree to visualize the upper levels as an overview.

### 4.2 Miniaturizing Visual Glyphs

Alternatively, emphasis can be placed on miniaturization of the visual glyphs generated by the visual mapping process. Tufte argues for increased data density in visual displays by maximizing the data per unit area of screen space and maximizing the data–ink ratio (Tufte, 2001). A higher data–ink ratio is accomplished by minimizing the quantity of "ink" required for each visual glyph and eliminating chart junk that wastes ink on unimportant nondata elements.

SeeSoft (Eick et al., 1992) (Figure 11*a*) provides an overview of textual software code using miniaturization. Similar to TableLens (Figure 4*a*), each line of code is reduced to a single line segment of colored pixels whose length is proportional to the number of characters in the line of code. In this way, large software projects of up to 50,000 lines of code can be overviewed in a single screen. Color coding can be used to reveal other attributes of the lines of code, such as which programmer wrote it, whether it has been tested, or the amount of CPU time required to execute the line (code profiling). Pixel Bar Charts (Keim et al., 2002) reduces the size of visual glyphs for tabular data to a single pixel, colored by one attribute and ordered on the display by another attribute. The Information Mural (Jerding and Stasko, 1998) (Figure 11*b*) takes miniaturization to the



(*a*)



(*b*)
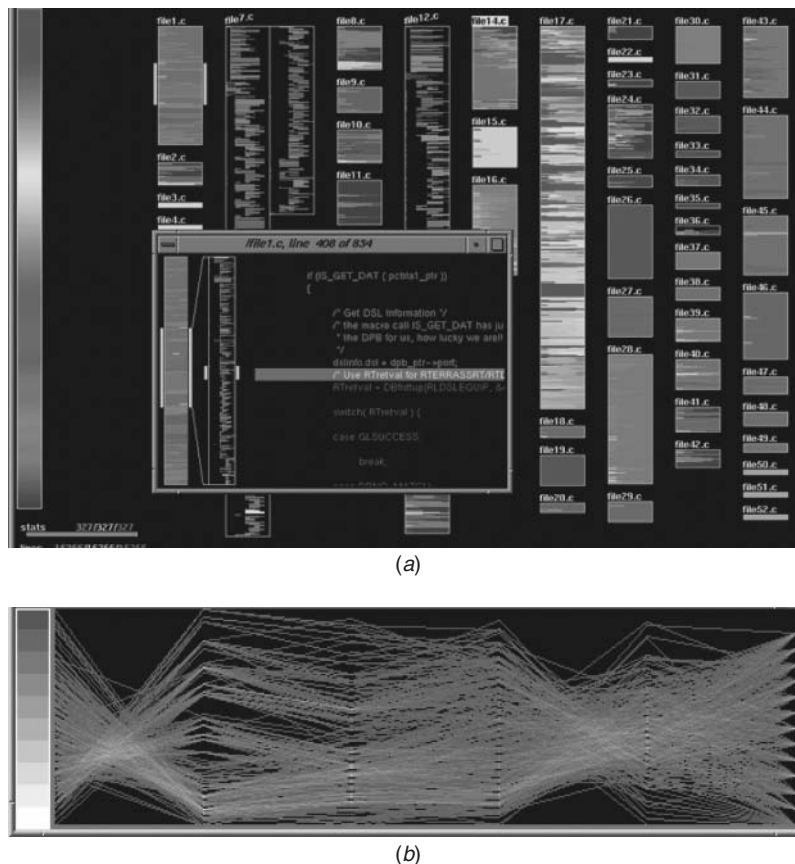
**Figure 11** (*a*) SeeSoft provides a miniaturized visual overview of software code, shading each line of code by an attribute such as date authored; (*b*) Information Mural shows the density of a parallel coordinates plot, enabling users to see hidden patterns that would otherwise be occluded within dense clutter as in Figure 4*b*. [(*a*) From Eick et al. (1992). Courtesy of Stephen Eick. (*b*) From Jerding and Stasko (1998). Courtesy of John Stasko.]

subpixel level. When many glyphs overlap and occlude each other, Mural visualizes the density of the glyphs like an X-ray image.

## 5 NAVIGATION STRATEGIES

After employing an overview strategy to provide a broad view of a large information space, the next design concern is that of navigation. Interactive methods are needed to support navigation between the broad overview and the details of the information. To support this need, three primary navigation design strategies have evolved: *zoom + pan*, *overview + detail*, and *focus + context*. These strategies reside at the third stage of the visualization pipeline, view transformation (see Figure 2).

These navigation strategies should be contrasted with the naive strategy called *detail only*. Detail only is the baseline strategy that does not employ an overview. It provides only the detail-level view of a portion of the information space (e.g., the spreadsheet in Figure 1*a*). Users can navigate by scrolling or panning to access the rest of the information space. In general, the detail-only strategy should be avoided. The principal disadvantage is disorientation due to lack of overview, leaving the user lost in the information space and wondering: Where am I? Where do I want to go? How do I get there?

### 5.1 Zoom + Pan

Zoomable visualizations begin with the overview and then enable users to zoom dynamically into the information space to reach details of interest. Users can zoom back out to return to the overview and zoom in again to another portion of detail. Users can also pan

across the space without zooming out. Zooming can be a smooth continuous navigation through the space as in Pad + + (Bederson et al., 1996) (Figure 12) or can be used to drill down through discrete levels of scale as in Treemaps (Johnson and Shneiderman, 1991). Although the zooming strategy provides an overview, disorientation remains when zooming in. It is easy for users to become lost in the space when zooming in and panning, since the overview is no longer present. This strategy is commonly found in online map browsing systems, since it well-suited to situations where there is not a clearly defined overview level in the information structure.

### 5.2 Overview + Detail

Overview + detail uses multiple views to display an overview and a detail view simultaneously. A field-of-view indicator in the overview indicates the location of the detail view within the information space. The views are bidirectionally linked such that manipulating the field of view in the overview causes the detail view to navigate accordingly. Similarly, when users navigate directly in the detail view, the field of view updates to provide location feedback. This strategy is commonly found in various digital imaging software (Plaisant et al., 1995) such as Photoshop. In SeeSoft (Eick et al., 1992), the miniaturized overview of text operates as a scrollbar for a detailed view of the actual text (Figure 11*a*, center). A zoom factor of 30:1 between overview and detailed view is the usability limit for navigating two-dimensional images, but intermediate views can be chained to reach higher total zoom factors (Plaisant et al., 1995). In navigating three-dimensional worlds, Worlds in Miniature (Stoakley et al., 1995) provides a small three-dimensional overview map attached to a
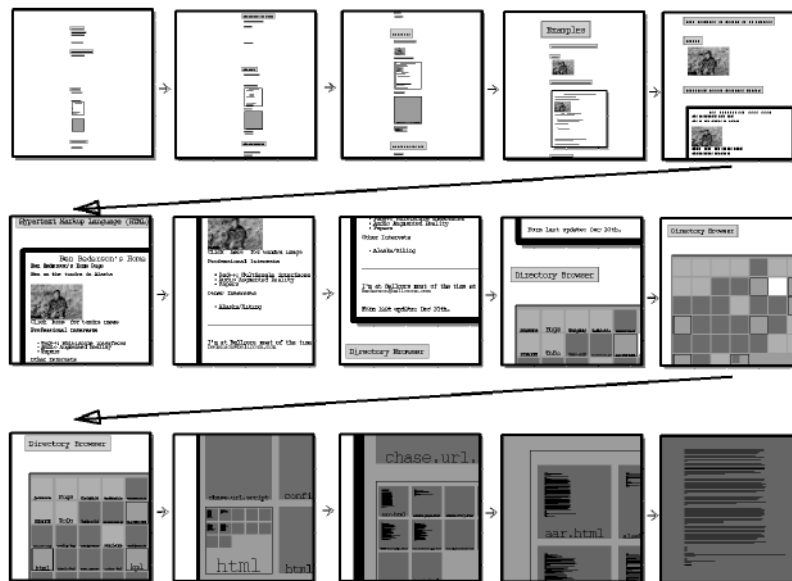


**Figure 12** Zooming sequence in Pad + + from a Web page to an embedded folder to an embedded text file. [From Bederson et al. (1996) Courtesy of Ben Bederson.]
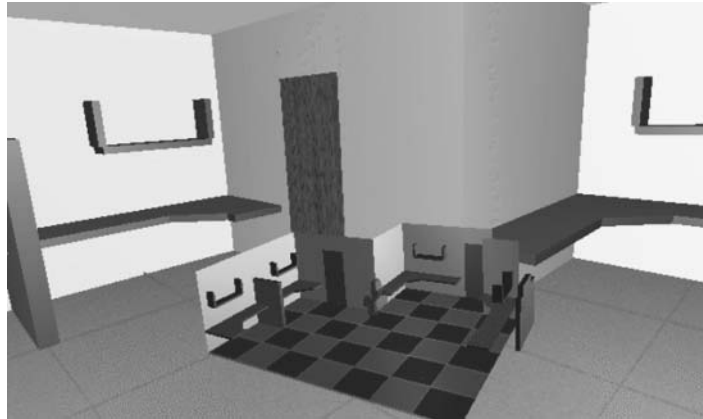
**Figure 13** Worlds in Miniature uses overview + detail for navigating three-dimensional virtual environments. [With permission from Stoakley et al. (1995). Copyright © 1995 ACM.]

virtual glove (Figure 13, bottom center) to help orient users within the world.

Overview + detail preserves overview to avoid disorientation in the detailed view but suffers from a visual discontinuity between the overview and detailed views. Ideally, the detail view should not overlap or occlude the overview, but pop-ups such as tooltips and magnifying glasses are reasonable for small amounts of temporary detailed information. In some cases, it is useful to provide multiple detail views to allow users to compare different entities. For example, a document collection visualization (as in Figure 8) can allow users to open multiple documents simultaneously in different windows. Overview + detail is particularly applicable in scenarios that require maintaining awareness of dynamic events in the overview, as in wargaming.

## 5.3  Focus + Context

Focus + context expands a focus region directly within the overview context. The focus is enlarged and magnified to provide detailed information for that portion of the information space. Users can navigate simply by sliding the focus across the overview to reveal details for other portions of the space. To make room for the expanded focus region, the surrounding overview must be pushed back partially by distorting or warping the overview. For this reason, this strategy is sometimes referred to as fisheye (Furnas, 1986) or distortion-oriented (Leung and Apperley, 1994) techniques. Without distortion, the magnified region would occlude the adjacent context like a magnifying glass. Since the near context is the most important part of the context, the magnifying glass effect is undesirable, and distortion is required to preserve the overview. In general, the focal point is magnified the most, and the degree of magnification decreases with distance from the focal point. Careful design based on a variety of metaphors can help to minimize the negative effects of the distortion.

Several variants of the focus + context strategy have been developed for navigating one- and two-dimensional spaces, including:

- *Bifocal* (Spence, 2001): uses two distinct levels of magnification, such as TableLens (Rao and Card, 1994) (Figure 14*a*)
- *Perspective*: wraps information on three-dimensional angled surfaces, such as Perspective Wall (Robertson et al., 1993) (Figure 14*a*)
- *Wide angle*: creates a classic visual fisheye effect, such as Hyperbolic Tree (Lamping et al., 1995)
- *Nonlinear*: uses more complex magnification functions to create a magnified bubble effect (Keahey and Robertson, 1996) (Figure 14*b*)

As an alternative to spatial distortion, focus + context screens (Baudisch et al., 2002) offer *resolution distortion*, which may provide a better match to the human visual system. Fisheyes have also been developed for navigating three-dimensional spaces (Carpendale et al., 1997). The focus + context strategy offers continuity of detail within overview context but suffers from disorientation caused by dynamic distortion. It is best applied to nonspatial information structures where preservation of spatial distances is not critical.

Although studies have repeatedly shown advantages of these three navigation strategies over the detail-only strategy, comparisons among the three are inconclusive and depend greatly on the specifics of the individual designs, data domains, and user tasks (e.g., Hornbæk et al., 2002). An analytic summary follows:

Zoom + pan:

+ Screen space efficient
+ Infinite scalability
– Lose overview when zooming in
– Slower navigation

Overview + detail:

+ Stable overview
+ Scalable; chained views; multiple overviews or foci

(a)



(b)

**Figure 14** (a) Perspective Wall wraps a one-dimensional time line around a bent wall. The front portion provides detailed information, and perspective sidewalls provide overview context. (b) Nonlinear magnification can create this bubble effect, which magnifies the focus region, squeezes the near context, and maintains an otherwise stable far context. [(a) From Robertson et al. (1993). Courtesy of PARC. (b) From Keahey and Robertson (1996). Copyright © 1996 IEEE.]

– Visual disconnect between views; back-and-forth glancing

– Views compete for screen space; smaller overview

Focus + context:

+ Detail visually connected to surrounding context

– Limited scalability; typically under 10:1 zoom factor

– Distortion; unstable overview

## 6  INTERACTION STRATEGIES

Interaction strategies support further scalability and complexity of visualized information. Although it is preferable to map all data onto the display visually in a form that effectively reveals all desired insights without interaction, this is generally impossible for data of even modest complexity. Interaction strategies overcome this limitation by enabling users to explore additional mappings and insights interactively over time. Many interactive techniques exist (Yi et al., 2007). A few major categories of interaction strategies should be considered in every visualization design.

Interaction occurs at each step in the visualization pipeline (Figure 2). At the data transformation step, users need interactive control for data manipulation and editing. Many data analysts use Excel for its spreadsheet model of data manipulation, enabling them to easily format data and perform computations. For example, NodeXL (Hansen et al., 2010) is a network visualization system implemented within Excel. At the view transform step, navigation is the primary form of interaction as discussed in Section 5. The following interactions primarily occur at the visual mapping step.

### 6.1  Selecting, Grouping, and Extracting

The most fundamental need in visualization is interactive selection of individual data entities or subsets of data entities. Users select entities to identify data that is of interest to them. This is useful for many reasons, including viewing detailed information about the entities (details on demand), highlighting entities that are obscured or occluded in a crowded display, grouping a set of related entities, or extracting entities for future use.

In general, there are two possible criteria by which users can specify selections. First, users can select data entities directly. Direct manipulation visualizations enable users to select entities in a visualization directly using a variety of techniques (Wills, 1996), such as pointing at individual entities' glyphs (as in Figure 4a) or lassoing a group of glyphs. Second, users can select data entities indirectly through selection criteria on information structures (Section 3). For example, Xmd-vTool (Ward, 1994) enables users to make selections in tabular data in parallel coordinates by specifying range criteria on data attributes. In Figure 4b, all American-made cars are highlighted by selecting the U.S. range on the origin axis. Other structure-based selection techniques include selecting an entire branch in a tree structure, selecting a path in a network structure, or selecting a ThemeView mountain (Wise et al., 1995) in a

document collection structure. Another very useful form of indirect selection is search, which enables finding specific entities in a crowded display by their textual content. For example, in the co-authorship network in Figure 7b, users will want to search for specific authors or their own name.

Selection techniques should be designed to enable users to easily select entities, add entities into the current selection, remove entities from the current selection, and clear the selection. Selecting is sometimes called *brushing*, because it is like painting glyphs with a special type of paintbrush that behaves according to the selection technique.

While most selections are ephemeral, there is also the need to preserve some selections. This is useful when users want to define groups of important or interesting entities and possibly extract them from the visualization for future reference or reuse as a future selection. For example, computer network security analysts want to identify suspicious Internet Protocol (IP) addresses in visualizations of network traffic, and biologists want to identify interesting genes in gene expression data. They want to drag the interesting entities out of the visualization into a container where they will be preserved, so that they can continue exploration in the visualization without losing these interesting entities.

### 6.2  Linking

Linking is useful to relate information interactively among multiple views (Baldonado et al., 2000; North et al., 2002). Information can be mapped differently into separate views to reveal different perspectives or different portions of the information. The most common form of linking is called *brushing and linking* (Becker and Cleveland, 1987). Interactive selections of entities in one view are propagated to other views to automatically highlight corresponding entities, enabling users to recognize relationships. This strategy enables users to take advantage simultaneously of the different strengths of different visual representations. This is particularly useful for relating between different information structures (Figure 15), essentially using one structure to query another. Users can select entities according to criteria in one structure, which then shows the distribution of those entities within the other structure. Although linking is commonly used to relate two views of the same data set in a one-to-one fashion, it can also be used to relate entities across many-to-many database relationships for more complex scenarios. Linking also helps users coordinate multiple views during navigation or other interactive operations, such as synchronized scrolling. Tools such as Snap-Together Visualization (North et al., 2002) and Improvise (Weaver, 2004) enable users to mix and match a wide variety of views to produce customized combinations of linked views.

### 6.3  Filtering

Interactive filtering enables users to dynamically reduce information quantity in the display and focus in on information of interest. *Dynamic queries* (Ahlberg and Wistrand, 1995) apply direct manipulation principles to
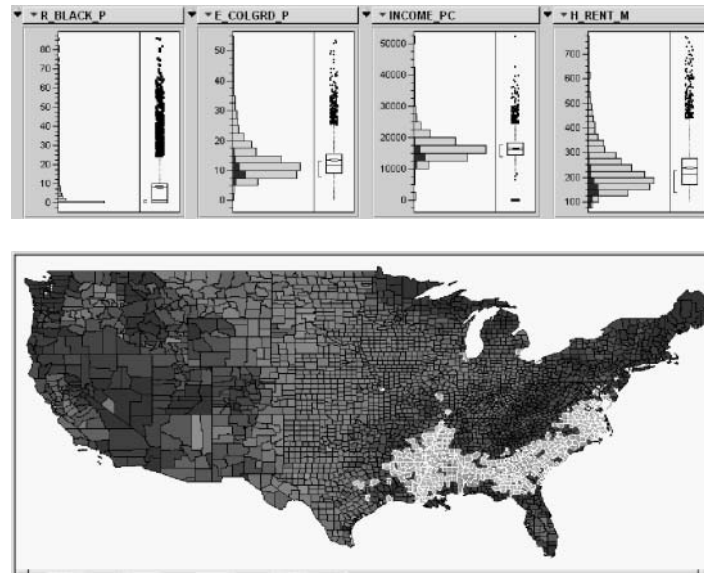
**Figure 15** Interactive brushing and linking between histogram plots (top) and a geographic map (bottom) of a census counties data set. Histograms (tabular information structure) show data distributions for four county attributes: percent population black, percent population with college degree, income per capita, and median rent. The map (two-dimensional spatial information structure) is colored by ''percent farmland.'' Selecting the counties that are more than 25% African American in the first histogram also highlights those counties in the other views. The highlighting in the map reveals that those counties are clustered in the southern and southeastern regions. Selecting counties in the map would similarly highlight them in the histograms. The histograms are generated by JMP (SAS, 2004) and the map by ArcView [Environmental Systems Research Institute (ESRI), 2004]. They are linked by Snap-Together Visualization (North et al., 2002).

querying attribute values. Visual widgets such as the range slider (Figure 1b, right) enable users to adjust query parameters rapidly and view filtered results in the visualization in real time. The widgets also provide a visual representation of the current query parameters. Because of the rapid feedback, dynamic query filters can be used not just to reduce information quantity but also to explore relationships between mapped attributes and query attributes. For example, in Figure 1, filtering with the query slider for "unemployment" to eliminate the low-unemployment counties from the display reveals that counties with high unemployment are all in the low-income and low-education area of the plot. The rapid query feedback also eliminates the difficulty of zero-hit or megahit query results, because users can quickly adjust the query parameters until a desirable number of hits is acquired. For example, by further filtering on "unemployment" in Figure 1, users find that there are 11 counties with an unemployment rate over 20%, most of which are located near the border with Mexico. Dynamic queries are the inverse of brushing; brushing highlights selected data, while dynamic queries elide unselected (filtered) data.

Magic Lenses (Fishkin and Stone, 1995) offers a spatially localized form of filter. For more advanced queries involving complex combinations of Boolean operations, metaphors such as Filter Flow (Young and Shneiderman, 1993) enable users to construct virtual pipelines of filters.

## 6.4 Rearranging and Remapping

Since a single mapping of information to visual form may not be adequate, it is straightforward to enable users to customize the mapping or choose among several mappings. Since the spatial layout is the most salient visual mapping, rearranging the spatial layout of the information is the most potent for generating different insights. For example, TableLens (Rao and Card, 1994) (Figure 4a) can spatially rearrange its view by choosing a different attribute to sort by, and Parallel Coordinates (Inselberg, 1997) (Figure 4b) can rearrange the left-to-right order of its axes. This enables users to explore relationships among attributes. In some visualizations, users may be able to directly manipulate data glyphs to create a custom arrangement manually. For example, some visualizations of network structures allow users to move nodes to new positions within the graph layout so that they can refine the results of the automatically generated layout.

In general, any part of the mapping process throughout the visualization pipeline can be under user control. For example, Spotfire (Ahlberg and Wistrand, 1995) users can customize the scatterplot view (Figure 1b) by choosing data attributes to map to various visual properties, such as $x$, $y$, color, and size. It also provides a variety of visual representations to choose from, including heat maps, parallel coordinates, histograms, and pie and bar charts. Visage (Roth et al., 1996) emphasizes a

technique called *data-centric interaction*, in which users can select data entities directly and drag them to different views to display them in new ways. At the extreme are systems such as Sage and SageBrush (Roth et al., 1994) that let users design new visual mappings for a data set using a set of basic primitives as described in Section 2. Sage can also automatically generate certain visual mappings for a given data set and task using a rule-based expert system.

## 6.5  History Keeping and Story Telling

During the course of extensive interactive exploration, users need to be able to undo interactions, backtrack, return to previous states, reuse common processes, keep bookmarks of important findings, annotate findings, and share results. Tracking a user's history of interaction is an important step to enabling these capabilities. History can be tracked at multiple levels, from low-level tracking of every interactive operation to high-level tracking of key findings. Tableau provides users with a visual history of snapshots of the visualization as they explore (Heer et al., 2008) (Figure 16). Clicking a history snapshot returns the visualization to that state. Snapshots can be taken automatically at every step or

only at major changes or manually whenever the user chooses. Snapshots can be presented as small thumbnail images of the visualization or as a textual script of interactive operations performed. Histories can be linear or branching.

Visualizations and histories can also be annotated by the user to track insights gained and tell stories. Storytelling spaces, such as Occulus's Sandbox (Wright et al., 2006), enable users to drag visualizations and entities into an editable storytelling space, where they can arrange snapshots and relevant information into a visual hypothesis for reporting purposes.

Annotated histories can be shared with others to support collaborative visualization activities. Data-sharing websites such as Swivel and Many Eyes (Viégas et al., 2007) enable users to upload data, choose appropriate visual mappings, and annotate visualizations, all in a public forum where others can participate in a distributed asynchronous social process. Many users can collaboratively build upon each others' exploration to develop much deeper stories.

Because there are many interaction strategies required in a flexible exploratory visualization system, it is critical that careful usability processes are applied in
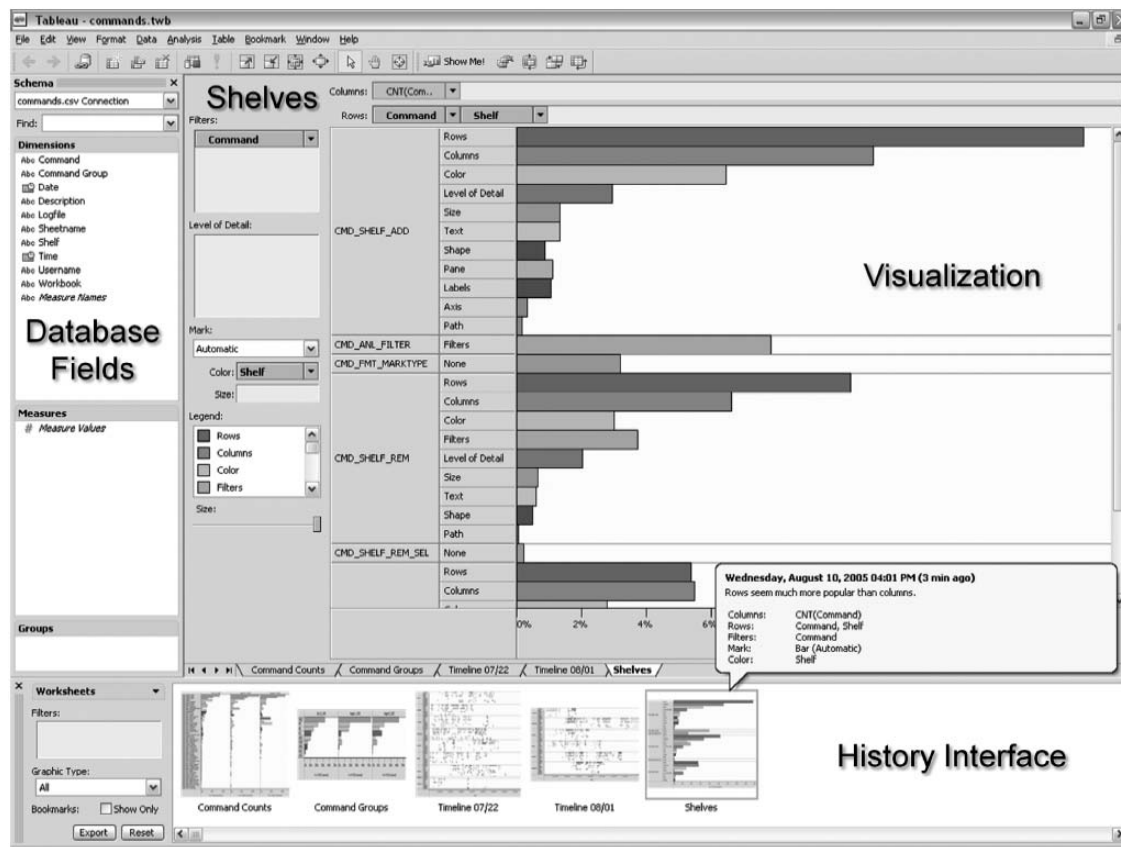


**Figure 16**  Tableau displays a history of snapshots below the main visualization, enabling users to quickly return to any previous state during their exploration. [From Heer et al., (2008). Courtesy of Jeff Heer.]

their design. The many interactions must be designed in an integrated, coherent, and consistent form along with the visual representation. Small usability problems with interaction techniques can significantly reduce the effectiveness of an otherwise well-designed visual representation (Saraiya et al., 2004).

## 7  VISUAL ANALYTICS

The recent rise of the new field of *visual analytics* (Thomas and Cook, 2005) brought forth new emphases in visualization. Visual analytics brings together several fields related to the analysis of data and thus considers the broader context within which visualization resides. Visual analytics seeks to unify the entire analytical process that data analysts encounter and places visualization as the user interface to the process. This has several important implications for the field of visualization.

First, visual analytics emphasizes the role of analysts' cognitive analytical reasoning and sensemaking processes. The sensemaking process model (Pirolli and Card, 2005) (Figure 17) identifies a broad range of analytic activities. Visualization has concentrated on the foraging loop portion of this process. Yet, the sensemaking process highlights the need for new tools to support the synthesis loop as well as the need for a *science of interaction* (Pike et al., 2009) that cleanly integrates all steps of this highly cyclical and fluid process. While past psychological research in visualization has focused primarily on perceptual issues, greater focus is now clearly needed on cognitive issues associated with visualization. Deeper theories are needed to understand how visualization supports analytical reasoning.

Second, visual analytics emphasizes the integration of visualization with computational analytical methods for large-scale data such as data processing, data transformation, data mining, and statistical analysis. Initial work in this area has applied visualization to enable users to control parameters of the computational methods and display the computed results. For example,

iPCA (Jeong et al., 2009) (Figure 18) provides an interactive form of principal-components analysis in which users can manipulate parameters such as the weighting of input dimensions to explore the high-dimensional space and produce insightful projections. New interaction models are needed to enable a deeper mixed-initiative style of interaction between the analyst and the computational methods and that make these complex algorithms usable by novices.

Third, visual analytics emphasizes the production and dissemination of analytical results. Visualization has previously focused on the exploration phase of data analysis, but new methods are needed to transition exploratory findings into presentations—the last step of the sensemaking process (Figure 17). New tools such as Active Reports (Chinchor and Pike, 2009) capture analytical process and provenance into the final report product so that readers can examine the process that led to the findings.

While many of these issues were previously recognized in visualization research, visual analytics has pushed them to the forefront of the research agenda.

## 8  THE FUTURE

Information visualization is a relatively young field (e.g., the IEEE Information Visualization Conference started in 1995). Significant further research is needed on new visual mappings, overview strategies, interaction and navigation strategies, evaluation methods, underlying theories, and guidelines. Among many grand challenges in visualization, a few critical areas of need that should be explored in the foreseeable future include:

1.  *Visualization of Massive Heterogeneous Data*. Applications in intelligence analysis and homeland security require new abilities to analyze terabytes of textual, voice, and video data in unstructured collections (Thomas and Cook, 2005). Bioinformatics is driving the need for
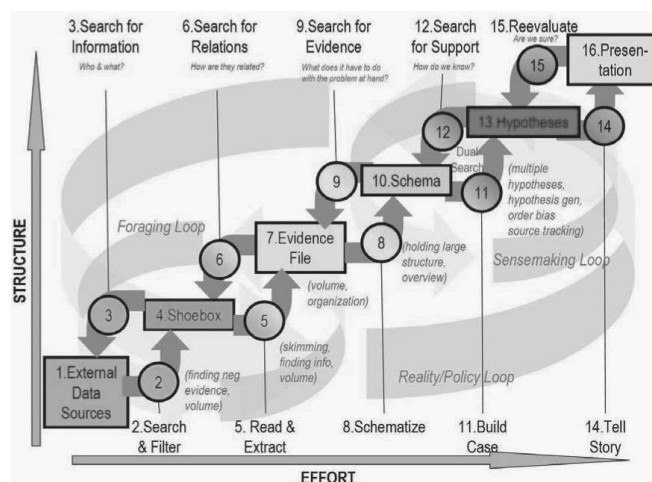


**Figure 17**  The sensemaking process for analysts. (Adapted from Pirolli and Card, 2005.)
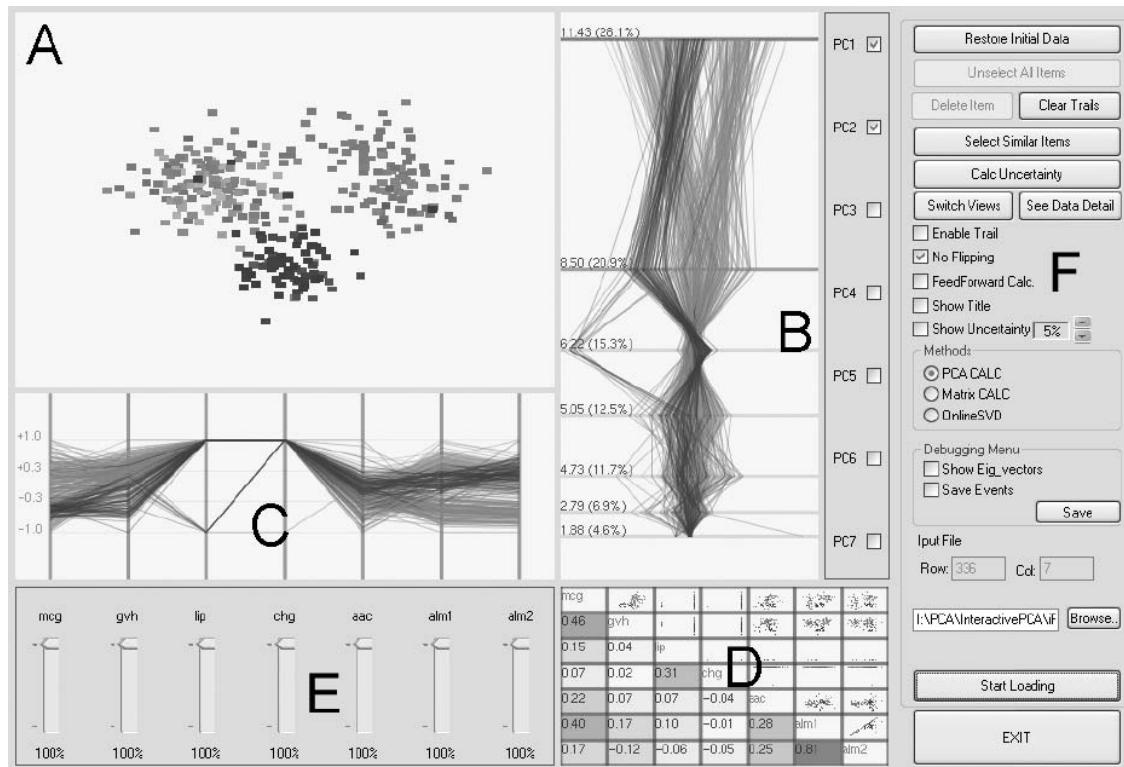
**Figure 18**    iPCA interactive principal-components analysis. [From Jeong et al. (2009). Courtesy of D. H. Jeong.]

new methods to visualize megadimensional tabular data sets, containing thousands or millions of data attributes and huge networks.

2. *Integrating Visualization with the Broader Analytic Context*. Visualization is not an independent task but must be integrated with data management, information retrieval, statistical analysis, data mining (Shneiderman, 2002), decision support, task management, and content authoring and publishing in support of visual analytics.

3. *Visualization with Novel Display and Interaction Devices*. Large high-resolution display technologies, multitouch tabletops, and mobile devices can fundamentally impact interactive visualization (Ni et al., 2006). New visualization strategies must be devised to expand the limits of visualization and exploit high-bandwidth interaction (Ball and North, 2007).

4. *Visualization Evaluation*. To support the iterative design of increasingly advanced visualizations, new evaluation methods are needed to identify and measure the long-term effect of visualizations on high-level insight generation and information analysis (North, 2006).

As researchers explore future visualization innovations and practitioners apply visualization design principles to new domains, the proliferation of effective information visualizations will lead to widespread improvements in the usability of information and to increased generation of valuable insight.

## REFERENCES

Ahlberg, C., and Wistrand, E. (1995), "IVEE: An Information Visualization and Exploration Environment," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 66–73.

Amar, R., Eagan, J., and Stasko, J. (2005), "Low-Level Components of Analytic Activity in Information Visualization," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 111–117.

Baeza-Yates, R., and Ribiero-Neto, B. (1999), *Modern Information Retrieval*, Addison-Wesley, Reading, MA.

Baldonado, M., Woodruff, A., and Kuchinsky, A. (2000), "Guidelines for Using Multiple Views in Information Visualization," in *Proceedings of the ACM Conference on Advanced Visual Interfaces*, ACM, New York, NY, pp. 110–119.

Ball, R., and North, C. (2007), "Visual Analytics: Realizing embodied interaction for visual analytics through large

displays," *Computers and Graphics,* Vol. 31, No. 3, pp. 380–400.

Baudisch, P., Good, N., Bellotti, V., and Schraedley, P. (2002), "Keeping Things in Context: A Comparative Evaluation of Focus Plus Context Screens, Overviews, and Zooming," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 259–266.

Becker, R., and Cleveland, W. (1987), "Brushing Scatterplots," *Technometrics*, Vol. 29, No. 2, pp. 127–142.

Becker, R. A., Eick, S. G., and Wilks, A. R. (1995), "Visualizing Network Data," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 1, No. 1, pp. 16–28.

Bederson, B., Hollan, J., Perlin, K., Meyer, J., Bacon, D., and Furnas, G. (1996), "Pad + +: A Zoomable Graphical Sketchpad for Exploring Alternate Interface Physics," *Journal of Visual Languages and Computing*, Vol. 7, No. 1, pp. 3–32.

Bederson, B., Shneiderman, B., and Wattenberg, M. (2002, October), "Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies," *ACM Transactions on Graphics*, Vol. 21, No. 4, pp. 833–854.

Bertin, J. B. (1983), *Semiology of Graphics: Diagrams, Networks, Maps*, translated by W. J. Berg, University of Wisconsin Press, Madison, WI.

Beshers, C., and Feiner, S. (1993), "AutoVisual: Rule-Based Design of Interactive Multivariate Visualizations," *IEEE Computer Graphics and Applications*, Vol. 13, No. 4, pp. 41–49.

Card, S., Robertson, G., and York, W. (1996), "The Web-Book and the Web Forager: An Information Workspace for the World-Wide Web," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, p. 111.

Card, S., Mackinlay, J., and Shneiderman, B. (1999), *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann, San Francisco, CA.

Carlis, J., and Konstan, J. (1998), "Interactive Visualization of Serial Periodic Data," in *Proceedings of the ACM Symposium on User Interface Software and Technology*, ACM, New York, NY, pp. 29–38.

Carpendale, M. S. T., Cowperthwaite, D. J., and Fracchia, F. D. (1997), "Extending Distortion Viewing from 2D to 3D," *IEEE Computer Graphics and Applications*, Vol. 17, No. 4, pp. 42–51.

Carswell, C. (1992), "Reading Graphs: Interactions of Processing Requirements and Stimulus Structure," in *Precepts, Concepts, and Categories: The Representation and Processing of Information*, B. Burns, Ed., Elsevier Science, Amsterdam, pp. 605–645.

Chambers, J., Cleveland, W., Kleiner, B., and Tukey, P. (1983), *Graphical Methods for Data Analysis*, Wadsworth, Belmont, CA.

Chen, C. (1999), *Information Visualisation and Virtual Environments*, Springer-Verlag, London.

Chen, C., and Yu, Y. (2000), "Empirical Studies of Information Visualization: A Meta-Analysis," *International Journal of Human–Computer Studies*, Vol. 53, No. 5, pp. 851–866.

Chernoff, H. (1973), "The Use of Faces to Represent Points in $k$-Dimensional Space Graphically," *Journal of the American Statistical Association*, Vol. 68, pp. 361–368.

Cheswick, B. (1998), "The Scenic Route," *Wired*, Vol. 6, No. 12; also, Internet Mapping Project, available: http://research.lumeta.com/ches/map/.

Chewar, C. M., McCrickard, D. S., Ndiwalana, A., North, C., Pryor, J., and Tessendorf, D. (2002), "Secondary Task Display Attributes: Optimizing Visualizations for Cognitive Task Suitability and Interference Avoidance," in *Proceedings of the Symposium on Data Visualisation*, Eurographics Association, Aire-la-Ville, Switzerland, pp. 165–171.

Chinchor, N., and Pike, W. (2009), "The Science of Analytic Reporting," *Information Visualization,* Vol. 8, No. 4, pp. 286–293.

Cleveland, W. (1993), *Visualizing Data*, Hobart, Summit, NJ.

Cleveland, W. S., and McGill, R. (1984), "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *Journal of the American Statistical Association*, Vol. 79, No. 387, pp. 531–554.

Conklin, N., Prabhakar, S., and North, C. (2002), "Multiple Foci Drill-Down through Tuple and Attribute Aggregation Polyarchies in Tabular Data," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 131–134.

Eick, S. G., Steffen, J. L., and Sumner, E. E., Jr. (1992), "SeeSoft: A Tool for Visualizing Line Oriented Software Statistics," *IEEE Transactions on Software Engineering*, Vol. 18, No. 11, pp. 957–968.

Elliott, E., and Davenport, G. (1994), "Video Streamer," in *Companion Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 65–68.

Environmental Systems Research Institute (ESRI) (2004), *ArcView Desktop GIS*, ESRI, Redlands, CA, available: http://www.esri.com/.

Fayyad, U. M., Grinstein, G., Wierse, A., and Fayyad, U. (2001), *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann, San Francisco, CA.

Feiner, S. (1988), "Seeing the Forest for the Trees: Hierarchical Displays of Hypertext Structures," in *Proceedings of the ACM Conference on Office Information Systems*, ACM, New York, NY, pp. 205–212.

Fekete, J. D., van Wijk, J., Stasko, J., and North, C. (2008), "The Value of Information Visualization", in *Information Visualization: Human-Centered Issues and Perspectives*, A. Kerren, J. Stasko, J. D. Fekete, and C. North (Eds.). Lecture Notes In Computer Science, Vol. 4950. Springer-Verlag, Berlin, pp. 1–18.

Fishkin, K., and Stone, M. (1995), "Enhanced Dynamic Queries via Movable Filters," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 415–420.

Furnas, G. (1986), "Generalized Fisheye Views," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 16–23.

Furnas, G., and Zacks, J. (1994), "Multitrees: Enriching and Reusing Hierarchical Structure," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 330–336.

Gillan, D., Wickens, C., Hollands, J., and Carswell, C. (1998), "Guidelines for Presenting Quantitative Data in HFES Publications," *Human Factors*, Vol. 36, pp. 419–440.

Grosjean, J., Plaisant, C., and Bederson, B. (2002), "SpaceTree: Supporting Exploration in Large Node Link Tree, Design Evolution and Empirical Evaluation," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 57–64.

Hansen, D., Shneiderman, B., and Smith, M. (2010), *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*, Morgan Kaufmann Publishers Inc., San Francisco, CA.

Harrower, M., and Brewer, C. (2003), "ColorBrewer.org: An Online Tool for Selecting Color Schemes for Maps," *Cartographic Journal,* Vol. 40, No. 1, pp. 27−37.

Healey, C. G., Booth, K. S., and Enns, J. T. (1996), "High-Speed Visual Estimation Using Preattentive Processing," *ACM Transactions on Human−Computer Interaction*, Vol. 3, No. 2, pp. 107−135.

Hearst, M. (1995), "TileBars: Visualization of Term Distribution Information in Full Text Information Access," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 59−66.

Heer, J., Mackinlay, J., Stolte, C., and Agrawala, M. (2008), "Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation," *IEEE Transactions on Visualization and Computer Graphics,* Vol. 14, No. 6, pp. 1189−1196.

Henry, N., Fekete, J. D., and McGuffin, M. (2007), "NodeTrix: A Hybrid Visualization of Social Networks," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 13, No. 6, pp. 1302−1309.

Herman, I., Melancon, G., and Marshall, M. S. (2000), "Graph Visualization and Navigation in Information Visualization: A Survey," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 6, No. 1, pp. 24−43.

Hetzler, B., Harris, W. M., Havre, S., and Whitney, P. (1998), "Visualizing the Full Spectrum of Document Relationships," in *Structures and Relations in Knowledge Organization: Proceedings of the 5th International ISKO Conference*, Advances in knowledge organization, no. 6. Mustafa El Hadi, W., Maniez, J., & Pollitt, A. S. (Eds.). Ergon Verlag, Würzburg, Germany, pp. 168−175.

Holten, D., and van Wijk, J. (2009), "Force-Directed Edge Bundling for Graph Visualization," *Computer Graphics Forum*, Vol. 28, No. 3, pp. 983−990.

Hornbæk, K., Bederson, B., and Plaisant, C. (2002), "Navigation Patterns and Usability of Zoomable User Interfaces with and without an Overview," *ACM Transactions on Computer−Human Interaction*, Vol. 9, No. 4, pp. 362−389.

Inselberg, A. (1997), "Multidimensional Detective," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 100−107.

Jeong, C., and Pang, A. (1998), "Reconfigurable Disc Trees for Visualizing Large Hierarchical Information Space," in *Proceedings of the IEEE Symposium on Information Visualization,* IEEE Computer Society, Washington, DC, pp. 19−25.

Jeong, D. H., Ziemkiewicz, C., Fisher, B., Ribarsky, W., and Chang, R. (2009), "iPCA: An Interactive System for PCA-based Visual Analytics," *Computer Graphics Forum*, Vol. 28, No. 3, pp. 767−774.

Jerding, D. F., and Stasko, J. T. (1998), "The Information Mural: A Technique for Displaying and Navigating Large Information Spaces," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 4, No. 3, pp. 257−271.

Johnson, B., and Shneiderman, B. (1991), "Treemaps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures," in *Proceedings of the IEEE Conference on Visualization*, IEEE Computer Society, Washington, DC, pp. 284−291.

Kandogan, E. (2000), "Star Coordinates: A Multi-Dimensional Visualization Technique with Uniform Treatment of Dimensions," in *Extended Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 9−12.

Keahey, T. A., and Robertson, E. (1996), "Techniques for Non-Linear Magnification Transformations," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 38−45.

Keim, D., Hao, M., Dayal, U., and Hsu, M. (2002), "Pixel Bar Charts: A Visualization Technique for Very Large Multi-Attribute Data Sets," *Information Visualization*, Vol. 1, No. 1, pp. 20−34.

Kniss, J., Kindlmann, G., and Hansen, C. (2001), "Interactive Volume Rendering Using Multidimensional Transfer Functions and Direct Manipulation Widgets," in *Proceedings of the IEEE Conference on Visualization*, IEEE Computer Society, Washington, DC, pp. 255−262.

Lamping, J., Rao, R., and Pirolli, P. (1995), "A Focus + Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 401−408.

Leung, Y. K., and Apperley, M. D. (1994), "A Review and Taxonomy of Distortion-Oriented Presentation Techniques," *ACM Transactions on Computer−Human Interaction*, Vol. 1, No. 2, pp. 126−160.

Lin, X. (1992), "Visualization for the Document Space," in *Proceedings of the IEEE Conference on Visualization*, IEEE Computer Society, Washington, DC, pp. 274−281.

Liu, Z., Nersessian, N., and Stasko, J. (2008), "Distributed Cognition as a Theoretical Framework for Information Visualization," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 14, No. 6, pp. 1173−1180.

Mackinlay, J. (1986), "Automating the Design of Graphical Presentations of Relational Information," *ACM Transactions on Graphics*, Vol. 5, No. 2, pp. 110−141.

Malinowski, S. (2004), "Music Animation Machine," available: http://www.musanim.com/.

Mihalisin, T., Timlin, J., and Schwegler, J. (1991), "Visualizing Multivariate Functions, Data, and Distributions," *IEEE Computer Graphics and Applications*, Vol. 11, No. 3, pp. 28−35.

Miller, J. (2004), "Daisy Analytics," available: http://www.daisy.co.uk/index.html.

Munzner, T., Guimbretiere, F., Tasiran, S., Zhang, L., and Zhou, Y. (2003), "TreeJuxtaposer: Scalable Tree Comparison Using Focus+Context with Guaranteed Visibility," *ACM Transactions on Graphics,* Vol. 22, No. 3, pp. 453−462.

Ni, T., Schmidt, G., Staadt, O., Livingston, M., Ball, R., and May, R. (2006), "A Survey of Large High-Resolution Display Technologies, Techniques, and Applications," in *Proceedings of the IEEE Conference on Virtual Reality,* IEEE Computer Society, Washington, DC, USA, 223−236.

North, C. (2006), "Visualization Viewpoints: Toward Measuring Visualization Insight," *IEEE Computer Graphics & Applications*, Vol. 26, No. 3, pp. 6−9.

North, C., Shneiderman, B., and Plaisant, C. (1996), "User Controlled Overviews of an Image Library: A Case Study of the Visible Human," in *Proceedings of the ACM Digital Libraries Conference*, ACM, New York, NY, pp. 74−82.

North, C., Conklin, N., and Saini, V. (2002), "Visualization Schemas for Flexible Information Visualization," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 15–22.

Nowell, L., Schulman, R., and Hix, D. (2002), "Graphical Encoding for Information Visualization: An Empirical Study," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 43–50.

Olsen, K., Korfhage, R., Sochats, K., Spring, M., and Williams, J. (1993), "Visualization of a document collection: The vibe system," *Information Processing and Management*, Vol. 29, No. 1, pp. 69–81.

Pike, W., Stasko, J., Chang, R., and O'Connell, T. (2009), "The Science of Interaction," *Information Visualization*, Vol. 8, No. 4, pp. 263–274.

Pirolli, P., and Card, S. (1999), "Information Foraging," *Psychology Review*, Vol. 106, No. 4, pp. 643–675.

Pirolli, P., and Card, S. (2005), "The Sensemaking Process and Leverage Points for Analyst Technology as Identified through Cognitive Task Analysis," in *Proceedings of International Conference on Intelligence Analysis 2005, MITRE Corporation, McLean, VA*, available: https://analysis.mitre.org/.

Plaisant, C. (2004), "The Challenge of Information Visualization Evaluation," in *Proceedings of the ACM Conference on Advanced Visual Interfaces*, ACM, New York, NY, pp. 109–116.

Plaisant, C., Carr, D., and Shneiderman, B. (1995), "Image Browsers: Taxonomy, Guidelines, and Informal Specifications," *IEEE Software*, Vol. 12, No. 2, pp. 21–32.

Plaisant, C., Milash, B., Rose, A., Widoff, S., and Shneiderman, B. (1996), "LifeLines: Visualizing Personal Histories," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 221–227.

Polys, N., Bowman, D., North, C., Laubenbacher, R., and Duca, K. (2004), "PathSim Visualizer: An Information-Rich Virtual Environment Framework for Systems Biology," in *Proceedings of the ACM Web3D Symposium*, ACM, New York, NY, pp. 7–14.

Purchase, H., Carrington, D., and Allder, J. (2002), "Empirical Evaluation of Aesthetics-Based Graph Layout," *Empirical Software Engineering*, Vol. 7, pp. 233–255.

Rao, R., and Card, S. (1994), "The TableLens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 318–322.

Rayson, R. (1999), "Aggregate Towers: Scale Sensitive Visualization and Decluttering of Geospatial Data," in *Proceedings of the IEEE Symposium on Information Visualization*, ACM, New York, NY, pp. 92–99.

Rencher, A. (2002), *Methods of Multivariate Analysis*, 2nd ed., Wiley, New York.

Robertson, G., Card, S., and Mackinlay, J. (1993), "Information Visualization Using 3D Interactive Animation," *Communications of the ACM*, Vol. 36, No. 4, pp. 57–71.

Robertson, G., Czerwinski, M. Larson, K., Robbins, D., Thiel, D., and van Dantzich, M. (1998), "Data Mountain: Using Spatial Memory for Document Management," in *Proceedings of the ACM Symposium on User Interface Software and Technology*, ACM, New York, NY, pp. 153–162.

Robertson, G., Cameron, K., Czerwinski, M., and Robbins, R. (2002), "Polyarchy Visualization: Visualizing Multiple Intersecting Hierarchies," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 423–430.

Rosenblum, L., Earnshaw, R., Encarnacao, J., Hagen, H., Kaufman, A., Klimenko, S., Nielson G., Post F., and Thalmann, D. (1994), *Scientific Visualization: Advances and Challenges*, Academic Press, San Diego, CA, in association with IEEE Computer Society, Los Alamitos, CA.

Rosson, M. B., and Carroll, J. (2001), *Usability Engineering: Scenario-Based Development of Human–Computer Interaction*, Morgan Kaufmann, San Francisco, CA.

Roth, S. F., Kolojejchick, J., Mattis, J., and Goldstein, J. (1994), "Interactive Graphic Design Using Automatic Presentation Knowledge," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 112–117.

Roth, S. F., Lucas, P., Senn, J. A., Gomberg, C. C., Burks, M. B., Stroffolino, P. J., Kolojejchick, J. A., and Dunmire, C. (1996), "Visage: A User Interface Environment for Exploring Information," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 3–12.

Saraiya, P., North, C., and Duca, K., (2004), "An Evaluation of Microarray Visualization Tools for Biological Insight," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 1–8.

SAS (2004), *JMP: The Statistical Discovery Software*, SAS Institute, Cary, NC, available: http://www.jmp.com/.

Shneiderman, B. (1996), "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization," in *Proceedings of the IEEE Symposium on Visual Languages*, IEEE Computer Society, Washington, DC, pp. 336–343.

Shneiderman, B. (2002), "Inventing Discovery Tools: Combining Information Visualization with Data Mining," *Information Visualization*, Vol. 1, No. 1, pp. 5–12.

Shneiderman, B., and Aris, A. (2006), "Network Visualization by Semantic Substrates," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 12, No. 5, pp. 733–740.

Shneiderman, B., and Plaisant, C. (2005), *Designing the User Interface: Strategies for Effective Human–Computer Interaction*, 4th ed., Addison-Wesley, Reading, MA.

Spence, R. (2001), *Information Visualization*, Addison-Wesley, Reading, MA.

Stasko, J., Catrambone, R., Guzdial, M., and Mcdonald, K. (2000), "An Evaluation of Space-Filling Information Visualizations for Depicting Hierarchical Structures," *International Journal of Human–Computer Studies*, Vol. 53, No. 5, pp. 663–694.

Stoakley, R., Conway, M. J., and Pausch, R. (1995), "Virtual Reality on a WIM: Interactive Worlds in Miniature," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 265–272.

Stolte, C., Tang, D., and Hanrahan, P. (2002), "Polaris: A System for Query, Analysis and Visualization of Multi-Dimensional Relational Databases," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 8, No. 1, pp. 52–65.

Tableau Software (2010), available: http://www.tableausoftware.com/.

Thomas, J., and Cook, K. (2005), *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, available: http://nvac.pnl.gov/agenda.stm.

Tory, M., and Möller, T. (2004a), "Human Factors in Visualization Research," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 10, No. 1, pp. 72−84.

Tory, M., and Möller, T. (2004b), "Rethinking Visualization: A High-Level Taxonomy," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 151−158.

Tufte, E. (2001), *The Visual Display of Quantitative Information*, 2nd ed., Graphics Press, Cheshire, CT.

van Ham, F., and van Wijk, J. (2004), "Interactive Visualization of Small World Graphs," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 199−206.

van Wijk, J., and van de Wetering, H. (1999), "Cushion Treemaps: Visualization of Hierarchical Information," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 73−80, available: http://www.win.tue.nl/sequoiaview/.

van Wijk, J., and van Liere, R. (1993), "HyperSlice: Visualization of Scalar Functions of Many Variables," in *Proceedings of the IEEE Visualization Conference*, IEEE Computer Society, Washington, DC, pp. 119−125.

Viégas, F., Wattenberg, M., van Ham, F., Kriss, J., and McKeon, M. (2007), "Many Eyes: A Site for Visualization at Internet Scale," *IEEE Transactions on Visualization and Computer Graphics,* Vol. 13, No. 6, pp. 1121−1128.

Ward, M. (1994), "XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data," in *Proceedings of the IEEE Conference on Visualization*, IEEE Computer Society, Washington, DC, pp. 326−333, available: http://davis.wpi.edu/~xmdv/.

Ward, M. (2002), "A Taxonomy of Glyph Placement Strategies for Multidimensional Data Visualization," *Information Visualization*, Vol. 1, No. 3, pp. 194−210.

Ware, C. (2004), *Information Visualization: Perception for Design*, Morgan Kaufmann, San Francisco, CA.

Ware, C., Purchase, H., Colpoys, L., and McGill, M. (2002), "Cognitive Measurements of Graph Aesthetics," *Information Visualization*, Vol. 1, No. 2, pp. 103−110.

Weaver, C. (2004), "Building Highly-Coordinated Visualizations in Improvise," in *Proceedings of the IEEE Symposium on Information Visualization 2004*, IEEE Computer Society, Washington, DC, pp. 159−166.

Wehrend, S., and Lewis, C. (1990), "A Problem-Oriented Classification of Visualization Techniques," in *Proceedings of the IEEE Conference on Visualization*, IEEE Computer Society, Washington, DC, pp. 139−143.

Wickens, C., and Hollands, J. (2000), *Engineering Psychology and Human Performance*, Prentice Hall, Upper Saddle River, NJ.

Wilkinson, L. (1999), *The Grammar of Graphics*, Springer-Verlag, New York.

Wills, G. (1996), "Selection: 524,288 Ways to Say 'This Is Interesting,'" in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 54−60.

Wilson, M. (2002), "Six Views of Embodied Cognition," *Psychonomic Bulletin & Review*, Vol. 9, pp. 625−636.

Wise, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., and Crow, V. (1995), "Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 51−58.

Wiss, U., and Carr, D. A. (1999), "An Empirical Study of Task Support in 3D Information Visualizations," in *Proceedings of the IEEE International Conference on Information Visualization*, London, IEEE Computer Society, Washington, DC, pp. 392−399.

Woodruff, A., Landay, J., and Stonebraker, M. (1998), "Constant Information Density in Zoomable Interfaces," in *Proceedings of the Conference on Advanced Visual Interfaces*, ACM, New York, NY, pp. 57−65.

Wright, W., Schroh, D., Proulx, P., Skaburskis, A., and Cort, B. (2006), "The Sandbox for Analysis: Concepts and Methods," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 801−810.

Yang, J., Ward, M., and Rundensteiner, E. (2003), "Interactive Hierarchical Displays: A General Framework for Visualization and Exploration of Large Multivariate Data Sets," *Computers and Graphics*, Vol. 27, No. 2, pp. 265−283.

Yee, K. P., Fisher, D., Dhamija, R., and Hearst, M. (2001), "Animated Exploration of Dynamic Graphs with Radial Layout," in *Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, pp. 43−50.

Yi, J. S., Kang, Y. A., Stasko, J., and Jacko, J. (2007), "Toward a Deeper Understanding of the Role of Interaction in Information Visualization," *IEEE Transactions on Visualization and Computer Graphics,* Vol. 13, No. 6, pp. 1224−1231.

Young, D., and Shneiderman, B. (1993), "A Graphical Filter/Flow Representation of Boolean Queries: A Prototype Implementation and Evaluation," *Journal of the American Society of Information Science*, Vol. 44, No. 6, pp. 327−339.

Zhou, M., and Feiner, S. (1998), "Visual Task Characterization for Automated Visual Discourse Synthesis," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp. 392−399.

# CHAPTER 44

# HUMAN FACTORS IN ONLINE COMMUNITIES AND SOCIAL COMPUTING

**Panayiotis Zaphiris**
**Department of Multimedia and Graphic Arts**
**Cyprus University of Technology**
**Limassol, Cyprus**

**A. Ant Ozok**
**Department of Information Systems**
**University of Maryland Baltimore County**
**Baltimore, Maryland**

## 1 INTRODUCTION

The popularity of online communities is expanding. It was estimated that there were over 1.70 billion Internet users globally in 2009 (Pingdom, 2010) with one in four Internet users participating in chat rooms or online discussions (Madden and Rainie, 2003). Especially social networking sites have been exponentially increasing in population in the last few years. For example, just on the social network site Facebook, there are around 350 million users, and 50% of these users log in everyday (Pingdom, 2010).

In this chapter we try to provide a synopsis of the topic of human factors for online communities and social computing by first defining online communities and computer-mediated communication (CMC). This is followed by a review of the different types of CMC, with specific categories of online communities described in more depth. The chapter concludes with a brief summary and suggestions for new directions in the area of online communities.

## 2 DEFINITION OF ONLINE COMMUNITIES

Online communities emerge through the use of CMC applications. The term online community is multidisciplinary in nature, means different things to different people, and is slippery to define (Preece, 2000). There are a number of different definitions of online communities One provided by Rheingold (1993, p. 5) states that "[online] communities are social aggregations that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace."

The cyberspace is the new frontier in social relationships, and people are using the Internet to make friends, colleagues, lovers, as well as enemies (Suler, 2004). As Korzeny pointed out, even as early as 1978, online communities are formed around interests and not physical proximity (Korzenny, 1978). People with common interests, such as hobbies, ethnicity, education, and beliefs are brought together through online communities

to discuss, debate, and share knowledge about these issues. As Wallace (1999) points out, meeting in online communities eliminates prejudging based on someone's appearance, and thus people with similar attitudes and ideas are attracted to each other.

Like any other technology, CMC has its benefits as well as its limitations. For instance, CMC discussions are often potentially richer than face-to-face discussions. However, users with poor writing skills may be at a disadvantage when using text-based CMC (SCOTCIT, 2003).

## 3 BRIEF HISTORY OF COMMUNITIES IN SOCIETY AND THEIR FUNCTION

The communities concept probably dates back several millennia to the days of ancient people who realized that working together and communicating help in their accomplishments.

There have been significant disagreements concerning the definition of an online community. Preece (2001) indicated that the online community concept means different things to different people. While some people may see an online community as a virtual place to exchange ideas and opinions, others may see it as a virtual environment to share the daily happenings in their lives, yet others may see it as an area where they can sell their goods and make a profit. At their beginnings, online communities have also been defined as virtual environments where it was easier to create networks of hatred and support deviant behavior (Preece, 2001).

While taking into consideration the disagreements concerning what exactly an online community entails, the idea of coming together on a networked environment can be traced back to the early days of the user network (usenet) systems in the early 1990s. It can be argued that with the exponential growth of the World Wide Web starting in 1994 a large number of these networks built around special-interest, demographic, or occupational groups moved to the Web-based environment. At the beginning most online communities were by invitation only or required a strict process to join, but as the demand grew, the sign-up processes became easier and online communities reached millions of online members.

Whereas the most common early Web-based online communities were made up of professional and special-interest groups, over the years the variety of services that online communities offered increased dramatically. Towards the end of the 1990s, communities that cater to individuals and their social activities and interactions increased at a rapid pace. The number of gaming communities has also grown rapidly in the last 15 years, with multiuser dungeons (MUDs) and multiuser object-oriented dungeons (MOOs) in the 1990s and early 2000s being largely replaced by massive multiplayer online role-playing games (MMORPGs). The online gaming concept easily qualifies as an online community as it allows interaction between the players in a number of ways, including live chat and video.

## 4 TODAY'S ONLINE COMMUNITIES AND THEIR FUNCTIONS

Today, the social structure of the World Wide Web largely relies on the online communities framework, where the social aspect of communities is easily part of the goal definitions in most sites. Any site with a bulletin board can be considered an online community site. Other types of communications for qualifying to be an online community site include having an online chat or e-mail option, a news group, list server, or other type of community marketing option. In the last decade, Wellman (1997) identified two types of relationships on online communities: strong-tie and weak-tie relationships. In today's communities, relationships among members vary to a great extent. On a social networking site such as MySpace® or Facebook®, interacting members know each other on a first-name basis, and most members know each other before becoming "friends" on the online community. This is an example of a strong-tie relationship with participants personally knowing each other. On a business-oriented online community, however, participants may only be interested in interacting on the basis of their job identities and to exchange information for hiring and other job-related purposes. This may be an example of a weak-tie relationship where participants do not know each other personally. An example potentially difficult to categorize would be a health care–related online community for cancer patients where community members may exchange information concerning their illness without knowing their personal identities for the purpose of protecting their identity. Communities for practice can exchange information with the ultimate goal of making a profit. In short, due to the sheer number of communities, the relationships between their members vastly vary, with some community members having a tight relationship while others have a professional-only relationship while many other communities have relationships between the members somewhere in between. Whereas initially online communities were sometimes categorized based on the technologies they used to communicate (such as bulletin boards, newsgroups, or "chat networks,"), now online communities are able to combine multiple, sometimes comprehensive communication technologies to provide their members with a wide variety of options to communicate with each other. Additionally, the advent of multimedia capabilities in online environments allows some online communities to turn into "multimedia hubs" where participants can post any image, sound bite or video they want. This allows for more interactive as well as information-rich interactions among members.

Online communities of today deliver a number of services to their customers, which include individuals as well as companies. There are a number of functionalities; the list below presents a sample of the most common ones. They are organized in three broad categories: communities for individuals, communities for professionals, and communities for organizations:

### Communities for Individuals
- Social communities to share their social lives and socialize

- Health care communities to share health-related advice, information, and experiences
- Cultural and special-interest group communities to share information (book reading, movie communities, etc.)
- Shopping communities to shop and share information and experiences (Amazon, E-Bay, etc.)
- Educational communities (e.g., online classes)
- Political communities

### Communities for Professionals

- To share information regarding their professional skills for collaboration and employment possibilities (LinkedIn®, Pipl®, etc.)
- For members of the same profession to exchange information (community for spinal surgery, community for real estate sales professionals)

### Communities for Organizations

- Health care organizations
- Retailers and other for-profit organizations
- Social work
- Government
- International organizations
- Customer- and company-specific portals

In the next chapter, the most significant types of communities are discussed with a human factors emphasis in more detail.

## 5 TYPES OF COMPUTER-MEDIATED COMMUNICATION AND ONLINE COMMUNITIES

### 5.1 Communities for Interest Groups

Different interest groups can build their own community, for example, those interested in business, sports, books, movies, and music. The types of online communities presented in this section cannot be considered comprehensive as the very high number of interest groups have resulted in thousands of online community types. In designing online communities for interest groups, designers need to keep in mind that the users of these communities may potentially have little in common except for the concept of the online community. Therefore, a broad group of user specifications need to be considered in designing these communities with several options that can be presented to the users for the purposes of customization. Interface designs in these communities can be kept simple with minimal use of multimedia. The usability and human computer interaction principles developed by Shneiderman (1992) and Nielsen (1993) can to some great extent be applied to these types of communities, taking into consideration the broad user range, resulting in a design that can be usable for different demographic and cultural groups such as users from different countries as well as from different age groups and education levels.

Different interest groups will have different expectations from the online communities design. For example,

users of an interest group on literature may deal with largely text-based interfaces while an interest group on movies would have an online community interface which would be heavy on video clips and images. A basic set of usability principles can be followed in design for online communities at a minimum. Today, just as in other types of online communities, special-interest group communities also offer customization features where users can choose interface elements such as colors, layout, and text size.

### 5.2 Communities for Health Care

In recent years, the number of communities specializing in health care has increased dramatically. Health care communities can target users with certain conditions, healthy users, or organizations. Health-focused communities can be for-profit or nonprofit organizations, while they are mostly free to individual members except for cases where some special premium services are offered (e.g., finding a doctor for a patient). Health care communities targeting users with certain conditions (such as cancer or epilepsy) are in most cases informational and private. Users can interact with each other without knowing each others' names. In these types of health care communities, most information exchanged deals with certain conditions and symptoms relating to ailments and their solutions. Additionally, the sites can provide some advice and recommendations for the users to improve their quality of life, either through automatically generated suggestions or by getting help from health care professionals. Similarly, health care community sites targeting healthy users usually focus on health care advice and recommendations to lead a healthy life. Recently, sites like Google Health® and Microsoft Health Vault® target the healthy population with the goal of allowing interactions for a healthy life style for different age groups, in some cases specifically the elderly population.

Usability and human factors issues in health care communities should be approached with caution as some health care communities will have target users with specific ailments that may require special design compensations for their user groups (e.g., online communities targeting users with visual ailments or blindness). The design issues in health care communities may need to be considered on a case-by-case basis depending on the nature of the community and the types of users it targets. Whereas most design issues may deal with visual elements and navigation, if the target audience consists of common individuals (sometimes referred to as "health consumers"), design issues for populations with disabilities need to be given special emphasis in the design if the target population generally has such a disability. For example, if the user group consists of people with low vision, the design of the online community Web pages should compensate for this by providing strong contrast and larger text and image sizes as well as ensuring that the sites are compatible for screen readers for the blind.

### 5.3 Online Virtual Game Communities

With the advent of ubiquitous broadband Internet connection and the increasing graphical processing power

of personal computers, a new paradigm of gaming has emerged. A paradigm of gaming that allows players to remotely play together the same computer game, theMMORPGs, has changed the game industry dramatically. MMORPGs provide a fictional setting where a large group of users voluntarily immerse themselves in a graphical virtual environment and interact with each other by forming a community of users.

Although the concept of multiplayer gaming is not new, the game world of most local network multiplayer games, as opposed to MMORPG, is simplistic and can accommodate only around 16 concurrent players in a limited space.

A MMORPG enables thousands of players to simultaneously play in an evolving virtual world over the Internet. The game world is usually modeled with highly detailed three-dimensional (3D) graphics, allowing individuals to interact not only with the gaming environment but also with other players. Usually this involves the players representing themselves through the use of avatars—the visual representation of the player's identity in the virtual world.

The MMORPG environment is a new paradigm in computer gaming in which players are part of a persistent world, a world that exists independent of the users (Yee, 2005). Unlike other games where the virtual world ceases to exist when players switch off the game, in an MMORPG, the world exists before the user logs on and continues to exist when the user logs off. More importantly, events and interactions occur in the world even when the user is not logged on as there are many other players who are constantly interacting, thus transforming the world. To accommodate the large number of users, the worlds in MMORPGs are vast and varied in terms of "geographical locations," characters, monsters, items, and so on. More often than not, new locations or items are added by the game developers from time to time according to the demands of the players.

An MMORPG, like any role-playing game (RPG), involves killing monsters, collecting items, developing characters, and so on. However, it also contains an extra aspect of internal sociability. Unlike single-player games which rely on other external modes of communication (such as mailing lists, discussion forums outside the game) to form the gaming culture, the culture is formed within the MMORPG environment itself.

In such a way, these MMORPG virtual worlds represent the persistent social and material world, which is structured around narrative themes (usually fantasy) where players are engaged in various activities: slaying monsters, attacking castles, scavenging for goods, trading merchandise, and so on. On one hand, the game's virtual world represents the escapist fantasy; on another, it supports social realism (Kolbert, 2001).

This means games are no longer meant to be a solitary activity played by a single individual. Instead, the player is expected to join a virtual community that is parallel with the physical world in which societal, cultural, and economical systems arise. It has gradually become a world that allows players to immerse themselves into experiences which closely match those of

the real world: seek virtual relationships, hold virtual marriages, set up virtual shops, and so on.

Such games are ripe for cultural analysis of the social practices around them. Although fundamentally MMORPGs are video games with virtual spaces where the players interact, they should be regarded not just as a game software but as a community, a society, and, if you wish, a culture. These games are becoming the most interesting interactive computer-mediated communication and networked activity environments (Taylor, 2002). Thus, understanding the pattern of participation in these game communities is crucial, as such virtual communities function as a major mechanism of socialization of the players to the norms of the game culture that emerges.

Such communities that formed around the game can be broadly divided into two categories: in-game and out-of-game communities. Most MMORPGs are created to encourage long-term relationships among the players through the features that support the formation of in-game communities. One of the most evident examples is the concept of guilds. Guilds are a fundamental component of the MMORPG culture for people who are natural organizers to run a virtual association which has formalized membership and rank assignments to encourage participation. Sometimes, a player might join a guild and get involved in a guild war in order to fight for the castle. Each guild usually has a leader and several guilds can team up in a war. This involves complicated leader–subordinate and leader–leader relationships.

Apart from relatively long-term relationships such as guild communities, MMORPGs also provide many opportunities for short-term relationship experiences. For example, a player could team up with another player to kill monsters in order to develop the abilities of their avatars (level up) or some more expert players could help newer players get through the game.

When trying to win the game, players often need to get information from other resources: guidebooks, discussion forums, other players, and so on. Therefore, game playing is generally more concerned with player–player interaction than with player–game interaction. What is at first confined to the game alone soon spills over into the virtual world beyond it (e.g., websites, chat rooms, email) and even life off-screen (e.g., telephone calls, face-to-face meetings).

Apart from these external communities around the game which are mediated through e-mails or online forums (which also exist in many other games), there is an interesting phenomenon that fuses the internal and external game communities. The participation in an external community starts to break the magic circle of the game—the game space is no longer separate from real life—as the out-of-game community trades in game items for real money.

For example, Norrath, the world of EverQuest, was estimated to have the seventy-seventh largest economy in the real world based on buying and selling in online auction houses (Castronova, 2001, p. 1):

About 12,000 people call it their permanent home, although some 60,000 are present there at any given time. The nominal hourly wage is about USD3.42

per hour, and the labors of the people produce a GNP per capita somewhere between that of Russia and Bulgaria. A unit of Norrath's currency is traded on exchange markets at USD0.0107, higher than the Yen and the Lira.

Having illustrated the social phenomenon around such a playful virtual community, it is believed that it is fruitful to research such communities as we might be able to derive some useful implications on how successful computer-supported collaborative work (CSCW) and computer-supported collaborative learning (CSCL) environments can be designed. For this reason, in Section 6 we will describe some of the methodologies that can be used in such studies, and in Section 7 we will present the application of some of these methods to two case studies.

### 5.4 Communities for Profit

A community for profit can be in three main forms, two of which employ the community aspects and processes directly to generate revenue, while the online community aspect may be a supporting factor in generating revenues for the profit-making entity.

By definition, online communities are made up of people with common interest or profession coming together electronically to exchange ideas and allow collaboration. Although online communities aimed at providing the electronic environment to support this communication mostly target individuals, companies can also play a role as customers for such online communities. Due to this focus on individual users, a large number of commonly known online communities generate revenue based on advertisements. Large communities such as Microsoft Network (MSN®), Google®, Facebook®, and Yahoo®, while varying in the types of services they deliver to individual customers, rely heavily on customized advertisements (Krammer, 2008). However, security and privacy concerns become critical issues when sites provide advertisements which are based on the content of the user-provided input, as these types of advertisements require that what the users type or do on the site is recorded without the user's knowledge in order to be used for advertisement purposes later (Hu et al., 2007). This issue should be considered in providing advertisements for individual users of online communities as part of the trade-off between revenue stream and violation of user rights concerning privacy.

The second method employed by the online communities for revenue generation includes charging users membership or one-time fees to use the online community site services. Because most social networking sites are free to end users and the aversion of common Internet users to pay for basic services such as e-mail services, bulletin boards, common information such as news, as well as other services offered by online communities, most community sites are free for basic membership. Those online communities that charge a fee for membership usually offer additional services on top of the common services or when the services provided are at a higher level and involve some additional costs to the online community provider. One example is a health community that provides advice to users from actual physicians on a case-by-case basis or experts in the area to answer user questions. Furthermore, sites can provide user-specific services such as recommending a doctor based on the patient-provided information. These "premium" services can be subject to a fee, although it can be argued that the majority of the major online communities are free to the end users and generate their revenue primarily from advertisers.

The third potential revenue generation method is arguably based on the notion that any electronic commerce company that employs a method such as user feedback on products can be considered an online community as it fits the definition of the provider of an electronic service that allows individuals with something in common (in this particular case, individuals that are interested in a product). Companies like Amazon® and eBay® largely rely on customers who provide feedback regarding their products for others to read, and this feedback plays a large role in other potential buyer's decisions to purchase the products. Electronic commerce companies see interaction among shoppers concerning products as a major component of customer relationship management (CRM), with the ultimate goal of positive product recommendations from other shoppers allowing shoppers to purchase the product, come back to shop for other products, and provide favorable feedback (Ozok et al., 2007). While the community issues may not be in the foreground for retail electronic commerce companies in general, this and some other studies suggest that treating the targeted consumers as a community and providing community-like services on their pages is becoming increasingly popular among online retailers.

One last relationship can be described between online communities and revenue generation in the form of portals. Consumer portals started as search engines for individual users to look up information relating to the keywords they entered, but since then, search engines have evolved to a rich information repository where users can look for the information they need. Portals like Google®, Yahoo!®, and Bing® can therefore also be seen as having some online community aspect where users can exchange information directly or indirectly. As the online community aspects of portals in this regard play an indirect role in revenue generation, it is difficult to categorize the design issues involving portals as they relate to these aspects. However, the online community aspects of portals may need to be taken into consideration in future portal design issues.

## 6 ANALYZING ONLINE COMMUNITIES: FRAMEWORKS AND METHODOLOGIES

Various aspects and attributes of CMCs can help us better understand online communities: for instance, analysis of the frequency of exchanged messages and the formation of social networks or analysis of the content of the exchanged messages and the formation of virtual communities. To achieve such an analysis a number of theoretical frameworks have been developed and proposed. For example, Henri (1992) provides an

analytical model for cognitive skills that can be used to analyze the process of learning within messages exchanged between students of various online e-learning communities. Mason's (1991) work provides descriptive methodologies using both quantitative and qualitative analysis. Furthermore, five phases of interaction analysis are identified in Gunawardena et al.'s (1997) model:

   I.   Sharing/comparing of information
   II.  Discovery and exploration of dissonance or inconsistency among ideas, concepts, or statements
  III.  Negotiation of meaning/coconstruction of knowledge
  IV.  Testing and modification of proposed synthesis or coconstruction
   V.  Agreement statement(s)/applications of newly constructed meaning

Some of the methods used are as follows:

*Interviews.* An interview can be defined as a type of conversation that is initiated by the interviewer in order to obtain relevant information. Interviews are usually carried out on a one-to-one basis where the interviewer collects information from the interviewee. Interviews can take place by telephone and face to face (Burge and Roberts, 1993). There are three types of interviews: (a) structured interviews: consist of predetermined questions asked in fixed order like a questionnaire; (b) semistructured interviews: questions are determined in advance but may be reordered, reworded, omitted, and elaborated upon; (c) unstructured interviews: are not based on predetermined questions but instead the interview has a general area of interest and the conversation may develop freely.

Interviews can be used to gain insights about general characteristics of the participants of an online community and their motivation for participating in the community under investigation. The data collected come straight from the participants of the online communities, whereby they are able to provide feedback based on their own personal experiences, activities, thoughts, and suggestions.

*Questionnaires.* A questionnaire is a self-reporting query-based technique. Questionnaires are typically produced on printed paper, but due to recent technologies and in particular the Internet, many researchers engage in the use of online questionnaires, thus saving time and money and eliminating the problem of a participant's geographical distance. There are three types of questions that can be used with questionnaires: open questions, where the participants are free to respond however they like; closed questions, which provide the participants with several choices for the answer; and scales where the respondents must answer on a predetermined scale.

*Log Analysis.* A log, also referred to as web-log, server log, or log-file, is in the form of a text file and is used to track the users' interactions with the computer system they are using. The types of interactions recorded include key presses, device movements, and other information about the user activities. The data are collected and analyzed using specialized software tools and the range of data collected depends on the log settings. Logs are also time stamped and can be used to calculate how long a user spends on a particular task or how long a user has lingered in a certain part of the website (Preece et al., 2002). In addition, an analysis of the server logs can help us find out: when people visited the site, the areas they navigated, the length of their visit, the frequency of their visits, their navigation patterns, from where they are connected, and details about the computer they are using.

*Content and Textual Analysis.* Content analysis is an approach to understanding the processes that participants engage in as they exchange messages (McLoughlin, 1996). There have been several frameworks created for studying the content of messages exchanged in online communities.

*Social Network Analysis* (*SNA*). According to Krebs (2004, p. 1), "Social Network Analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organizations, computers or other information/knowledge processing entities. The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships." Preece (2000) adds that it provides a philosophy and a set of techniques for understanding how people and groups relate to each other and has been used extensively by sociologists (Wellman, 1982, 1992), communication researchers (Rice, 1994; Rice et al., 1990), and others. Analysts use SNA to determine if a network is tightly bounded, diversified, or constricted; to find its density and clustering; and to study how the behavior of network members is affected by their positions and connections (Garton et al., 1997; Henneman, 1998; Scott, 2000).

There are two approaches to SNA:

*Ego-Centered Analysis.* Focuses on the individual as opposed to the whole network, and only a random sample of the network population is normally involved (Zaphiris et al., 2003). The data collected can be analyzed using standard computer packages for statistical analysis (Garton et al., 1997).

*Whole-Network Analysis.* The whole population of the network is surveyed and this facilitates conceptualization of the complete network (Zaphiris et al., 2003). The data collected can be analyzed using microcomputer programs like UCINET and Krackplot (Garton et al., 1997).

The following are important units of analysis and concepts of SNA (Garton et al., 1997; Wellman, 1982, 1992; Hanneman, 2001; Zaphiris et al, 2003):

- Nodes: The actors or subjects of study.
- Relations: The strands between actors. They are characterized by content,
- Direction and strength.
- Ties: Connect a pair of actors by one or more relations.
- Multiplexity: The more relations in a tie, the more multiplex the tie is.
- Composition: This is derived from the social attributes of both participants.
- Range: The size and heterogeneity of the social networks.
- Centrality: Measures who is central (powerful) or isolated in networks.
- Roles: Network roles are suggested by similarities in the network members' behavior.
- Density: The number of actual ties in a network compared to the total amount of ties that the network can theoretically support.
- Reachability: In order to be reachable, connections that can be traced from the source to the required actor must exit.
- Distance: The number of actors that information has to pass through to.
- Connect one actor with another in the network.
- Cliques: Subsets of actors in a network who are more closely tied to each other than to other actors who are not part of the subset.

Usability issues involving online community sites can be concluded to not be radically different from those that involve design of commercial websites. General design guidelines involving usability design of Web pages can to a great extent be observed in online community design. Major design issues specific to online communities may include those involving the potential cognitive and physical limitations of the targeted user group to which the community caters. However, as indicated earlier, producing general usability guidelines for online communities is difficult due to the different user groups for which the many online communities provide services.

## 7  CASE STUDIES

In this section we present two case studies that demonstrate the use of theoretical and analytical techniques for studying online communities. In the first case study, we demonstrate how the results from an attitude towards thinking and learning questionnaire can be combined with SNA to describe the dynamics of a computer-aided language learning (CALL) online community. In the second case study, we present a theoretical activity model that can be used for describing interactions in online game communities.

## 7.1  Computer-Aided Language Learning Communities

In the first case study we demonstrate a synthetic use of quantitative (SNA) and qualitative (questionnaire) methods for analyzing the interactions that take place in a CALL course. Data were collected directly from the discussion board of the "Learn Greek Online" (LGO) course (Kypros-Net, 2005).

LGO is a student-centered e-learning course for learning Modern Greek and was built through the use of a participatory design and distributed constructionism methodology (Zaphiris and Zacharia, 2001). In an ego-centered SNA approach, we have carried out an analysis of the discussion postings of the first 50 actors (in this case the students of the course) of LGO.

To carry out the SNA, we used "NetMiner" (Cyram, 2004), a tool which enables us to obtain centrality measures for our actors. The "in- and out-degree centrality" was measured by counting the number of interaction partners per individual in the form of discussion threads (e.g., if an individual posts a message to three other actors, then his or her out-degree centrality is 3, whereas if an individual receives posts from five other actors, then his or her in-degree is 5).

Due to the complexity of the interactions in the LGO discussion, we had to make several assumptions in our analysis:

Posts that received no replies were excluded from the analysis. This was necessary in order to obtain meaningful visualizations of the interaction.

Open posts were assumed to be directed to everyone who replied.

Replies were directed to all the existing actors of the specific discussion thread unless the reply or post was specifically directed to a particular actor.

In addition to the analysis of the discussion board interactions we also collected subjective data through the form of a survey. More specifically, the students were asked to complete an Attitudes Towards Thinking and Learning Survey (ATTLS). The ATTLS measures, through the use of 20 Likert scale questions, the extent to which a person is a "connected knower" (CK) or a "separate knower" (SK). People with higher CK scores tend to find learning more enjoyable and are often more cooperative, more congenial, and more willing to build on the ideas of others, while those with higher SK scores tend to take a more critical and argumentative stance to learning (Galotti et al., 1999).

The out-degree results of the social network analysis are depicted in Figure 1 in the form of a sociogram, and the in-degree results are depicted in Figure 2. Each node represents one student (to protect the privacy and anonymity of our students their names have been replaced by a student number). The position of a node in the sociogram is representative of the centrality of that actor (the more central the actor, the more active). As can be seen from Figure 1, students S12, S7, S4, and S30 (with out-degree scores ranging from 0.571 to 0.265) are at the center of the sociogram and possess the
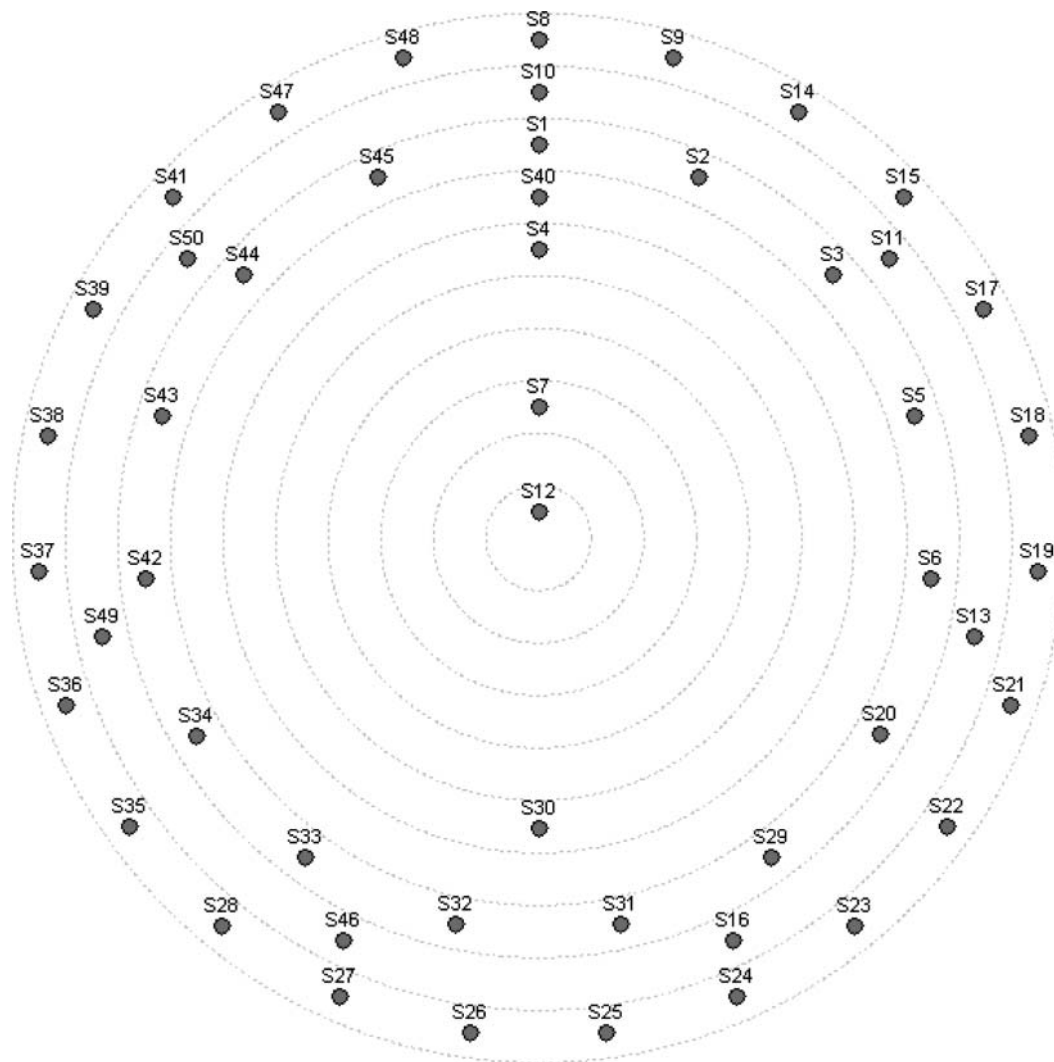
**Figure 1**  Out-degree analysis sociogram.

highest out-degree. The same students also possess the highest in-degree scores (Figure 2). This is an indication that these students are the most active members of this online learning community, posting and receiving the largest number of postings. In contrast, participants in the outer circle (e.g., S8, S9, S14) are the least active with the smallest out-degree and in-degree scores (all with 0.02 out-degree scores).

In addition, a clique analysis was carried out (Figure 3) and it showed that 15 different cliques (the majority of which are overlapping) of at least three actors each have been formed in this community.

As part of the ego-centered analysis for this case study we look in more detail at the results for two of our actors: S12, who is the most central actor in our SNA analysis, that is, with the highest out-degree score,

and S9, an actor with the smallest out-degree score. It is worth noting that both members joined the discussion board at around the same time.

First, through a close look at the clique data (Table 1), we can see that S12 is a member of 10 out of the 15 cliques, whereas S9 is not a member of any, an indication of the high interactivity of S12 versus the low interactivity of S9. In an attempt to correlate the actors' position in the SNA sociogram with their self-reported attitudes toward teaching and learning, we looked more closely at the answers these two actors (S12, S9) provided to the ATTLS. Actor S12 answered all 20 questions of the ATTLS with a score of at least 3 (on a 1–5 Likert scale) whereas S9 had answers ranging from 1 to 5. The overall ATTLS score of S12 is 86 whereas that of S9 is 60. A clear dichotomy of opinions
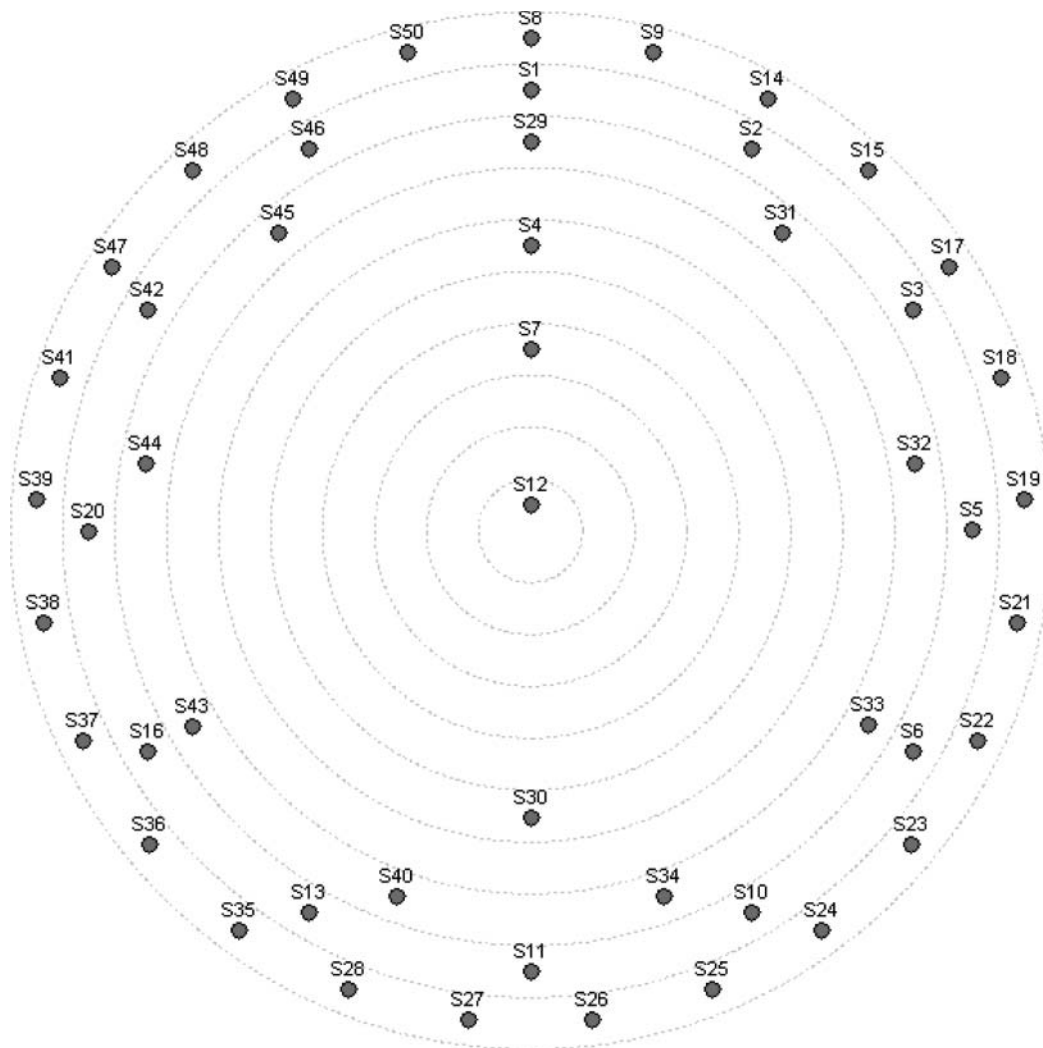
**Figure 2** In-degree analysis sociogram.

occurred on 5 of the 20 questions of the ATTLS. S12 answered all 5 of those questions with a score of 5 (strongly agree) whereas S9 answered them with a score of 1 (strongly disagree). More specifically, S12 strongly agreed that he or she:

1. Is more likely to try to understand someone else's opinion than to try to evaluate it.
2. Often finds herself or himself arguing with the authors of books read, trying to logically figure out why they're wrong.
3. Finds that he or she can strengthen his or her position through arguing with someone who disagrees with them.
4. Feels that the best way to achieve his or her own identity is to interact with a variety of other people.

5. Likes playing devil's advocate—arguing the opposite of what someone is saying.

S9 strongly disagreed with all of the above statements. These are all indications that S12 is a CK whereas S9 is a SK.

This case study showed that the combination of quantitative and qualitative techniques can facilitate a better and deeper understanding of online communities.

## 7.2 Game Communities and Activity Theoretical Analysis

The main motivation of the second case study arises from the more general area of computer game-based learning. Game-based learning has focused mainly on how the game itself can be used to facilitate learning activities, but we claim that the educational
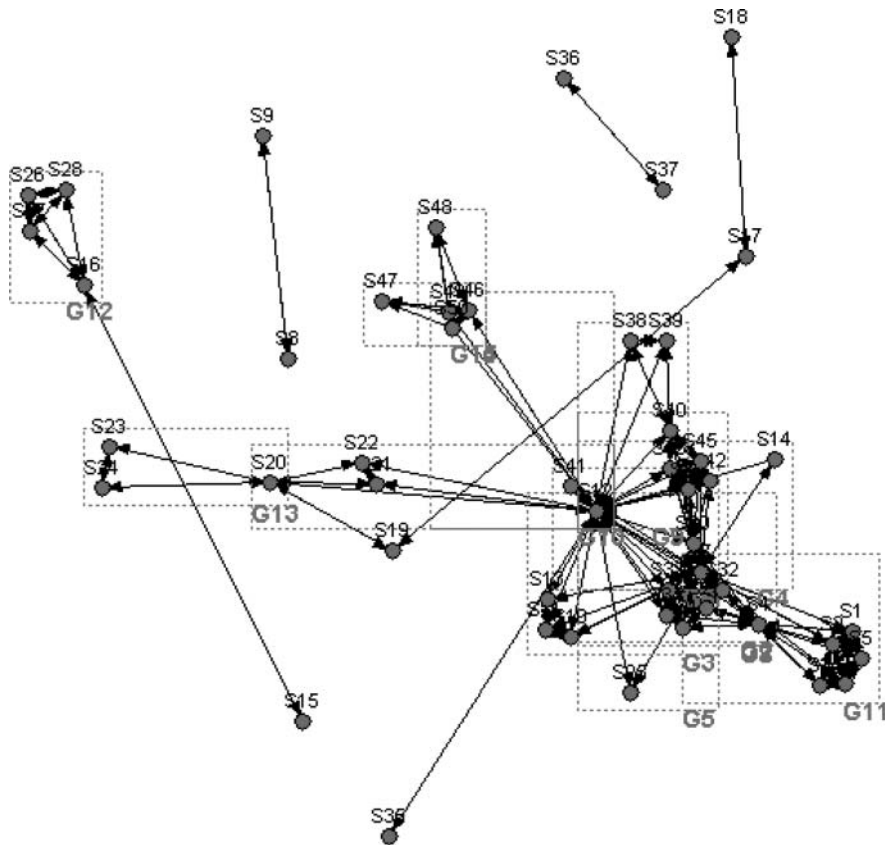
**Figure 3**   Clique analysis sociogram.

opportunity in computer games stretches beyond the learning activities in the game per se. Indeed, if you observe most people playing games, you will likely see them downloading guidelines from the Internet and

**Table 1  Clique Analysis of LGO Discussions**

| Cliques | Actors |
| --- | --- |
| K1 | S12, S7, S30, S40, S43, S44, S45 |
| K2 | S12, S7, S30, S4 |
| K3 | S12, S7, S10, S11, S13 |
| K4 | S12, S7, S14 |
| K5 | S12, S7, S25 |
| K6 | S12, S7, S41 |
| K7 | S12, S20, S21, S22 |
| K8 | S12, S29, S4, S30, S31, S32, S33, S34 |
| K9 | S12, S38, S39, S40 |
| K10 | S12, S46, S49, S50 |
| K11 | S1, S2, S3, S4, S5, S6, S7 |
| K12 | S16, S26, S27, S28 |
| K13 | S23, S20, S24 |
| K14 | S47, S46, S49, S50 |
| K15 | S48, S46, S49, S50 |

participating in online forums to talk about the game and share strategies. In actuality, almost all game playing could be described as a social experience, and it is rare for a player to play a game alone in any meaningful sense (Kuo, 2004). This observation is even more evident in MMORPGs, which have been discussed earlier in this chapter. For example, the participation in a MMORPG is constituted through language practice within the in-game community (e.g., in-game chatting and joint task) and out-of-game community (e.g., the creation of written game-related narratives and fan-sites). The learning is thus not embedded in the game, but it is in the community practice of those who inhabit it.

We believe that the study of computer games should be expanded to include the entire game community. Computer game communities can be categorized into three classes which we have identified (Figure 4) (Ang et al., 2005) as:

*Single Game Play Community*. This refers to a game community formed around a single-player game. Although players of a single-player game like The Sims 2 and Final Fantasy VII play the game individually, they are associated with an out-of-game community which discusses the game either virtually or physically.

**Figure 4** Types of game communities: (*a*) single game play community; (*b*) social game play community; (*c*) distributed game play community.

*Social Game Play Community*. This refers to multiplayer games which are played together in the same physical location. It creates game communities at two levels: in-game and out-of-game. Occasionally, these two levels might overlap. The out-of-game interaction might be affected by issues beyond the specific game system; for example, the community starts exchanging information about another game.

*Distributed Game Play Community*. This is an extension of the social game play community, but it emphasizes the online multiplayer game in which multiple sessions of a game are established in different geographical locations.

The study of game communities, especially out-of-game communities, from the perspective of education is still very much unexplored. We believe the potential of games in education is not limited to what is going on in the game. Educators could benefit by studying games as a social community because games are now becoming a culture that permeates the life of everyone, especially the younger generation. Black (2004) has investigated the interactions among participants in a virtual community of Japanese comic fans which involve a lot of reading and writing throughout the site. She examines how the fans in the community help each other with English language writing skills and with cross-cultural understanding. In this section we have pointed out that game communities can emerge from both single-player and multiplayer games. We believe that, by further studying the social interaction in the game community, we will be able to utilize games in learning in a more fruitful way. In the next section, we apply and evaluate one of these models of game communities to a specific scenario in knowledge building using activity theory.

## 8 FUTURE ISSUES IN ONLINE COMMUNITY DESIGN

Online communities have evolved very rapidly in the last 15 years. The most popular online community in the world today (according to mostpopularwebsites.net), Facebook®, is only six years old. While it is difficult to foresee the future and next steps in online communities in terms of popularity as well as in terms of human factors and design issues, it would be safe to say that online communities will continue their rapid change for the foreseeable future. It is likely that more interactive features will become prevalent for online communities serving individuals. Users will be able to communicate more commonly via videoconferencing, and the reach of relatives, friends, and colleagues will become easier with the help of continuous connection to online communities. Microblogging provided by sites such as Twitter® is likely to improve its features, allowing video- and audio-blogging. Additionally, mobile devices are already extremely prevalent, and online communities are likely to allow users to be constantly connected to their communities via their cell phone or other mobile device interfaces. Online communities are also likely to increase in the role of connecting people and professionals with each other, with capabilities of communities improving in terms of speed and accuracy.

With the increase in popularity and capabilities of online communities, some challenges in online community design from human factors perspectives also need to be acknowledged. As users are likely to share private information on online communities which is intended for only those individuals that the users approve of, privacy and security are becoming major concerns in online communities. The concerns of users need to be addressed via community policies as well

as technologies that ensure that the information being shared in online communities is secure and intrusion does not happen. As online communities evolve, users with more limited technology may have a difficult time keeping pace with newly developing technologies or may move on to other communities. In the highly competitive environment of online communities, users have many options, and communities can lose popularity easily. Research in human factors and human computer interaction will allow community providers to keep track of user needs, preferences, and limitations and provide optimal communities for all types of users.

# REFERENCES

Ang, C. S., Zaphiris, P., and Wilson, S. (2005), "Social Interaction in Game Communities and Second Language Learning," paper presented at The 19th British HCI Group Annual Conference, Edinburgh, United Kingdom.

Black, R. W. (2004), "Anime-Inspired Affiliation: An Ethnographic Inquiry into the Literacy and Social Practices of English Language Learners Writing in the Fan-Fiction Community," paper presented at 2004 Meeting of American Educational Research Association, San Diego, CA.

Burge, E. L., and Roberts, J. M. (1993), *Classrooms with a Difference: A Practical Guide to the Use of Conferencing Technologies*, University of Toronto Press, Ontario.

Castronova, E. (2001, December), "Virtual Worlds: A First-Hand Account of Market and Society on the Cyberian Frontier," CESifo Working Paper Series No. 618, Center for Economic Studies and Ifo Institute for Economic Research, California State University, Fullerton, CA.

Cyram (2004), "Netminer for Windows," available: http://netminer.com.

Galotti, K. M., Clinchy, B. M., Ainsworth, K., Lavin, B., and Mansfield, A. F. (1999), "A New Way of Assessing Ways of Knowing: The Attitudes towards Thinking and Learning Survey (ATTLS)," *Sex Roles*, Vol. 40, No. 9/10, pp. 745–766.

Garton, L., Haythorthwaite, C., and Wellman, B. (1997), "Studying On-Line Social Networks," in *Doing Internet Research*, S. Jones, Ed., Sage, Thousand Oaks CA.

Gunawardena, C., Lowe, C., and Anderson, T. (1997), "Analysis of a Global Online Debate and the Development of an Interaction Analysis Model for Examining Social Construction of Knowledge in Computer Conferencing," *Journal of Educational Computing Research*, Vol. 17, No. 4, pp. 397–431.

Hanneman, R. A. (2001), "Introduction to Social Netwok Methods," available: http://faculty.ucr.edu/~hanneman/SOC157/TEXT/TextIndex.html, accessed November 9, 2004.

Henri, F. (1992), "Computer Conferencing and Content Analysis," in *Collaborative Learning through Computer Conferencing: The Najaden Papers*, A. R. Kaye, Ed., Springer-Verlag, Berlin, pp. 117–136.

Hu, J., Zeng, H.-J., Li, H., Niu, C., and Chen, Z. (2007), "Demographic Prediction Based on User's Browsing Behavior," in *Proceedings of the 16th International Conference on World Wide Web*, Banff, Alberta, Canada.

Kolbert, E. (2001, May 28), "Pimps and Dragons: How an Online World Survived a Social Breakdown," *The New Yorker*, available: http://www.newyorker.com/fact/content/?010528fa_FACT, accessed February 15, 2005.

Korzenny, F. (1978), "A Theory of Electronic Propinquity: Mediated Communication in Organizations," *Communication Research*, Vol. 5, pp. 3–23.

Krammer, V. (2008), "An Effective Defense against Intrusive Web Advertising," paper presented at the PST '08 Sixth Annual Conference on Privacy, Security and Trust, Fredericton, New Brunswick, Canada.

Krebs, V. (2004), "An Introduction to Social Network Analysis," available: http://www.orgnet.com/sna.html, accessed November 9, 2004.

Kuo, J. (2004), "Online Video Games in Mental Health," paper presented at the Annual Meeting of the American Psychiatry Association, New York.

Kypros-Net (2005), "The World of Cyprus," available: http://kypros.org.

Madden, M. and Rainie, L. (2003), "Pew Internet & American Life Project Surveys," Pew Internet & American Life Project, Washington, DC.

Mason, R. (1991), "Analyzing Computer Conferencing Interactions," *Computers in Adult Education and Training*, Vol. 2, No. 3, pp. 161–173.

McLoughlin, C. (1996), "A Learning Conversation: Dynamics, Collaboration and Learning in Computer Mediated Communication," in *Proceedings of the Third International Interactive Multimedia Symposium*, Promaco Conventions, Perth, Western Australia, January 21–25, C. McBeath and R. Atkinson, Eds., pp. 267–273.

Nielsen, J. (1993), *Usability Engineering*, Academic Press Professional, Chestnut Hill, MA.

NUA (2004), "NUA Internet Surveys," available: http://www.nua.ie/surveys/how_many_online/index.html, accessed October 20, 2004.

Özman, M. (2007), "Network Formation and Strategic Firm Behaviour to Explore and Exploit," Journal of Artificial Societies and Social Simulation, Vol. 11, No. 17, available: http://jasss.soc.surrey.ac.uk/11/1/7.html.

Ozok, A. A., Oldenburger, K. and Salvendy, G. (2007), "Impact of Consistency in Customer Relationship Management on E-Commerce Shopper Preferences," *Journal of Organizational Computing and E-commerce,* Vol. 17, No. 4, pp. 283–309.

Papert, S. (1980), *Mindstorms: Children,Computers, and Powerful Ideas*, Basic Books, New York, NY.

Pingdom (2010), "Internet 2009 in Numbers," available: http://royal.pingdom.com/2010/01/22/internet-2009-in-numbers/.

Preece, J. (2000), *Online Communities: Designing Usability, Supporting Sociability*, Wiley, Chichester.

Preece, J. (2001), "Sociability and Usability: Twenty Years of Chatting Online," *Behavior and Information Technology Journal*, Vol. 20, No. 5, pp. 347–356.

Preece, J., Rogers, Y., and Sharp, H. (2002), *Interaction Design: Beyond Human-Computer Interaction*, Wiley, New York.

Rheingold, H. (1993), *The Virtual Community: Homesteading on the Electronic Frontier*, Addison-Wesley, Reading, MA.

Rice, R. (1994), "Network Analysis and Computer Mediated Communication Systems," in *Advances in Social Network Analysis*, S. W. J. Galaskiewkz, Ed., Sage, Newbury Park, CA.

Rice, R. E., Grant, A. E., Schmitz, J., and Torobin, J. (1990), "Individual and Network Influences on the Adoption and Perceived Outcomes of Electronic Messaging," *Social Networks*, Vol. 12, pp. 17–55.

SCOTCIT (2003), "Enabling Large-Scale Institutional Implementation of Communications and Information Technology (ELICIT): Using Computer Mediated Conferencing," available: http://www.elicit.scotcit.ac.uk/modules/cmc1/welcome.htm, accessed November 2, 2004.

Scott, J. (2000), *Social Network Analysis: A Handbook*, 2nd ed., Sage, London.

Shneiderman, B. (1992), *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, Addison-Wesley, Reading, MA..

Suler, J. (2004), "The Final Showdown between In-Person and Cyberspace Relationships," available: http://www1.rider.edu/~suler/psycyber/showdown.html, accessed November 3, 2004.

Taylor, T. L. (2002), "Whose Game Is This Anyway?: Negotiating Corporate Ownership in a Virtual World", in F. Mayra (ed.) *CGDC Conference Proceedings*, Tampere, Tampere University Press, pp. 227–242.

Wallace, P. (1999), *The Psychology of the Internet*, Cambridge University Press, Cambridge.

Wellman, B. (1982), "Studying Personal Communities," in *Social Structure and Network Analysis*, P. M. N. Lin, Ed., Sage, Beverly Hills, CA.

Wellman B. (1992), "Which Types of Ties and Networks Give What Kinds of Social Support?" *Advances in Group Processes*, Vol. 9, pp. 207–235.

Wellman, B. (1997), "An Electronic Group Is Virtually a Social Network," in *Culture of the Internet*, Kiesler, S., Ed., Lawrence Erlbaum Associates. Mahwah, NJ, pp. 179–205.

Yee, N. (2005), "The Demographics and Derived Experiences of Users of Massively Multi-User Online Graphical Environments," paper presented at The 55th Internal Communication Association Annual Conference, New York, May 2005.

Zaphiris, P., and Zacharia, G. (2001), *Design Methodology of an Online Greek Language Course*, Ext. Abstracts CHI 2001, ACM Press, New York.

Zaphiris, P., Zacharia, G., and Rajasekaran, M. (2003), "Distributed Constructionism through Participatory Design," in *E-Education Applications: Human Factors and Innovative Approaches*, Ghaoui, C., Ed., Idea Group Publishing, London.

# CHAPTER 45

# HUMAN FACTORS AND INFORMATION SECURITY

**E. Eugene Schultz**
**Emagined Security**
**San Carlos, California**

## 1   OVERVIEW

In this chapter we delve into the relationship between information security (sometimes called *computer security*) and usability design. The goals of *information security* are to protect the confidentiality, integrity, and availability of systems, information, applications, and network devices as well as to prevent repudiation (untruthful denial) of electronically based transactions. Security-related breaches manifest themselves in a variety of forms, including intrusions into systems, worm and virus infections, misuse, denial of service, integrity compromises in systems and/or data, scams, hoaxes, and many others. A taxonomy of the major security-related tasks that people must perform includes the following tasks: identification and authentication; assurance of integrity, confidentiality, availability, and system integrity; and intrusion detection. *Usability* flaws in a number of corresponding areas—password selection, third-party authentication, file access control, Web server configuration, firewall configuration, encryption of sensitive information, electronic commerce transactions, auditing and logging, and intrusion detection—are analyzed. These flaws are almost certainly also linked to the most costly form of security-related incident—damage and disruption to and/or theft of systems and data due to insiders. Employees and contractors who are disgruntled may, for example, be less motivated than other users to overcome usability hurdles in computing systems, something that may escalate damage and/or disruption considerably. Better default parameters in operating system and application software and the availability of settings that produce pervasive changes in the security of systems and applications, thus obviating the need to interact with systems many times or with more difficulty to tighten these settings, would go a long way in addressing the usability problems discussed in this chapter.

## 2   INTRODUCTION

Using computers is a way of life almost everywhere around the world. Without computers, life in virtually every country would be radically different. Although appreciating how different life would be without computers is easy, many people fail to appreciate what happens if computers are unreliable for a variety of reasons. In some cases, such as when computers are used to regulate energy flow within buildings, computer unreliability might not radically disrupt computing activity because people could in these circumstances simply take manual control of functions normally performed by computers. But in other cases, such as in air traffic control systems, the ramifications of computers becoming unreliable can potentially be considerably more draconian.

Computers become unreliable for a variety of well-known reasons: electrical failure, damage due to water or fire, and hardware and software flaws ("bugs") that threaten the normal operation of computers, to name a few. Strangely, until relatively recently, people have been largely unaware of the many security-related reasons for unreliability in computing systems, such as remote access by unauthorized users or the execution of malicious programs, even though security threats may be more costly and disruptive than others, such as power outages.

Protecting computers, the information residing in them and sent over networks, applications, and network components such as routers and switches against security-related threats falls within the purview of the field of information security, also known as computer security (Garfinkle et al., 2003). Included in the goals of information security are ensuring confidentiality of information; protecting the integrity of systems, data, and applications; ensuring that systems, data, and applications are available when needed; and ensuring that anyone who initiates an electronic transaction cannot repudiate or deny having done so afterward. Although once an obscure area, information security has become increasingly important over the last two decades as the number, magnitude, and impact of security breaches have grown. Available statistics show that:

- According to a recent Ponemon Institute (2009) survey, the average cost of an incident in which personally identifiable information (PII) has been compromised was $6,655,758 in 2009. This amount has grown every year since the Ponemon Institute started collecting statistics on this subject four years ago.
- According to a recent Computer Security Institute (CSI, 2009) Annual Computer Crime Survey, the average cost of a security breach in 2009 was $234,000.
- According to the Internet Crime Complaint Center (IC3, 2009), in 2009, 336,655 crime complaints were filed, a 22.3% increase over 2008. The vast majority of referred cases were related to fraud, which amounted to $559.7 million (up from $264.6 million in total reported losses in 2008).

Federal and state/provincial laws as well as regulations in the private sector require the implementation of information security measures to protect certain types of information such as PII that could be used to perpetrate identity theft. These laws and compliance regulations, which typically prescribe penalties such as fines for computer-related actions such as gaining unauthorized access to systems, have also greatly contributed to the growth of information security in numerous countries.

## 3 SECURITY BREACHES

A security incident is one in which an actual or possible adverse outcome due to a breach or bypass in a security mechanism has occurred (Schultz and Shumway, 2001). The nature of security breaches varies considerably. The most common types of security breaches include:

1. *Intrusions into Systems and Network Devices*. These are commonly known as *hacker attacks*. In most of these attacks, perpetrators break into user or system administrator's accounts using passwords captured ("sniffed") as they are entered during local log-ins or as they traverse networks during remote log-ins. *Brute-force attacks*, in which attackers run programs that

enter one password after another until one finally succeeds, are another variation of this type of attack. Alternatively, many attackers run programs that attempt to exploit vulnerabilities in systems and applications to gain unauthorized access. Once intruders break into a system, they often engage in activities such as reading users' files and email and planting Trojan horse programs that allow them once again to gain access to the victim system later.

2. *Malicious Code Infections*. Malicious code (also called "malware") is a program intended to subvert or bypass security functions built into systems and applications. Viruses are self-replicating programs that spread because of user actions, whereas worms are self-replicating programs that spread independently of user actions (Schultz and Shumway, 2001). The fact that worms work independent of users makes them particularly troublesome; from late 2008 through the present, for example, the Conficker worm and its variants have infected approximately 15 million PCs connected to the Internet (Schultz, 2009). Trojan horses are programs that are intended to be covert so that they are unlikely to be noticed and then eradicated. Trojan horse programs have grown disproportionately compared to viruses and worms, because computer crime is increasingly being perpetrated by desire for financial gain. To make money in this manner requires stealth—Trojan horses are stealthy, whereas viruses and worms are not (Schultz, 2006).

3. *Misuse and Subversion by Trusted Individuals, Such As Employees and Contractors*. Misuse and subversion are a less common but in many cases the most costly category of security breaches. These types of malfeasance are often due to motives such as greed or revenge.

4. *Denial-of-Service (DoS) Attacks*. Among the most common of all security breaches, these are intended to shut down or disrupt computing activities. According to the most recent CSI Computer Crime Survey, from 2008 to 2009 DoS attacks grew more than other types of attack—a growth rate of over 29% (CSI, 2009). They also are among the most costly of all because of organizations' dependence upon computing services. A particularly severe type of DoS attack is a distributed denial-of-service (DDoS) attack. In this type of attack malicious clients called bots are installed in systems throughout a network. When the "botmaster" sends the command, all bots in the botnet respond by sending volumes of malicious traffic that severely disrupts or brings down the network.

5. *Integrity Compromises*. Integrity compromises occur when perpetrators place malicious programs in systems they have accessed or modify

files within these systems. Web page deface-
ments are the most widely known type of
integrity compromise, although unauthorized
modification of system files and critical data
such as financial data are generally much more
costly than is typical.

6. *Social Engineering*. Social engineering means
"conning" someone to reveal information [e.g.,
a password or bank account number and
personal identification number (PIN)] desired
by a perpetrator. Many social engineering
attacks are perpetrated via email; others are via
phone calls to intended victims.

7. *Scams*. Scams are schemes in which email,
electronic messaging, websites, or chat rooms
are used to con unsuspecting users out of some-
thing (usually, money). *Phishing* (in reality
a type of social engineering attack) is cur-
rently the most common type of scam. One
kind of phishing attack involves a perpetra-
tor sending email that threatens people such
as bank customers with disruption of services
if recipients do not enter personal and/or finan-
cial information in a form on a Web page that
appears to belong to a legitimate company.
Other scams offer recipients of messages that
appear to come from people such as deposed
African political figures a large commission in
return for helping transfer what is described as
millions of dollars to the United States. The
catch is that recipients must first send a sum of
money to a designated address as a "measure
of good faith" before they allegedly received
any money.

8. *Hoaxes*. In hoaxes, bogus information is dis-
seminated electronically. For example, certain
network postings falsely claim that Win-
dows operating systems contain routines that
covertly glean data stored in these systems for
the National Security Agency (NSA).

9. *SQL Injection Attacks*. Relational database
management systems (RDBMSs) typically have
built-in controls to limit the potential unau-
thorized access to the information they store.
However, if database applications are not pro-
grammed in accordance with principles of
secure code development, perpetrators can sub-
mit specially crafted Structured Query Lan-
guage (SQL) statements to a database that cause
commands that retrieve database information
to be executed—an "SQL injection attack."
Perpetrators might also exploit SQL injection
vulnerabilities to compromise the database host
machine itself and use it as a "pivot point" to
break into other systems in the same network.

10. *Session Hijacking*. In a "session hijacking"
attack, perpetrators monitor traffic sent over
networks to obtain information such as Internet
Protocol (IP) addresses of clients and servers,
the state of the interaction between a Web
browser and a Web server, and so on. Using

this information, perpetrators may be able
to gain control of an existing connection or
create a new connection with exactly the same
characteristics as another one. With control over
the connection, the perpetrator has the same
privileges and access rights that the legitimate
user has.

11. *Spamming*. Spam is unwanted email ("junk
mail") and pop-up messages. Although sending
email does not in and of itself constitute a
security breach, the fact that the overwhelming
majority of spam has a sender address that has
been falsified does. Spam is thus, in effect, often
a type of repudiation attack. Furthermore, spam
now constitutes such a large proportion of the
email that users receive that organizations often
lose large amounts of money each year from lost
productivity (because each user must read and
then delete each spam message).

Despite the growing importance of information
security, system administrators and users often resist
using measures that improve security. User resistance
toward systems with which users must interact is a
well-known phenomenon (Turnage, 1990), as is the
fact that systems with poor user interaction methods
lead to greater user resistance than do other systems
(e.g., Markus, 1983; Al-Ghatani and King, 1999). User
resistance manifests itself in many ways. For example,
requirements for passwords that are so long and/or
complex can lead to users writing down their passwords,
something that may easily lead to computer account
compromises, as explained shortly.

Although usability design in systems is generally less
than optimal, poor usability design abounds in comput-
ing systems and devices designed to improve informa-
tion security, as explained shortly. The weaknesses in
this design may be "the straw that broke the camel's
back," in that measures used to raise security too often
create usability barriers that cause people to neglect or
abandon them, leaving their systems, applications, data,
and network devices vulnerable to all types of attacks.

## 4 TAXONOMY OF INFORMATION SECURITY TASKS

Analyzing the tasks that system administrators and users
must perform in securing systems, applications, and data
is a good starting point for examining usability issues in
information security. Schultz et al.'s (2001) taxonomy of
security tasks provides an analysis of six major security-
related tasks:

1. *Identification and Authentication*. *Identification*
means proving one's identity. *Authentication*,
very similar in meaning to identification, means
proving one's identity for the purpose of access-
ing a system or network. The most common
type of identification and authentication task
is entering a password, although many other
identification and authentication methods (such
as inserting a smart card and then entering a

short PIN) exist. Effective identification and authentication help prevent perpetrators from masquerading as other users.

2. *Protecting Data Integrity*. Although numerous data integrity protection methods exist, the most commonly used method is setting file system permissions to prevent all but very few users from being able to change, replace, or delete files and directories. Software for detecting changes in files and directories (often called *tripwire software*) is also becoming used more frequently. Data integrity protection methods help to prevent unauthorized deletion of and/or changes in data.

3. *Protecting Data Confidentiality*. Setting file permissions appropriately is also the most commonly used method of protecting data confidentiality. Another is encrypting data at rest (i.e., data stored on a system) as well as data in motion (e.g., data sent over a network). Controlling against privilege escalation in systems by drastically limiting the number of "superusers" (privileged users) is another data confidentiality assurance method because superusers can read every file on their system, regardless of what permissions have been set. Data confidentiality methods help prevent unauthorized disclosure and/or possession of information.

4. *Ensuring Data Availability*. Assuring that data and the applications that use them are available is another critical information security task. Tasks that help achieve this goal include making system and data backups as well as using other measures, such as fault-tolerant storage systems.[*] These tasks help in guarding against unauthorized deletion of or denying access to information and the programs that use them.

5. *Ensuring System Integrity*. The methods discussed previously that are used to protect data integrity are used in ensuring system integrity. Additionally, installing patches for vulnerabilities in systems and applications helps prevent unauthorized modification of system files. Inspecting system files for unauthorized changes is still another often-used method. These measures help to prevent unauthorized deletion and/or changes in system files.

6. *Accountability*. Accountability means being able to link user actions on computer systems, networks, and applications to individual users. One of the major methods of achieving accountability is inspecting the output of intrusion detection systems (IDSs). Intrusion detection means identifying attacks that have occurred and their outcomes (in terms of their success or lack thereof). Although intrusion detection

is not really a security countermeasure per se in that it does not help directly in preventing attacks, it is nevertheless a much used measure in that it enables technical experts to quickly identity and thwart attacks that are under way, thereby minimizing their impact. The most basic form of intrusion detection is inspecting system audit logs, a labor-intensive task at best. Intrusion detection is often performed by special software and hardware; even so, user interaction tasks are necessary.

Now that the basic kinds of tasks that must be performed in information security have been introduced, we'll explore the types of usability hurdles in information security and the impact they have.

## 5  FLAWS IN USABILITY DESIGN

As mentioned earlier, the area of information security abounds with examples of poor usability design. We will look at the usability design of password-based authentication, third-party authentication, file access control methods, Web configuration, firewall configuration, encryption of sensitive information, electronic commerce transactions, auditing and logging, and intrusion detection.

### 5.1  Password Selection and Memorability

Previous work by Proctor et al. (2000) demonstrates that entering a user name–password combination to log in to a system or network is not at all difficult from a human factors perspective. The fact that users become highly practiced over time in entering passwords when they see the appropriate prompt helps overcome the few usability hurdles that this task poses. A task analysis of generic password-based log-ins shows that users must engage in only a few relatively simple actions:

1. Visually sight the dialog box and the prompts and input field within (see Figure 1).
2. Use the pointing device to align the cursor/pointer to the correct location.
3. Home both hands at the keyboard.
4. Recall the password.
5. Enter the password by pressing the appropriate keystroke sequence.
6. Click on <OK> or press the <ENTER> key.

Although entry of a log-in name–password sequence is relatively easy for users, there is a more difficult human factors problem of which few users are aware. Hackers and password-cracking tools have become so efficient that passwords that users normally choose can be compromised in a very short time (Skoudis, 2004). Users can choose stronger (more difficult-to-guess) password, but doing so makes them more difficult to remember (Proctor et al., 2002). For example, "password" would be an easy-to-remember password, but it would be trivial to guess or crack. "6f*2Sl&," which has just as many characters as "password," would

---

[*] RAID, the redundant array of independent drives, is the most frequently used fault tolerance solution. RAID distributes data across multiple drives; if any drive fails, the data will thus be available on another.
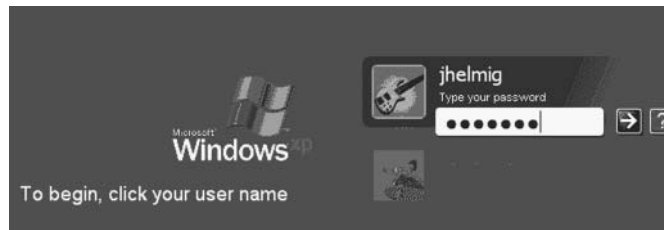
Figure 1    Log-on screen in Windows 7.

be much more difficult to guess or crack, but few users would be able to remember this password. Many users are not aware of what differentiates good passwords from ones that are easy to guess or crack; even if they know how to create good passwords, they nevertheless often choose weak ones (such as combinations of characters that include their user account names) because of the additional effort needed to create good passwords (Bishop and Klein, 1995). To compensate for the inability to remember more-difficult-to-remember passwords, users could write them down, but doing so would enable anyone who found the slips of paper or whatever else on which the passwords are written to use them to gain unauthorized access to the users' accounts.

Proctor et al. (2002) conducted studies testing the effects of proactive constraints, that is, limiting the types of passwords that users could create, and password length on the ability to crack passwords. As expected, longer passwords were more difficult to crack than were shorter ones. They found that proactive constraints produced different effects for shorter passwords than for longer passwords, however. When passwords were shorter (five characters in length), however, constraints elevated resistance to cracking more than when passwords were longer (eight characters in length). Constraints and password length together required the most effort on the part of users and produced only slightly better resistance to password cracking than did password length alone, suggesting that requiring longer passwords but not imposing additional constraints represents a good trade-off point between usability and security.

Yan et al. (2004) conducted a study designed to test the effectiveness of passphrase passwords versus randomly assigned passwords. Passphrases involve using the first characters of each word within well-known phrases, for example, the passphrase for "now is the time for all good men to come to the aid of their country" would be "nittfagmtcttaotc." Longer and more complex passwords are harder to crack, but they are also harder to remember. Passphrases are an attempt to solve the problem of users being unable to remember longer and more complex phrases. A control group was instructed to create a password of at least seven characters in length, of which one character could not be alphabetic. The number of passwords that were cracked using a combination of password cracking methods was significantly highest for the control group, but no significant difference between the randomly assigned and passphrase groups was found. However, users reported that passphrases were easier to remember than randomly

assigned passwords. User ratings indicated that users found passphrases significantly less difficult to deal with than random passwords but that passphrases and user-selected passwords were not significantly different in rated difficulty. This study provides evidence that security and usability may not necessarily be orthogonal to each other—passphrases represent a good balance between security and usability.

Vu et al. (2007) performed a set of experiments to determine how mnemonic methods could be used to enhance both the memorability and security of passwords. In one study, users were tasked with creating a sentence and combining the first letters of the words into a password. In one condition, participants were instructed to use only the first letters. In a second condition users were told to also include a digit and special character in the sentence to form a password. The additional constraint to include a digit and special character lowered password memorability but substantially decreased susceptibility to cracking. The lc5 password-cracking tool cracked 62% of passwords without a digit and special character but only 2% from the group that had additional constraints. In this and another, similar experiment, password recall was better with a one-week retention interval when recall was also tested after 5 min. However, having to immediately reenter a password after creating it, something that is commonly required, did not appear to improve long-term retention. Additionally, the results bring into question the effectiveness of passphrases. Only when a special character and digit were embedded in passphrases did passphrases produce better recall than in the condition in which participants created passwords without having to use passphrases.

The results of studies on password creation and memorability described so far repeatedly point to the fact that conventional passwords fall far short when the trade-off between password strength and usability/memorability is considered. Aware of the shortfalls of requiring users to create and recall strings of alphanumeric characters and symbols, researchers in recent years have also explored the possibility of users creating and using graphical-based passwords. Chiasson et al. (2007) conducted studies in which participants were required to create graphical "passwords" by sequentially clicking on points within graphical displays. After a delay, participants were then tested for their ability to authenticate by clicking on the same points within these displays that they had originally chosen. These researchers found that users were able to authenticate accurately and reasonably rapidly, although accuracy

and speed were somewhat faster in a laboratory setting as compared to a real-world setting. Additionally, participants who had to authenticate using multiple graphical "passwords" tended to experience interference that lowered their performance. Users also tended to rate graphical "passwords" favorably. These results correspond to those of an earlier study of the usability of graphical passwords by Wiedenbeck et al. (2005), in which participants interacted with systems under more limited conditions than in the Chiasson et al. study. Habibilashkari and Farmand (2009) conducted surveys on usability and security of graphical log-ons, the results of which demonstrate that these log-ons result in acceptable levels of usability *and* security.

These studies show the viability of graphical interaction tasks as an alternative to conventional passwords. Graphical "passwords" are stronger than conventional passwords in terms of resistance to cracking yet are considerably more usable. The long-established fact that recall (as is required for conventional passwords) is generally more difficult than recognition (as required for "graphical" passwords) (cf. Eagle and Leiter, 1964) goes far in explaining why graphical log-ons tend to be easier for users. Additionally, memory for images tends to be superior to memory for verbal materials (Shepard, 1967).

## 5.2  Third-Party Authentication Tasks

Many information security experts believe that passwords are now too dangerous to use and that other forms of authentication must supplant password-based authentication. Third-party authentication is any form of authentication that is not built into operating systems but that is, instead, provided through vendor products or methods. Password-based authentication is built into virtually every commercial off-the-shelf operating system, but smart card–based authentication, one of the most common types of third-party authentication, is not. Smart cards are actually miniature computers with chips that are built into an object such as a plastic card (Corcoran, 2000). To authenticate using smart cards, users must insert a smart card into a smart card reader, which is normally attached to a keyboard. Proctor et al. (2000) performed a task analysis on a typical smart card–based user interaction sequence. They found that to authenticate using smart cards users had to:

1. Visually sight a prompt on the display terminal that directs the user to proceed with the smart card authentication process.
2. Visually sight the smart card.
3. Use several fingers and the thumb to grasp the smart card.
4. Visually sight the smart card reader.
5. Use the hand and arm to move the smart card toward the smart card reader until it is in close proximity.
6. Rotate the hand until the smart card is at the proper angle to be inserted.
7. Insert the smart card until it fits inside the smart card reader.

8. Visually sight the display terminal (or listen for auditory feedback) for confirmation that the smart card was read successfully.
9. Grasp the smart card in several fingers and the thumb, moving the hand and finger away from the smart card reader.
10. Confirm that smart card data have been processed properly and are valid.
11. Use the hand to place the smart card on a surface, such as a table surface.
12. Let go of the smart card with the fingers and thumb.
13. Read a prompt that begins the "normal" log-in name–password entry sequence.
14. Home the hand on the keyboard.
15. Engage in the steps normally required for a password-based log-in but enter the PIN instead.

The task sequence for this generalized smart card–based authentication task sequence involves many steps that are not included in password-based log-ins, showing that the former type of authentication presents additional levels of difficulty for users. Although the exact tasks vary in different smart card implementations, one thing seems clear—users who are accustomed to password-based authentication are likely to resist smart card–based authentication because a greater amount of work and more opportunity for error are inherent in the latter. Given the far greater strength of smart card versus password-based authentication, this is unfortunate. This is just one of the many cases in information security in which security and usability are orthogonal to each other. Other forms of third-party authentication, such as biometric authentication, authentication based on fingerprints, retinal patterns, facial shape, and so on, are available, but do they require less work on the part of users? A task analysis on user interaction with a generic fingerprint-based biometric authentication device showed that this method involved substantially more steps than does password-based authentication but somewhat fewer steps than smart card–based authentication (Proctor et al., 2000). The usability limitations in third-party authentication thus are not limited to task sequences involving smart card–based authentication.

Cranor and Garfinkle (2005) conducted a study in which participants had to secure the contents of email using a smart card and two types of universal serial bus (USB) devices, one a "base token"(a simple kind of memory stick) and the other an "advanced token." The order of tasks was counterbalanced across participants. The researchers found that smart cards required approximately twice as long as the USB base or advanced tokens and that the error rate for smart cards was seven times higher than for any other condition. Participants also rated smart card interaction as the most unfavorable. The authors attributed the results to the fact that more than one hardware component was involved in smart card–based interaction, whereas a single piece of hardware was involved in the USB device–based interaction tasks.

Once again, research points to inadequate usability in user interaction in performing security-related tasks, in this case, third-party authentication. Third-party authentication methods are potentially advantageous in that they require little or no memory on the part of the user. Faulty design of interaction sequences, particularly the number of steps involved in third-party authentication tasks, is a yet-to-be corrected problem.

### 5.3   File Access Control

File access control mechanisms are used to protect information from not only unauthorized disclosure and possession but also unauthorized modification and deletion. Virtually every modern operating system includes a file system that offers permissions such read, write, and execute to control access to files and directories by users and/or groups. Write access is particularly potentially dangerous; it generally allows not only modification of content but also deletion of files and/or directories. Setting file and directory permissions is, however, far from optimal from a usability perspective. One of the problems is that an excessive number of user interaction task steps is often required. Consider, for example, the following interaction steps required for changing the permissions of a file or directory in Windows NT (Schultz et al., 2001). To change the permissions for a single user, the system administrator or owner of a file must:

1. Inspect the desktop visually to find the icon for Windows Explorer or visually sight the Windows and E keys on the keyboard.
2. Bring up Windows Explorer by double clicking on the desktop icon or pressing the Windows and E keys simultaneously.
3. Use a pointing device to scroll through groups of icons for files and directories until sighting the desired one.
4. Right click on the selected icon using the pointing device.
5. Click on Properties using the pointing device.
6. Visually scan the tabs at the top of the new screen that appears to find the Security tab.
7. Click on Security.
8. Of the options that appear, click on Permissions.
9. Click on Add.
10. Scroll through the list of groups and users.
11. Click on the Show Users box.
12. Scroll down to the desired user name.
13. Click on the user name to highlight.
14. Scroll through Type of Access.
15. Highlight the desired type of access (e.g., Full Control, Change, Read) and release the selection button on the pointing device.
16. Click on OK on three different screens.

Worse yet, the standard Windows NT file permissions are only the beginning when it comes to securing files and directories because they are in effect high-level permissions intended mainly for the convenience of system administrators and users. Many *individual* or *advanced permissions* are also available to allow for more granular file and directory access control. In Windows NT there are 7 such permissions: Read, Write, Execute, Delete, Change Permissions, Take Ownership, and Full Control. In Windows 2000 there are 14 such permissions: Traverse Folder/Execute File, List Folder/Read Data, Read Attributes, Read Extended Attributes, Create Files/Write Data, Create Folders/Append Data, Write Attributes, Write Extended Attributes, Delete Subfolders and Files, Delete, Read Permissions, Change Permissions, Take Ownership, and Synchronize. To add or delete any of these additional permissions, additional user interaction steps, almost none of which are intuitive to novice users, must be performed.

Various methods of changing file permissions without having to use a graphical user interface (GUI) also exist. These alternative methods may involve considerably fewer steps on the part of the user but require the entry of commands and flags in a syntax that is extremely unforgiving. For example, in Windows NT, 2000, XP, and Server 2003, someone must enter the following command to change file permissions on a file named Payroll in a folder named Finance for a user (named Brown in this example) from change (modify) to read only:

```
cacls  D:\Finance\Payroll  / E  / R  Brown:C  / G
   brown:R
```

In Linux and Unix systems users must enter the following command to remove Read and Write access from the group that owns the file "foo" in user Brown's home directory as well as from world (other):

```
chmod og-rw /home/Brown/foo
```

Experienced system administrators would have little trouble entering any of these commands to change file permissions, but only because of prolonged practice. Less experienced system administrators and users would experience extreme difficulty entering these commands correctly because of the nonintuitive nature of the command primitives and syntax.

### 5.4   Configuring Web Servers

The universal popularity of the World Wide Web (WWW) makes optimizing usability design in Web servers a necessity. Websites face enormous competition, to the point that user interaction with the majority of them must be satisfactory to users if users are going to be willing to visit (and, in particular, revisit) them. The main usability limitations in the Web arena thus instead involve webmasters' interactions with Web servers. Interaction sequences involved in setting up and administering Web servers tend to be extremely nonintuitive. For example, consider how to deny default access to directories in Apache Web servers (Schultz, 2002a). The webmaster must insert the following directives (commandlike objects) in the server's configuration:

<Directory />
Order Deny,Allow
Deny from all
</Directory>

The webmaster must next make exceptions by specially permitting access to the directories intended for Web access by users. For example, the following directives allow access to the/usr/users/*/public_html and/usr/local/httpd directories:

<Directory /usr/users/*/public_html>
Order Deny,Allow
Allow from all
</Directory>
<Directory /usr/local/httpd>
Order Deny,Allow
Allow from all
</Directory>

Although experienced webmasters can deal readily with Apache directories, novices are not so fortunate. Directives have a difficult syntax, one that once again illustrates the usability problems that plague information security. Ironically, Microsoft's Internet Information Server (IIS) Web server is considerably easier to use because of well-designed control panels that obviate the need for recalling complex syntactic conventions such as Apache directives. At the same time, however, IIS webmasters must often go through four to six levels of menus to get to a particular function even though menu depth produces a longer menu navigation time than does menu breadth (Schultz and Curran, 1986).

## 5.5 Configuring Firewalls

Firewalls are barriers between one network and another that are put in place to insulate one network from attacks from another (Cheswick et al., 2003). Implementing well-designed and well-maintained firewalls is one of the most important security measures that an organization can put in place. Firewalls vary considerably in their functionality; some do little more than block certain kinds of incoming packets bound for certain IP addresses and/or ports, whereas others analyze packets very thoroughly to determine whether or not they constitute desirable input to applications before sending them on via an entirely new connection that they create.

Regardless of the functionality, most commercially available firewalls have one thing in common: poor usability design. Consider, for example, the following access control entries in a Cisco PIX firewall:

#access-list acl_out permit tcp any any eq telnet

#access-list acl_out deny tcp any any

#access-list acl_out deny udp any any

#access-list acl_in permit tcp any host 128.13.23.9
eq ftp

#access-list acl_in permit tcp any host 128.13.23.9
eq
netbios-ssn

Each of these entries controls packet traffic in a unique manner. For example, the topmost entry says in effect that all telnet packets (i.e., packets sent in connection with the telnet service that allows one system to connect to another) are allowed to go outbound from the network in which PIX is placed, regardless of where they originated and where they are being sent. The second and third rules say that all other outbound traffic based on TCP (Transmission Control Protocol) and UDP (User Datagram Protocol) is blocked, again independently of the source or destination. The fourth rule says that all inbound FTP (File Transfer Protocol) packets bound for IP address 128.13.23.9 are allowed. The fifth and final rule says that all inbound NetBIOS packets, packets often used in connection with Windows network sessions, bound for the same system are also allowed.

In this case, one might expect that a very experienced firewall administrator would readily understand each of these rules, although this might not be true for a less experienced firewall administrator. Even if these assumptions are true, however, this might not make as much difference when it comes to making errors in firewall configurations as one might expect. Wool (2004) has conducted studies that have shown that firewall administrators often make errors in firewall rules that leave internal networks exposed to attacks that well-configured firewalls would block. Wool asserts that these errors are due to the fact that firewall interfaces deal with directionality of packets, that is, whether they are inbound or outbound, differently from how firewall administrators think about traffic flow through firewalls. Worse yet, some vendors fail to provide explanations of directionality in documentation provided with firewalls. This incongruity, according to Wool, results in poor usability that leaves firewall administrators confused and error prone when they configure firewalls.

The fact that the order of rules within an access control list (ACL) is extremely important in determining exactly how the rules work is another critical human factors consideration, as is the likelihood that the ACL will be extremely long, sometimes as much as 6000 (or even more) entries. Furthermore, several firewalls, many versions of PIX included, do not allow firewall administrators to edit the ACLs directly. Instead, they must add new rules that affect the existing ACL list. This makes the job of obtaining exactly the right rule set in the correct order even more difficult.

So far the usability analysis of firewalls has been limited to network firewalls, firewalls that filter traffic between one network and another. Another type of firewall is a "personal firewall," one designed to filter out individual systems from potentially malicious traffic that may be sent to them. Herzog and Shahmehri (2007) evaluated the usability and security of 13 different personal firewalls. Once again significant usability limitations were identified. The investigators concluded that providing better user guidance and making the design of the application behind personal firewalls more transparent to users would substantially improve usability.

**Figure 2**   PGP public key list display.

## 5.6   Encrypting Messages Sent over the Network

Anyone who expects the content of any message sent across a network or any information sent during a network session to be safe from unauthorized reading is badly deluded. Attackers often use hardware or software to capture the content of all packets going over the network, thereby enabling them to glean not only cleartext passwords but also potentially valuable information such as credit card numbers. *Encryption*, which means systematically scrambling the content of a cleartext message using a key and then applying a key to unscramble it (Schneier, 1998), provides a potentially very strong type of protection against unauthorized reading or possession of messages and network session content.

Despite the many advantages of encryption, the use of encryption by everyday users is rare, once again because of associated usability problems. Whitten and Tygar (1999) showed how deficient the usability design of a well-known encryption program, PGP (Pretty Good Privacy), is. PGP can be used to encrypt the content of messages and files sent across the network. The fact that free versions of this software exist and that PGP can run on a variety of operating systems makes it a good candidate for widespread use. To use this tool to send an encrypted message to another user, users of Windows systems must double click on an icon (a gray PGP padlock) on the desktop and choose a menu selection called Encrypt to encrypt the contents of a message (or possibly the contents of the clipboard). Once the user decides which to encrypt, the screen shown in Figure 2 appears.

If the user wants to encrypt a message sent to someone such as someone in Figure 2, the user must first scroll to that user's name and then confirm that the person has the same type of encryption,such as Diffie-Hellman/Digital Signature Standard (DH/DSS) public key encryption, that the originator of the message has. Very few users understand enough about encryption to make this decision. The user must next double click on the other user's name, making the name and key information appear in a window below the list of names.

If the user clicks only once, the error dialog box shown in Figure 3 appears.

The error message provides no meaningful feedback to users. Once the user double clicks on the name of another user, an additional set of nonintuitive choices appears related to whether the user wants a secure viewer and/or conventional encryption. If the user checks both of the radio buttons representing these choices, another error message that provides no meaningful feedback appears. If the user manages to click on the correct radio buttons, the message content looks like the one shown in Figure 4. The user can (finally) send the message. The fact that even experienced PGP users often have trouble using PGP and that adding, deleting, and generating keys needed for encryption and decryption involve additional long and conceptually different interaction sequences attests further to the usability problems inherent in the use of PGP.

Not all encryption is as difficult to use as is PGP. Users of Windows 2000 and XP systems can, for example, encrypt the contents of any file of which they are the owner by performing the following interaction steps:

1.   Bring up Windows Explorer.
2.   Sight the icon for the to-be-encrypted file.
3.   Right click on the icon to Properties.
4.   Click on Advanced.
5.   Click on Encrypt Contents to Secure Data.
6.   Click on Apply.
7.   Click on Encrypt the File Only.
8.   Click on OK twice.



**Figure 3**   PGP error dialog box.

```
-----BEGIN PGP MESSAGE-----
Version: PGPfreeware 6.5.8 for non-commercial use <http://www.pgp.com>

qANQR1DBwE4DH9zEXh/Jxd0QA/0WR5pJGWQ8Nao5mp88JroWivxGe8/P4ruqBZqk
HfHCrp/G5Z5kEFB4OEQTkqcGhX5/7/tNnfPuEiTj83DrYh7y6tEkMjF6Wjqt0p6I
RcyEMpHqVeuarfPaYk1hIoB26g8p8eqt2XURXnoYw/Kp0nj/PLEO1eYMQlkb4jht
xczSIAP/cABA8Vz3ggje/DeKQnNMz8UojVhxsa8H2X6a0aWLIn480/eDK5DySEVG
Qf0cE/anygdtlkrCYmcXcGPTopqvPhuI5BR2hLCXQG1KU5W/6z5Wjmavl5hDfueM
1Jm0895IGy7qCeySIKtBGCqGdAi4jsKh9I7NgGxzFEu8zDmbr7JwLDllf1MC8mr
YrYzODfrG5SO4x9ZkPnDyE2hmYdbgMSgTZBacubwy2PgGJqPCgwv58+tnJAF0c4h
UYTp2DR7tBZNdP78FV3uM7BCUEVYV6RhJE6zauyjd3nlBp4kTNZxEN2MO6Hn//j1
Qoakp2xCa1pRGeBNh5ralRI1Xf3mLFjHeUSIdNx3IgPsBE5PcX7bO+yrKUlLtMCD
xNpFuEpr+qqQKYG11AMNstX5bSUw9rpaQmwk/8M2nv9FIttuRLXGpEgSghCojRZQ
QLj5KT9jLVXxeZ1sBlntC6hDvXMrWKoDxVrVfp+9EoxXLUIFzBJXk++CUPhfGIrw
1ndU1NSiS95BTllZ5Lwa70XMwZ+Q1ZdEHTYpPogXkYRjCC/3R6v6j/kV8iBlnCK5
bpnlXI5ghzWA5gP6pBDAMr4MJ4WRHjpJDQIAFcgJh3713xdy3oeA9BNe4E3kxLSg
9paoEes1gClA9s8oMgLlqwEnySriAZnmxA===
=GJxi
-----END PGP MESSAGE-----
```

**Figure 4**   PGP-encrypted message.

At this point all might seem well as far as the user goes, but there is a serious problem of which most users are unlikely to be aware. Although the file will be encrypted transparently whenever the user closes it and it will also be decrypted transparently whenever the user opens it, if something happens to the user's key—if it should be deleted or corrupted—the contents of the file will be unrecoverable because they cannot be decrypted. As a precaution, the user (or the system administrator acting on behalf of the user) needs to engage in additional tasks related to making an "escrow" key for file recovery purposes. Doing so is not a trivial task from a human–computer interaction standpoint, another case in point for the conclusion that information security and usability requirements are often opposed to each other.

### 5.7   Electronic Commerce Transactions

Electronic commerce transactions require secrecy, integrity, and nonrepudibility more than anything else. Many ways of achieving secrecy exist, but secrecy alone is not enough in many such transactions. Several corporations created a special protocol, the Secure Electronic Transaction (SET) protocol, to address all three needs at once by encrypting all traffic generated in connection with transactions, using strong user authentication, confirming credit card numbers, and approving each transaction. SET does not merely encrypt network traffic, however; it also keeps personal information obtained from merchants as well as the specific types of purchases made from financial institutions that process the transactions.

Despite the inherent goodness of SET from an information security perspective, SET's usability design weaknesses have made it an extremely unpopular protocol with those who use it. As Schultz (2011) has pointed out, to start a SET transaction, each customer must request and then fill in entries in an electronic wallet or digital certificate, a kind of electronic credit card that contains information about a customer and

that customer's credentials. The issuing institution (normally, a bank or credit card company) provides copies of certificates issued to third-party merchants. These certificates contain the public keys[*] of both the merchant and the issuing institution. The customer initiates a transaction, causing the customer's Web browser to receive and validate the merchant's certificate. The browser uses the merchant's public key to encrypt a message related to the transaction and the issuing institution's public key to encrypt the payment information. This information, as well as information that uniquely links payment to this particular transaction, is sent to the issuing institution and the merchant. The merchant confirms the identity of the customer by verifying the customer's digital signature[*] contained within the customer's certificate. Next, the merchant transmits the order message to the issuing institution. The order message contains the issuing institution's public key, customer payment information, and merchant's certificate. The issuing institution confirms the identity of the merchant and verifies the message itself. The issuing institution verifies the payment portion of the message and then digitally signs and sends authorization back to the merchant, who can then supply the goods or services specified in the customer's order.

If you are confused at this point, you can readily understand how those who use SET often feel. Many of the major steps in a SET transaction can be broken down into multiple individual user interaction tasks. Many of the steps involving the merchant and issuing institution are automated, however, so it is the customer

---

[*] A public key is one half of a public–private key pair in which encryption of data is performed using one key and decryption is performed using the other. This type of encryption is often called "public key encryption." A digital signature is a public key–based cryptographic method used to uniquely identify each person using public key encryption as well as other cryptographic methods. Digital signatures help protect against repudiation.

who is faced with the majority of the human–computer interaction tasks, many of which are not very intuitive. SET shows once again that information security and usability design are very often orthogonal to each other.

SSL (secure sockets layer) encryption is the most frequently used kind of encryption in electronic commerce (as well as in Web-based transactions in general). Using SSL should not in theory be difficult, because the number of user interaction steps involved is small. Analysis and research show otherwise, however. For example, Mannan and van Oorschot (2007) observed computer-savvy users perform electronic banking-related tasks, many of which involved the use of SSL. The researchers found that many users did not understand the purpose of SSL and the certificates that it requires for mutual authentication between the Web browser and Web server. They also found that users often did not notice the padlock icon that appears in the lower right corner in most browsers whenever SSL encryption is in place, most likely because this icon is typically very small and thus easy to overlook. Hol (2008) also points out that when this icon is displayed, it provides no information about the identity of the Web server to which the browser is connected. Furthermore, perpetrators may create fake Web pages with fake padlocks completely unbeknownst to users. Finally, the warning that is displayed to users when an unknown or untrusted SSL certificate is being evaluated (see Figure 5 below) or another potentially unsafe condition could occur is difficult for users to understand. Consequently, they are likely to simply click on "Allow" rather than halting their electronic commerce transaction to determine what

might be wrong at that point. And, according to Hol, even if they take the time to examine the content of a certificate, they are unlikely to understand what it means.

Worse yet, users typically do not know what "dangerous" and "safe" Web content is, what a malicious cookie is, or what aspects of a particular Extensible Markup Language (XML) or Active X object are "unsecure" or "secure" (Schultz, 2011). They thus might go so far as to turn off all Web-related warnings. If they do not do so, they are at least likely to have an implicit trust for events that happen in connection with Web servers and thus keep clicking "Yes," "Allow," and similar options when warnings are displayed.

## 5.8 Auditing and Logging

Auditing and logging in operating systems enable system administrators to examine what users and applications have done, something that can lead to corrective action such as disabling accounts of malicious users as well as modifying information security policy. Accessing audit logs is not generally very difficult in most operating systems. In Unix and Linux, for example, the root user (the superuser) needs only to enter the who and last commands to discover who is currently logged in and the log-in and log-out times of each user, respectively. A major exception is in Novell NetWare. To view Netware audit reports for volume auditing, the system administrator must enter AUDITCON and then select the Change current server to choose the appropriate service. From the AUDITCON main menu, the system administrator must select Change current volume to designate the volume of interest and then
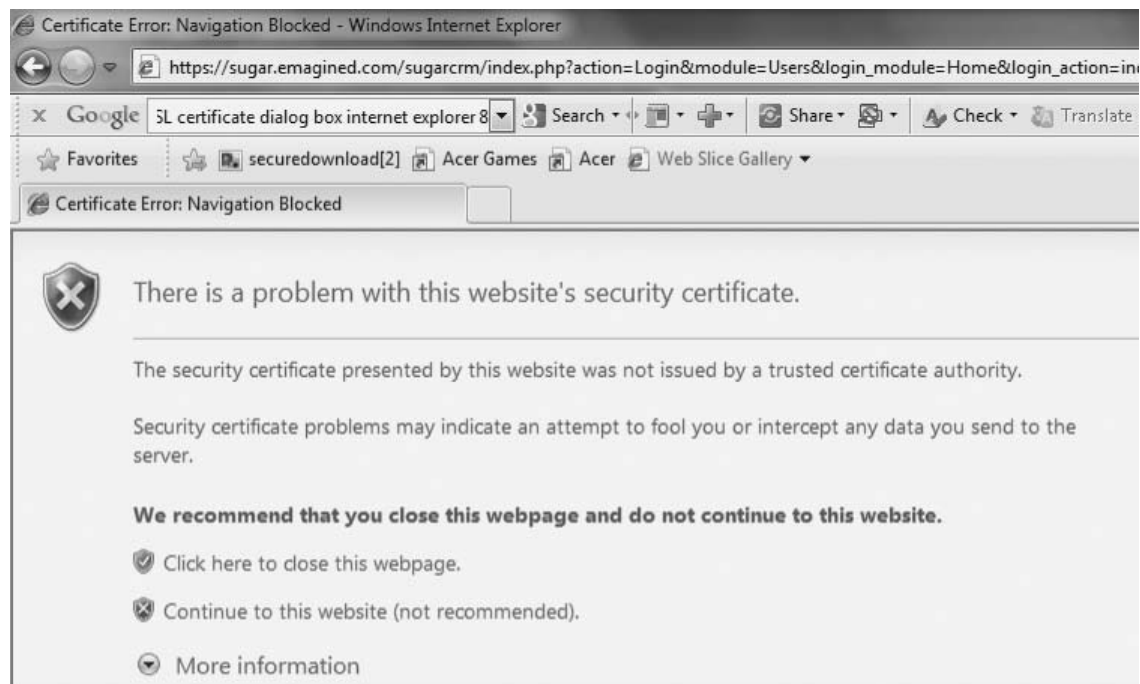


**Figure 5**  An Internet Explorer warning.

select Auditor volume log-in from the AUDITCON main menu. The Enter volume password input box will appear. The system administrator must enter the auditing password for the chosen volume and then press <ENTER> and then select "Auditing reports," a selection within the Available audit options menu that will be displayed. The difficulty of performing these tasks is self-explanatory.

The generally more difficult part from a human–computer interaction standpoint is configuring logging. Unix and Linux system logging (syslog) is controlled by configuring a file, /etc/syslog.conf. There are eight priorities of logging: emerg (highest), alert, crit, err, warning, notice, info, debug (lowest). There are also seven types of logable messages, each concerning a different part of or function within the system: kernel, user, mail, daemon, auth, lpr, and local. The syslog messages can be sent to one or more of the following: the system console, a central log server, and/or to a file within the system in which syslog has been enabled.

The entries in an /etc/syslog.conf file in a Unix system are as follows:

```
*.crit;kern.debug;auth.info      /dev/console
*.alert;user.notice              root
auth.debug                       /var/adm/authlog
mail.notice                      /var/adm/maillog
```

The first line specifies that any type of event that has a priority of critical or higher[*] will be sent to the terminal (console) on which the event occurred. Any kernel-related event with a priority of debug or higher as well as any authentication-related event with a priority of information or higher will also be sent to the terminal. The second line entries cause any event with a priority of alert or higher to be sent to the root (superuser) account in the system in which this event has occurred; additionally, any user-related event with a priority of notice or higher will also be sent to root. The third line specifies that any authentication-related event will be sent to a local file, /var/adm/authlog. The fourth line causes all mail-related events with a priority of notice or higher to be sent to the mail log, /var/adm/maillog. Regardless of any specific entry, there is nothing in the format of the /etc/syslog.conf file that makes configuring system logging straightforward.

The built-in graphical user interface in Windows 2000, XP, Vista, Windows 7, and Windows Server 2003 and 2008 makes configuring security logging (auditing) in these systems somewhat more intuitive. Nevertheless, the number of user interaction steps involved is still excessive. On a Windows XP Professional workstation, for example, one must:

1. Go from Start to Control Panel.
2. Double click on Administrative Tools.
3. Double click on Local Security Policy.

4. Enumerate the Security Settings container.
5. Enumerate the Local Policies container.
6. Click on Audit Policy.
7. For each setting or change desired, double click on the name of the audit category (Audit account log-on events, Audit policy changes, Audit privilege use, and so on).
8. Click on Success and/or Failure.
9. Click on Apply.
10. Click on OK.
11. Repeat steps 7–10 for each additional audit category.

Figure 6 shows the Audit Policy configuration screen. Even though configuring Security Logging in Windows systems is more intuitive than in Unix and Linux, the number of interaction steps, especially when multiple audit categories must be selected, is fairly large.

## 5.9 Intrusion Detection Monitoring

Intrusion detection means identifying security breaches that occur. Intrusion detection has become an important component of most organizations' information security program. Intrusion detection enables technical staff to identify and respond readily to incidents, thereby minimizing the amount of financial and other types of loss (Endorf et al., 2004).

IDSs automate the process of detecting intrusions, thereby increasing proficiency and reducing the number of personnel needed.

Although most IDSs are not all that difficult to configure, reading the output of these systems can be quite a challenge. Consider, for example, the following output from Snort, the most widely used IDS today (Caswell and Foster, 2003):

```
[**] SCAN-SYN FIN [**]
11/02-16:01:36.792199 109.10.0.1:21 ->
    16.16.90.1:21
TCP TTL:24 TOS: 0x0 ID:39426
**SF**** Seq: 0x27896E4 Ack: 0xB35C4BD Win:
    0x404
```

Even a proficient technical person would have difficulty understanding what this output means without careful study of Snort documentation. Following is another example of Snort output:

```
[**] [1:1959:1] RPC portmap request NFS UDP [**]
[Classification: Decode of an RPC Query] [Pri-
    ority: 2]
08/14-04:12:43.991442 109.10.0.1:46637 ->
16.16.90.1:111
UDP  TTL:250  TOS:0x0  ID:38580  IpLen:20
    DgmLen:84 DF
Len: 56
```

Snort is only one of many IDSs. Bro (see ftp://ftp.ee.lbl.gov/.vp-bro-pub-0.7a90.tar.gz), another IDS,

---

[*] This is not necessarily true, however. In some flavors of Unix, choosing a priority of debug results in only debug-related events being logged.
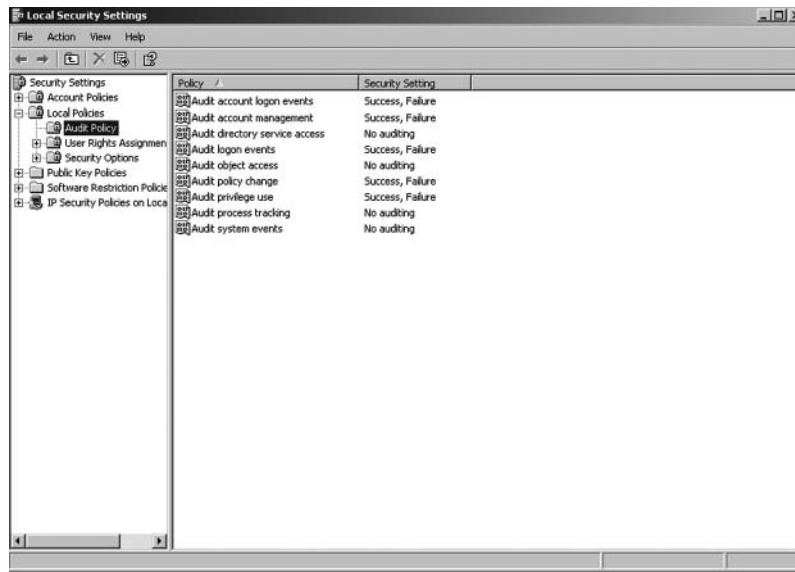
**Figure 6** Audit Policy screen in Windows XP.

also illustrates the problem of poor usability design in information security with output such as the following:

Nov 16 03:08:39 AddressDropped dropping address spock.bcc.com.pl (ftp)

Nov 16 03:15:01 WeirdActivity 218.73.102.106/ 1039

> nsx/dns: repeated_SYN_with_ack

Nov 16 03:31:23 AddressDropped low port trolling a213-22-132-227.netcabo.pt 258/tcp

Nov 16 04:50:44 AddressDropped dropping address 12.31.179.246 (4000/tcp)

Nov 16 06:25:23 SensitivePortmapperAccess rpc: cs4/917 > guacamole.cchem.berkeley.edu/portmap pm_dump: (done)

Nov 16 06:30:48 SensitivePortmapperAccess rpc: jackal.icir.org/1721 > arg/portmap pm_dump: (nil)

Nov 16 06:30:49 AddressScan 66.243.211.244 has scanned 10000 hosts (445/tcp)

Nov 16 06:30:50 PortScan 218.204.91.85 has scanned 50 ports of siblys.dhcp

Nov 16 06:30:50 AddressDropped dropping address 216.101.181.5 (4000/tcp)

Nov 16 06:30:50 SensitiveConnection hot: neutrino 200b > 147.8.137.149/telnet 463b 14.2s "root"

Nov 16 06:30:50 WeirdActivity p508c7fc5.dip.t-dialin.net -> 131.243.3.162: excessively_large_fragment

Nov 16 06:30:50 SensitiveConnection hot: p508d918a.dip.t-dialin.net 0b }2 muaddib/IRC ?b

0.6s inbound IRC

Nov 16 06:30:50 OutboundTFTP outbound TFTP: sip000d28083467.dhcp -> inoc-dba.pch.net

Nov 16 06:30:52 SensitiveConnection hot: 198.128.27.21 560b > 208.254.3.160/https 4202b

0.5s <IRC source sites>

Nov 16 06:30:53 WormPhoneHome worm phone-home

signature mcr-88-4 -> 218.146.108.51/9900

The output of Bro almost seems to be designed to be confusing to everyone but people who thoroughly understand how the system works. Bro's user interface is not at all atypical of today's IDSs.

Werlinger et al. (2008) conducted interviews with technical staff who used IDSs and also observed them as they actually used IDSs. The authors asserted that the sheer complexity of IDSs makes them difficult for even technically proficient users to use. Some of the areas Werlinger et al. singled out as in need of improvement from a usability perspective were:

- Installation—too complex
- Initial configuration—too complex and nonintuitive
- Lack of automated discovery of computers and devices on networks, something that necessitates considerable work on IDS analysts' part
- Reporting—not sufficiently flexible in providing information IDS analysts need
- Error messages—too terse and uninformative
- Lack of tools for adjusting alarm thresholds

As mentioned earlier, intrusion detection is in and of itself a difficult task, even for technically accomplished individuals. The usability problems in this area thus provide disproportionally difficult hurdles for IDS users.

## 6  IMPLICATIONS

Failure to pay sufficient attention to effective usability design in the information security arena has caused a plethora of dire consequences, the most apparent of which is failure to implement measures needed to defend systems and networks against attack. If too much effort, confusion, error, and/or frustration results from attempting to engage in security-related tasks, users will simply refrain from engaging in these tasks or will perform them inadequately. Consequently, stronger forms of authentication than password-based authentication will not be implemented, firewalls will not be inadequately configured, file system permissions will allow too much access by too many users, sensitive data and passwords will traverse networks in cleartext, operating system patches will never get installed, auditing will not be enabled or will be configured inadequately, intrusion detection data will be ignored, and so on.

Why does the world of information security repeatedly turn its proverbial back on the principles of effective human–computer interaction? The "Sherman M51 Tank" analogy may help in understanding what may be happening. Over a half century ago the Sherman M51 tank represented a major advance in warfare from the standpoint of the firepower it delivered, but poor human factors design greatly hampered tank crews' ability to operate it. The weaponry-related advantages apparently outweighed the human factors–related disadvantages, as judged by the many M51 tanks that were built and deployed. In information security the same kinds of trade-offs apply. When one considers the advantages of third-party authentication, such as smart card–based authentication compared to normal password-based authentication, information security professionals will readily endorse third-party authentication. At the same time, however, users and system administrators (and, in particular, managers) may not understand just how advantageous smart card–based authentication is compared to the usability disadvantages, leading them to favor conventional (password-based) authentication.

Additional negative consequences of failing to consider usability design also need to be considered. Vendors of products designed to improve security are in general not exactly reaping record profits. Vendors could considerably boost their sales by redesigning the user interfaces to their products.

Perhaps most important in the information security arena, however, is the potential relevance of the usability problems documented in this chapter to the insider threat. Security breaches instigated by insiders—employees, contractors, and consultants—account for far more financial and other forms of loss than any other source (Schultz, 2002b). User resistance to interaction tasks with poor usability design is well documented. This resistance

surfaces in a variety of ways, including passive behavior, verbal behavior, hesitance to continue in interacting with computers, loss of attention to tasks, and many others (Martinko et al., 1996). A study by the Information Security Forum (http://www.securityforum.org) in 2000 shows that inadequate security behavior of staff members rather than poor security measures per se account for as much as 80% of all security-related loss. Furthermore, even when staff members realize that security controls have been put in place with sound justification, they quickly reject controls that are ineffective, inefficient, and ambiguous (Leach, 2003). Poor usability design thus appears to be closely linked to insider attacks, internal misuse, and insider error that result in massive losses. Although the link between usability design and insider-related loss is currently indirect, empirical studies on these issues will in time provide more definitive results.

## 7  SOLUTIONS

Applying well-accepted principles of usability design in the information security arena is the most obvious solution to the problems presented in this chapter. Table 1 lists the problems presented in this chapter and possible usability solutions for each. A good high-level approach to an effective solution is to assume that most people who have security needs are not very aware of exactly what these needs are and what must be done to meet them. First and foremost, operating systems and applications need to have more secure settings right out of the box. Vendors typically use default settings that cause the least disruption to users rather than providing settings that raise security to at least an acceptable minimum. The unfortunate result is higher susceptibility to attacks. Allowing options in user interfaces that set security to a desired level without requiring that users know all the individual settings and what they mean is an excellent solution. A good example of how this can be done is the security options for Microsoft Internet Explorer (IE), a widely used Web browser, as shown in Figure 7.

A slide bar allows users to choose privacy levels varying from high to medium to low, thereby precluding the need to navigate to and choose settings from an excessive number of screens. In Figure 7 the user has chosen a level of security that is slightly below medium. The result is that third-party cookies (objects used to keep information about users in Web transactions) from websites that do not have a defined privacy policy will be blocked. The IE user interface in this example also is conducive to explorability—users can choose privacy levels without being locked in to a particular choice. Users may not fully understand the choices they explore and/or choose, but they will be better off than the way things are with the current user interfaces in the information security arena.

Table 1 summarizes the areas investigated in this chapter, usability problems associated with each area, and possible solutions. Information security has come a long way over the years. Unfortunately, the same cannot be said for usability design in systems used in protecting

**Table 1  Usability Problems in Information Security and Possible Solutions**

| Task | Usability Problems | Possible Solutions |
| --- | --- | --- |
| Password entry | Stronger passwords require more effort to create, are more difficult to remember. | Use an alternative form of authentication (e.g., graphical log-in) that does not require users to create and remember difficult-to-remember authentication credentials. |
| Third-party authentication | Task involves excessive number of user interaction steps. | Design more efficient interaction sequences. |
| Setting file access controls | The number and difficulty of user interaction steps are often overwhelming to users. | Design more efficient and simpler interaction sequences. |
| Configuring Web server security parameters | Syntax for changing parameters is often nonintuitive; interaction with menus may involve an excessive number of levels. | Syntax should be made more intuitive; menus should have greater breadth, not depth. |
| Configuring firewall security parameters | Syntax for changing parameters is often nonintuitive. | Syntax should be made more intuitive. |
| Encryption | User interaction may involve nonintuitive steps; may involve an excessive number of steps. | Design more efficient and simpler user interaction steps. |
| Electronic commerce transactions | User interaction may involve nonintuitive steps; number of steps may be excessive; error and warning messages may be confusing. | Design more efficient and simpler user interaction steps; error messages should be made clearer and more informative. |
| Configuring auditing and logging | Syntax for setting auditing and logging parameters is often nonintuitive; number of steps may be excessive. | Syntax should be made more intuitive; design more efficient user interaction steps. |
| Intrusion detection | Output may be cryptic and poorly formatted, installation and configuration of IDSs is complex, defining systems and devices user intensive, reporting is inflexible, error messages are terse. | Output should be easier to interpret and should be formatted in a manner that facilitates recognition of important data; fewer installation steps, more intuitive configuration settings, autodiscovery of network entities, flexible reporting, more detailed and understandable error messages. |



**Figure 7**   Dialog box for setting privacy level in Windows Internet Explorer.

systems and detecting when systems become attacked. The good news is that more research on usability and security is being conducted and published every year. How to improve usability while at the same time having sufficient levels of security is becoming increasingly clear. But both usability and security need to be integrated *early* in the systems development life cycle if both are to be optimal. Having to retrofit either generally results in expenditure of more resources as well as less efficient mechanisms and functions. The time to start integrating both usability and security into systems is during the *requirements* phase, not later.

## REFERENCES

Al-Ghatani, S. S., and King, M. (1999), "Attitudes, Satisfaction and Usage: Factors Contributing to Each in the Acceptance of Information Technology," *Behaviour and Information Technology*, Vol. 18, pp. 277–297.

Bishop, M., and Klein, D. V. (1995), "Improving System Security via Proactive Password Checking," *Computers and Security*, Vol. 14, pp. 233–249.

Caswell, B., and Foster, J. C. (2003), *Snort 2.0 Intrusion Detection*, Syngress, Rockland, MA.

Cheswick, B., Bellovin, S., and Rubin, A. (2003), *Firewalls and Internet Security: Repelling the Wily Hacker*, 2nd ed., Addison-Wesley, Reading, MA.

Chiasson, S., Biddle, R., and van Oorschot, P. C. (2007), "A Second Look at the Usability of Click-Based Graphical Passwords," in *Proceedings of the 3rd Symposium of Usable Privacy and Security,* pp. 1–12.

Computer Security Institute (2009), "2009 CSI Computer Crime & Security Survey," available: http://www.hlncc .com/docs/CSI_Survey_2009_Executive_Summary.pdf.

Corcoran, D. (2000), "Security-Related Exposures and Solutions in Smartcards," *Information Security Bulletin*, Vol. 5, No. 9, pp. 13–22.

Cranor, L. F., and Garfinkle, S. (2005), *Security and Usability: Designing Secure Systems That People Can Use*, O'Reilly, Sebastopol, CA, pp. 229–239.

Eagle, M., and Leiter, E. (1964), "Recall and Recognition in Intentional and Incidental Learning," *Journal of Experimental Psychology*, Vol. 68, pp. 58–63.

Endorf, C., Schultz, E. E., and Mellander, J. (2004), *Intrusion Detection and Prevention*, McGraw-Hill, New York.

Garfinkle, S., Spafford, G., and Schwartz, A. (2003), *Practical Unix and Internet Security*, 3rd ed., O'Reilly & Associates, Sebastopol, CA.

Habibilashkari, A., and Farmand, S. (2009), "A Survey on Usability and Security Features in Graphical User Authentication Algorithms," *International Journal of Computer Science and Network Security*, Vol. 9 No. 9, p. 195.

Herzog, A., and Shahmehri, N. (2007), "Usability and Security of Personal Firewalls," in *Proceedings of IFIP Security Conference 2007*, pp. 37–48.

Hol, K. J. (2008), "Security and Usability: An Introduction," available: http://www.kjhole.com/WebSec/PDF/Security Usability.pdf, accessed October 20, 2008.

Internet Crime Complaint Center (IC3) (2009), "2009 Internet Crime Report," available: http://www.ic3.gov/media/ annualreport/2009_IC3Report.pdf.

Leach, J. (2003), "Improving User Security Behavior," *Computers and Security*, Vol. 22, No. 8, pp. 685–692.

Markus, M. L. (1983), "Power, Politics, and MIS Implementation," *Communications of the ACM*, Vol. 26, pp. 430–444.

Mannan, M., and van Oorschot, P. C. (2007), "Security and Usability: The Gap in Real-World Online Banking," paper presented at the New Security Paradigms Workshop (NSPW), North Conway, NH, September 18–21 2007, pp. 1–14.

Martinko, M. J., Henry, J. W., and Zmud, R. W. (1996), "An Attributional Explanation of Individual Resistance to the Introduction of Information Technologies in the Workplace," *Behaviour and Information Technology*, Vol. 15, pp. 313–330.

Ponemon Institute (2009), "Fourth Annual US Cost of Data Breach Study," available: http://www.ponemon.org/ local/upload/fckjail/generalcontent/18/file/2008-2009%20 US%20Cost%20of%20Data%20Breach%20Report%20 Final.pdf.

Proctor, R. W., Lien, M. C., Salvendy, G., and Schultz, E. E. (2000), "A Task Analysis of Usability in Third-Party Authentication," *Information Security Bulletin*, Vol. 5, pp. 49–56.

Proctor, R. W., Lien, M. C., Vu, K.-P. L, Schultz, E. E., and Salvendy, G. (2002), "Improving Computer Security for Authentication of Users: Influence of Proactive Password Restrictions," *Behavior Research Methods, Instruments and Computers*, Vol. 34, pp. 163–169.

Schneier, B. (1998), *Applied Cryptography*, 2nd ed., Wiley, New York.

Schultz, E. E. (2002a, December), "Guidelines for Securing Apache Web Servers," *Network Security*, pp. 8–14.

Schultz, E. E. (2002b), "A Framework for Understanding and Predicting Insider Attacks," *Computers and Security*, Vol. 21, No. 6, pp. 526–531.

Schultz, E. E. (2011), "Web Security, Privacy,and Usability," in *Human Factors in Web Design*, 2nd ed., K.-P. L Vu and Robert Proctor, Eds., CRC Press, Boca Raton, FL, pp. 663–676.

Schultz, E. E. (2006), "Where Have Viruses and Worms Gone? New Trends in Malware," *Computer Fraud and Security*, No. 7, pp. 4–8.

Schultz, E. E. (2009), "Global Threat Update," State of California CISO Lecture Series, available: http://www.cio.ca .gov/OIS/Government/events/documents/Global_Threats .pdf, accessed September 30, 2009.

Schultz, E. E., and Curran, P. S. (1986), "Menu Structure and Ordering of Menu Selections: Independent or Interactive Effects?" *SIGCHI Bulletin*, Vol. 18, pp. 69–71.

Schultz, E. E., and Shumway, R. (2001), *Incident Response: A Strategic Guide for Handling Security Incidents*, New Riders, Indianapolis, IN.

Schultz, E. E., Proctor, R. W., Lien, M. C., and Salvendy, G. (2001), "Usability and Security: An Appraisal of Usability Issues in Information Security Methods," *Computers and Security*, Vol. 20, pp. 620–634.

Shepard, R. N., "Recognition Memory for Words, Sentences and Pictures, *Journal of Verbal Learning and Verbal Behavior*, Vol. 6, pp. 156–163.

Skoudis, E. (2004), *Malware: Fighting Malicious Code*, Prentice Hall, Upper Saddle River, NJ.

Turnage, J. J. (1990), "The Challenge of New Workplace Technology for Psychology," *American Psychologist*, Vol. 45, pp. 171−178.

Vu, K.-P., Proctor, R. W., Bhargav-Spantzel, A., Tai, B.-K., Cook, J., and Schultz, E. E. (2007), "Improving Password Security and Memorability to Protect Personal and Organizational Information," *International Journal of Human-Computer Studies*, Vol. 65, pp. 744−757.

Werlinger, R., Hawkey, K., Muldner, K., Jaferian, P., and Beznosov, K. (2008), "The Challenges of Using an Intrusion Detection System: Is It Worth the Effort?" in *Proceedings of the 4th Symposium on Usability in Privacy and Security*, pp 107−118.

Whitten, A., and Tygar, J. D. (1999), "Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0," in *Proceedings of the 8th USENIX Security Symposium*, Usenix Association, Berkeley, CA.

Wiedenbeck, S., Waters, J., Birget, J. C., Brodskiy, A., and Memon, N. (2005), "Authentication Using Graphical Passwords: Effects of Tolerance and Image Choice," in *Proceedings of Symposium on Usable Privacy and Security 2005*, pp. 1−12.

Wool, A. (2004), "A Quantitative Study of Firewall Configuration Errors," *IEEE Computer*, Vol. 37, No. 6, pp. 62−67.

Yan, J., Blackwell, A., Anderson, R., and Grant, A. (2004). "Password Memorability and Security: Empirical Results," *IEEE Security & Privacy*, Vol. 2, No. 5, pp. 25−31.

# CHAPTER 46

# USABILITY TESTING

**James R. Lewis**
**Software Group**
**International Business Machines Corporation**
**Boca Raton, Florida**

## 1 INTRODUCTION

Usability testing is an essential skill for usability practitioners—professionals whose primary goal is to provide guidance to product developers for the purpose of improving the ease of use of their products. It is by no means the *only* skill with which usability practitioners must have proficiency (Uldall-Espersen et al., 2008), but it is an important one. Surveys of experienced usability practitioners consistently reveal the importance of usability testing (Mao et al., 2005; Vredenburg et al., 2002).

One goal of this chapter is to provide an introduction to the practice of usability testing. This includes some discussion of the concept of usability and the history of usability testing, various goals of usability testing, and running usability tests. A second goal is to cover important statistical topics for usability testing, such as sample size estimation for usability tests, computation of confidence intervals, and the use of standardized usability questionnaires.

## 2 THE BASICS

### 2.1 What Is Usability?

The term *usability* came into general use in the early 1980s. Related terms from that time were *user friendliness* and *ease of use*, which *usability* (sometimes spelled *useability*) has since displaced in professional and technical writing on the topic (Bevan et al., 1991). Well before the 1980s, a refrigerator advertisement from March 8, 1936 cited usability as a feature (S. Isensee,

personal communication, January 17, 2010, see http://tinyurl.com/yjn3caa). The earliest scientific publication (of which I am aware) to include the word *usability* in its title was Bennett (1979).

It is the nature of language that words come into use with fluid definitions. Ten years after the first scientific use of the term *usability*, Shackel (1990, p. 31) wrote, "one of the most important issues is that there is, as yet, no generally agreed definition of usability and its measurement." Eight years later, Gray and Salzman (1998, p. 242) stated: "Attempts to derive a clear and crisp definition of usability can be aptly compared to attempts to nail a blob of Jell-O to the wall." Twenty years after Shackel, according to Alonso-Ríos et al. (2010, p. 53), "A major obstacle to the implantation of User-Centered Design in the real world is the fact that no precise definition of the concept of usability exists that is widely accepted and applied in practice."

There are several reasons why it has been so difficult to define usability. Usability is not a property of a person or thing. There is no thermometer-like instrument that can provide an absolute measurement of the usability of a product (Dumas, 2003; Hertzum, 2010; Hornbæk, 2006). Usability is an emergent property that depends on the interactions among users, products, tasks, and environments.

Introducing a theme that will reappear in several parts of this chapter, there are two major conceptions of usability. These dual conceptions have contributed to the difficulty of achieving a single agreed-upon definition. One conception is that the primary focus of usability should be on measurements related to the

accomplishment of global task goals (summative, or measurement-based, evaluation). The other conception is that practitioners should focus on the detection and elimination of usability problems (formative, or diagnostic, evaluation).

The first (summative) conception has led to a variety of similar definitions of usability, some embodied in current standards (which, to date, have emphasized summative evaluation). For example (Bevan et al., 1991, p. 652):

> The current MUSiC definition of usability is: the ease of use and acceptability of a system or product for a particular class of users carrying out specific tasks in a specific environment; where "ease of use" affects user performance and satisfaction, and "acceptability" affects whether or not the product is used.

Usability is the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [International Organization for Standardization (ISO), 1998, p. 2; American National Standards Institute (ANSI), 2001, p. 3]. As defined in ISO 9126-1, usability is one of several software characteristics that contribute to quality in use (in addition to functionality, reliability, efficiency, maintainability, and portability), and Bevan (2009) has recommended including flexibility and safety along with traditional summative conceptions of usability in a more complete quality-of-use model. The quality in use integrated measurement (QUIM) scheme of Seffah et al. (2006) includes 10 factors, 26 subfactors, and 127 specific metrics. Winter et al. (2008) proposed a two-dimensional model of usability that associates a large number of system properties with user activities. Alonso-Rios et al. (2010) published a preliminary taxonomy for the concept of usability that includes traditional and nontraditional elements, organized under the primary factors of Knowability, Operability, Efficiency, Robustness, Safety, and Subjective Satisfaction.

These attempts to provide a more comprehensive definition of usability have yet to undergo statistical testing to confirm their defined structures. An initial meta-analysis of correlations among prototypical summative prototypical usability metrics (effectiveness, efficiency, and satisfaction) that used published scientific studies from the human–computer interaction (HCI) literature found generally weak correlations among the different metrics (Hornbæk and Law, 2007). A replication using data from a large set of industrial usability studies, however, found strong correlations among prototypical usability metrics measured at the task level, with principal-components and factor analyses that provided statistical evidence for the underlying construct of usability with clear underlying objective (effectiveness, efficiency) and subjective (task-level satisfaction, test-level satisfaction) factors (Sauro and Lewis, 2009).

One of the earliest formative definitions of usability (ease of use) is from Chapanis (1981, p. 3):

> Although it is not easy to measure "ease of use," it is easy to measure difficulties that people have in using something. Difficulties and errors can be identified, classified, counted, and measured. So my premise is that ease of use is inversely proportional to the number and severity of difficulties people have in using software. There are, of course, other measures that have been used to assess ease of use, but I think the weight of the evidence will support the conclusion that these other dependent measures are correlated with the number and severity of difficulties.

Practitioners in industrial settings generally use both conceptualizations of usability during iterative design. Any iterative method must include a stopping rule to prevent infinite iterations. In the real world, resource constraints and deadlines can dictate the stopping rule (although this rule is valid only if there is a reasonable expectation that undiscovered problems will not lead to drastic consequences). In an ideal setting, the first conception of usability can act as a stopping rule for the second. Setting aside, for now, the question of where quantitative goals come from, the goals associated with the first conception of usability can define when to stop the iterative process of the discovery and resolution of usability problems. This combination is not a new concept. In one of the earliest published descriptions of iterative design, Al-Awar et al. (1981, p. 31) wrote: "Our methodology is strictly empirical. You write a program, test it on the target population, find out what's wrong with it, and revise it. The cycle of test–rewrite is repeated over and over until a satisfactory level of performance is reached. Revisions are based on the performance, that is, the difficulties typical users have in going through the program."

## 2.2 What Is Usability Testing?

Imagine the two following scenarios.

***Scenario 1*** Mr. Smith is sitting next to Mr. Jones, watching him work with a high-fidelity prototype of a Web browser for personal digital assistants (PDAs). Mr. Jones is the third person that Mr. Smith has watched performing these tasks with this version of the prototype. Mr. Smith is not constantly reminding Mr. Jones to talk while he works but is counting on his proximity to Mr. Jones to encourage verbal expressions when Mr. Jones encounters any difficulty in accomplishing his current task. Mr. Smith takes written notes whenever this happens and also takes notes whenever he observes Mr. Jones faltering in his use of the application (e.g., exploring menus in search of a desired function). Later that day he will use his notes to develop problem reports and, in consultation with the development team, will work on recommendations for product changes that should eliminate or reduce the impact of the reported problems. When a new version of the prototype is ready, he will resume testing.

***Scenario 2*** Dr. White is watching Mr. Adams work with a new version of a word-processing application. Mr. Adams is working alone in a test cell that looks almost exactly like an office, except for the large mirror

on one wall and the two video cameras overhead. He has access to a telephone and a number to call if he encounters a difficulty that he cannot overcome. If he places such a call, Dr. White will answer and provide help modeled on the types of help provided at the company's call centers. Dr. White can see Mr. Adams through the one-way glass as she coordinates the test. She has one assistant working the video cameras for maximum effectiveness and another who is taking time-stamped notes on a computer (coordinated with the video time stamps) as different members of the team notice and describe different aspects of Mr. Adams's task performance. Software monitors Mr. Adams's computer, recording all keystrokes and mouse movements. Later that day, Dr. White and her associates will put together a summary of the task performance measurements for the tested version of the application, noting where the performance measurements do not meet the test criteria. They will also create a prioritized list of problems and recommendations, along with video clips that illustrate key problems, for presentation to the development team at their weekly status meeting.

Both of these scenarios provide examples of usability testing. In scenario 1 the emphasis is completely on usability problem discovery and resolution (formative, or diagnostic, evaluation). In scenario 2 the primary emphasis is on task performance measurement (summative, or measurement-focused, evaluation), but there is also an effort to record and present usability problems to the product developers. Dr. White's team knows that they cannot determine if they've met the usability performance goals by examining a list of problems, but they also know that they cannot provide appropriate guidance to product development if they present only a list of global task measurements. The problems observed in the use of an application provide important clues for redesigning the product (Chapanis, 1981; Norman, 1983). Furthermore, as J. Karat (1997, p. 693) observed: "The identification of usability problems in a prototype user interface (UI) is not the end goal of any evaluation. The end goal is a redesigned system that meets the usability objectives set for the system such that users are able to achieve their goals and are satisfied with the product."

These scenarios also illustrate the defining properties of a usability test. During a usability test, one or more observers watch one or more participants perform specified tasks with the product in a specified test environment (compare this with the ISO/ANSI definition of usability presented earlier in this chapter). This is what makes usability testing different from other user-centered design (UCD) methods or marketing research (Dumas and Salzman, 2006). In interviews (including the group interview known as a focus group), participants do not perform worklike tasks. Usability inspection methods (such as expert evaluations and heuristic evaluations) also do not include the observation of users or potential users performing worklike tasks. The same is true of techniques such as surveys and card sorting. Field studies (including contextual inquiry) can involve the observation of users performing work-related tasks in target environments but restrict the control that

practitioners have over the target tasks and environments. Note that this is not necessarily a bad thing, but it is a defining difference between usability testing and field (ethnographic) studies.

This definition of usability testing permits a wide range of variation in technique (Wildman, 1995). Usability tests can be very informal (as in scenario 1) or very formal (as in scenario 2). The observer might sit next to the participant, watch through a one-way glass, or watch the on-screen behavior of a participant who is performing specified tasks at a location halfway around the world. Usability tests can be think-aloud (TA) tests, in which observers train participants to talk about what they're doing at each step of task completion and prompt participants to continue talking if they stop. Observers might watch one participant at a time or might watch participants work in pairs. Practitioners can apply usability testing to the evaluation of low-fidelity prototypes (MacKenzie and Read, 2007), high-fidelity prototypes, mixed-fidelity prototypes (McCurdy et al., 2006), Wizard of Oz (WOZ) prototypes (Dow et al., 2005; Kelley, 1985), products under development, predecessor products, or competitive products.

## 2.2.1 Where Did Usability Testing Come From?

The roots of usability testing lie firmly in the experimental methods of psychology (in particular, cognitive and applied psychology) and human factors engineering (Dumas and Salzman, 2006) with strong ties to the concept of iterative design. In a traditional experiment, the experimenter draws up a careful plan of study that includes the exact number of participants that the experimenter will expose to the different experimental treatments. The participants are members of the population to which the experimenter wants to generalize the results. The experimenter provides instructions and debriefs the participant, but at no time during a traditional experimental session does the experimenter interact with the participant (unless this interaction is part of the experimental treatment). The more formative (diagnostic, focused on problem discovery) the focus of a usability test, the less it is like a traditional experiment (although the requirements for sampling from a legitimate population of users, tasks, and environments still apply). Conversely, the more summative (focused on measurement) a usability test is, the more it should resemble the mechanics of a traditional experiment. Many of the principles of psychological experimentation that exist to protect experimenters from threats to reliability and validity (e.g., the control of demand characteristics, the Hawthorne effect) carry over into usability testing (Holleran, 1991; Macefield, 2007; Wenger and Spyridakis, 1989).

As far as I can tell, the earliest accounts of iterative usability testing applied to product design came from Alphonse Chapanis and his students (Al-Awar et al., 1981; Chapanis, 1981; Kelley, 1984) and had an almost immediate influence on product development practices at IBM (Kennedy, 1982; Lewis, 1982) and other companies, notably Xerox (Smith et al., 1982) and Apple (Williams, 1983). Shortly thereafter, John Gould and his associates at the IBM T. J. Watson Research Center

began publishing influential papers on usability testing and iterative design (Gould and Boies, 1983; Gould and Lewis, 1984; Gould et al., 1987; Gould, 1988), as did Whiteside et al. (1988) at DEC (Baecker, 2008; Dumas, 2007).

The driving force that separated iterative usability testing from the standard protocols of experimental psychology was the need to modify early product designs as rapidly as possible (as opposed to the scientific goal of developing and testing competing theoretical hypotheses). As Al-Awar et al. (1981, p. 33) reported: "Although this procedure [iterative usability test, redesign, and retest] may seem unsystematic and unstructured, our experience has been that there is a surprising amount of consistency in what subjects report. Difficulties are not random or whimsical. They do form patterns."

When difficulties of use become apparent during the early stages of iterative design, it is hard to justify continuing to ask test participants to perform the test tasks. There are ethical concerns with intentionally frustrating participants who are using a product with known flaws that the design team can and will correct. There are economic concerns with the time wasted by watching participants who are encountering and recovering from known error-producing situations. Furthermore, any delay in updating the product delays the potential discovery of problems associated with the update or problems whose discovery was blocked by the presence of the known flaws. For these reasons, the earlier you are in the design cycle, the more rapidly you should iterate the cycles of test and design.

### 2.2.2   Is Usability Testing Effective?

The widespread use of usability testing is evidence that practitioners believe that usability testing is effective. Unfortunately, there are fields in which practitioners' belief in the effectiveness of their methods does not appear to be warranted by those outside the field (e.g., the use of projective techniques such as the Rorschach test in psychotherapy) (Lilienfeld et al., 2000). In our own field, papers published since 1998 have questioned the reliability of usability problem discovery (Kessner et al., 2001; Molich et al., 1998, 2004; Molich and Dumas, 2008).

The common finding in these studies has been that observers (either individually or in teams across usability laboratories) who evaluated the same product produced markedly different sets of discovered problems. Molich et al. (1998) had four independent usability laboratories carry out inexpensive usability tests of a software application for new users. The four teams reported 141 different problems, with only one problem common among all four teams. Molich et al. (1998) attributed this inconsistency to variability in the approaches taken by the teams (task scenarios, level of problem reporting). Kessner et al. (2001) had six professional usability teams independently test an early prototype of a dialog box. None of the problems were detected by every team, and 18 problems were described by one team only. Molich et al. (2004) assessed the consistency of usability testing across nine independent organizations that evaluated the

same website. They documented considerable variability in methodologies, resources applied, and problems reported. The total number of reported problems was 310, with only 2 problems reported by six or more organizations, and 232 problems (61%) uniquely reported. The fourth comparative usability evaluation (CUE-4; Molich and Dumas, 2008) had a similar method and similar outcomes. "Our main conclusion is that our simple assumption that we are all doing the same and getting the same results in a usability test is plainly wrong" (Molich et al., 2004, p. 65).

This is important and disturbing research, but there is a clear need for more research in this area. A particularly important goal of future research should be to reconcile these studies with the documented reality of usability improvement achieved through iterative application of usability testing. For example, a limitation of research that stops with the comparison of problem lists is that it is not possible to assess the magnitude of the usability improvement (if any) that would result from product redesigns based on design recommendations derived from the problem lists (Hornbæk, 2010; Wixon, 2003). When comparing problem lists from many labs, one aberrant set of results can have an extreme effect on measurements of consistency across labs, and the more labs that are involved, the more likely this is to happen.

In the case of CUE-4 (Molich and Dumas, 2008), 17 professional usability teams evaluated the same website, with 9 teams conducting usability tests (5–15 participants per test) and 8 teams using expert review (1–2 reviewers per team). With one exception, the usability test teams used different sets of tasks for their evaluations. Across the 17 teams, there were 76 usability test participants and 10 expert reviewers, for a total of 86 individual experiences with the website. Using the binomial probability formula (see Section 3.1.2), it is possible to estimate the percentage of problems discovered with this sample size for problems of different likelihoods of discovery. For individual problems that would affect 10% of participants, the likelihood of having the problem turn up at least once in this study is about 99.99%, making their discovery virtually certain. For problems with a 1% probability of occurrence, the likelihood of discovery (at least once) with a sample size of 86 is about 58%, better than even odds. Even problems with probabilities of occurrence as low as 0.1% had about an 8% likelihood of discovery. It is not possible to know how many specific problems were available for discovery as a function of their probabilities of occurrence, but it seems reasonable that a mature website would have eliminated most high-probability problems, leaving a mass of less probable (hard-to-discover) problems, leading to little overlap in problem discovery across the teams. As Molich and Dumas (2008, p. 270) concluded, "The limited overlap could be interpreted as a sign that some of the teams . . . had conducted a poor evaluation. Our interpretation, however, is that the usability problem space is so huge that it inevitably leads to some instances of limited overlap." Furthermore, difficulties in matching problem descriptions can lead to an appearance of greater underlap than

occurs when observers have an opportunity to discuss problem matching (Hornbæk, 2010; Hornbæk and Frøkjær, 2008a, 2008b).

The interpretation of the results of these studies (Kessner et al., 2001; Molich et al., 1998, 2004; Molich and Dumas, 2008) as indicative of a lack of reliability (e.g., Law et al., 2005) stands in stark contrast to the published studies in which iterative usability tests (sometimes in combination with other UCD methods) have led to significantly improved products (Al-Awar et al., 1981; Kennedy, 1982; Lewis, 1982, 1996b; Kelley, 1984; Gould et al., 1987; Bailey et al., 1992; Bailey, 1993; Ruthford and Ramey, 2000). For example, in a paper describing their experiences in product development, Marshall et al. (1990, p. 243) stated: "Human factors work can be reliable—different human factors engineers, using different human factors techniques at different stages of a product's development, identified many of the same potential usability defects." Published cost–benefit analyses (Bias and Mayhew, 1994) have demonstrated the value of usability engineering processes that include usability testing, with cost–benefit ratios ranging from 1 : 2 for smaller projects to 1 : 100 for larger projects (C. Karat, 1997).

Most of the papers that describe the success of iterative usability testing are case studies (such as Høegh and Jensen, 2008; Marshall et al., 1990—for adaptation of usability testing to an Agile framework see Sy, 2007; Illmensee and Muff, 2009), but a few have described designed experiments. Bailey et al. (1992) compared two user interfaces derived from the same base interface: one modified via heuristic evaluation and the other modified via iterative usability testing (three iterations, five participants per iteration). They conducted this experiment with two interfaces, one character based and the other a graphical user interface (GUI), with the same basic outcomes. The number of changes indicated by usability testing was much smaller than the number indicated by heuristic evaluation, but user performance was the same with both final versions of the interface. All designs after the first iteration produced faster performance and, for the character-based interface, were preferred to the original design. The time to complete the performance testing was about the same as that required for the completion of multireviewer heuristic evaluations.

Bailey (1993) provided additional experimental evidence that iterative design based on usability tests leads to measurable improvements in the usability of an application. In the experiment, he studied the designs of eight designers, four with at least four years of professional experience in interface design and four with at least five years of professional experience in computer programming. All designers used a prototyping tool to create a recipes application (eight applications in all). In the first wave of testing, Bailey videotaped participants performing tasks with the prototypes, three different participants per prototype. Each designer reviewed the videotapes of the people using his or her prototype and used the observations to redesign his or her application. This process continued until each designer indicated that it was not possible to improve his or her application. All designers

stopped after three to five iterations. Comparison of the first and last iterations indicated significant improvement in measurements such as number of tasks completed, task completion times, and repeated serious errors.

In conclusion, the results of the studies of Molich et al. (1998, 2004; Molich and Dumas, 2008) and similar studies show that usability practitioners must conduct their usability tests as carefully as possible, document their methods completely, and show proper caution when interpreting their results. The limitations of usability testing make it insufficient for certain testing goals, such as quality assurance of safety-critical systems (Thimbleby, 2007). It can be difficult to assess complex systems with complex goals and tasks (Howard, 2008; Howard and Howard, 2009; Redish, 2007). On the other hand, as Landauer stated (1997, p. 204): "There is ample evidence that expanded task analysis and formative evaluation can, and almost always do, bring substantial improvements in the effectiveness and desirability of systems." This is echoed by Desurvire et al. (1992, p. 98): "It is generally agreed that usability testing in both field and laboratory is far and above the best method for acquiring data on usability."

## 2.3 Goals of Usability Testing

The fundamental goal of usability testing is to help developers produce more usable products. The two conceptions of usability testing (formative and summative) lead to differences in the specification of goals in much the same way that they contribute to differences in fundamental definitions of usability (diagnostic problem discovery and measurement). Rubin (1994, p. 26) expressed the formative goal as follows: "The overall goal of usability testing is to identify and rectify usability deficiencies existing in computer-based and electronic equipment and their accompanying support materials prior to release." Dumas and Redish (1999, p. 11) provided a more summative goal: "A key component of usability engineering is setting specific, quantitative, usability goals for the product early in the process and then designing to meet those goals."

These goals are not in direct conflict, but they do suggest different focuses that can lead to differences in practice. For example, a focus on measurement typically leads to more formal testing (less interaction between observers and participants), whereas a focus on problem discovery typically leads to less formal testing (more interaction between observers and participants). In addition to the distinction between diagnostic problem discovery and measurement tests, there are two common types of measurement tests: comparison against objectives and comparison of products.

### 2.3.1 Problem Discovery Test

The primary activity in diagnostic problem discovery tests is the discovery, prioritization, and resolution of usability problems. The number of participants in each iteration of testing should be fairly small, but the overall test plan should be for multiple iterations, each with some variation in participants and tasks. When the focus is on problem discovery and resolution, the assumption is that more global measures of user performance

and satisfaction will take care of themselves (Chapanis, 1981). The measurements associated with problem discovery tests are focused on prioritizing problems and include frequency of occurrence in the test, likelihood of occurrence during normal usage (taking into account the anticipated usage of the part of the product in which the problem occurred), and magnitude of impact on the participants who experienced the problem. Because the focus is not on precise measurement of the performance or attitudes of participants, problem discovery studies tend to be informal, with a considerable amount of interaction between observers and participants. Some typical stopping rules for iterations are a preplanned number of iterations or a specific problem discovery goal, such as "Identify 90% of the problems available for discovery for these types of participants, this set of tasks, and these conditions of use." As Lindgaard (2006, p. 1069) pointed out, "It is impossible to know whether all usability problems have been identified in a particular test or type of evaluation unless testing is repeated until it reaches an asymptote, a point at which no new problems emerge in a test. Asymptotic testing is not, and should not be, done in practice; it is as unfeasible as it is irrelevant in a work context." See the section below on sample size estimation and adequacy for more detailed information on setting and using these types of problem discovery objectives.

### 2.3.2 Measurement Test Type I: Comparison against Quantitative Objectives

Studies that have a primary focus of comparison against quantitative objectives include two fundamental activities (Jokela et al., 2006). The first is the development of the usability objectives. The second is iterative testing to determine if the product has met the objectives. A third activity (which can take place during iterative testing) is the enumeration and description of usability problems, but this activity is secondary to the collection of precise measurements.

The first step in developing quantitative usability objectives is to determine the appropriate variables to measure. As part of the work done for the European MUSiC (Measuring the Usability of Systems in Context) project, Rengger (1991) produced a list of potential usability measurements based on 87 papers out of a survey of 500 papers. He excluded purely diagnostic studies and also excluded papers if they did not provide measurements for the combined performance of a user and a system. He categorized the measurements into four classes:

- *Class* 1: goal achievement indicators (such as success rate and accuracy)
- *Class* 2: work rate indicators (such as speed and efficiency)
- *Class* 3: operability indicators (such as error rate and function usage)
- *Class* 4: knowledge acquisition indicators (such as learnability and learning rate)

In a later discussion of the MUSiC measures, Macleod et al. (1997) described measures of effectiveness (the level of correctness and completeness of goal achievement in context) and efficiency (effectiveness related to cost of performance, typically the effectiveness measure divided by task completion time). Optional measures were of productive time and unproductive time, with unproductive time consisting of help actions, search actions, and snag (negation, canceled, or rejected) actions.

Macleod et al.'s (1997) description of the measures of effectiveness and efficiency seem to have influenced the objectives expressed in ISO 9241-11 (ISO, 1998, p. iv): "The objective of designing and evaluating visual display terminals for usability is to enable users to achieve goals and meet needs in a particular context of use. ISO 9241-11 explains the benefits of measuring usability in terms of user performance and satisfaction. These are measured by the extent to which the intended goals of use are achieved, the resources that have to be expended to achieve the intended goals, and the extent to which the user finds the use of the product acceptable."

In practice [and as recommended by ANSI (2001)], the fundamental global measurements for usability tasks are successful task completion rates (for a measure of effectiveness), mean task completion times [for a measure of efficiency—either the arithmetic mean or, as recently suggested by Sauro and Lewis (2010), the geometric mean], and mean participant satisfaction ratings (collected either on a task-by-task basis or at the end of a test session; see Section 3.3 for more information on measuring participant satisfaction). There are many other measurements that practitioners could consider (Nielsen, 1997; Dumas and Redish, 1999), including but not limited to (1) the number of tasks completed within a specified time limit, (2) the number of wrong menu choices, (3) the number of user errors, and (4) the number of repeated errors (same user committing the same error more than once).

After determining the appropriate measurements, the next step is to set the goals. Ideally, the goals should have an objective basis and shared acceptance across the various stakeholders, such as marketing, development, and test groups (Lewis, 1982). The best objective basis for measurement goals is data from previous usability studies of predecessor or competitive products. For maximum generalizability, the historical data should come from studies of similar types of participants completing the same tasks under the same conditions (Chapanis, 1988). If this information is not available, an alternative is for the test designer to recommend objective goals and to negotiate with the other stakeholders to arrive at a set of shared goals.

According to Rosenbaum (1989, p. 211): "Defining usability objectives (and standards) isn't easy, especially when you're beginning a usability program. However, you're not restricted to the first objective you set. The important thing is to establish some specific objectives immediately, so that you can measure improvement. If the objectives turn out to be unrealistic or inappropriate, you can revise them." Such revisions, however, should take place only in the early stages of gaining experience and taking initial measurements with a product. It is important not to change reasonable goals to accommodate an unusable product.

When setting usability goals, it is usually better to set goals that make reference to an average (mean) of a measurement than to a percentile. For example, set an objective such as "The mean time to complete task 1 will be less than 5 minutes" rather than "95% of participants will complete task 1 in less than 10 minutes." The statistical reason for this is that sample means drawn from a continuous distribution are less variable than sample medians (the 50th percentile of a sample), and measurements made away from the center of a distribution (e.g., measurements made to attempt to characterize the value of the 95th percentile) are even more variable (Blalock, 1972). Cordes (1993) conducted a Monte Carlo study comparing means and medians as measurements of central tendency for time-on-task scores and determined that the mean should be the preferred metric for usability studies (unless there is missing data due to participants failing to complete tasks, in which case the mean from the study will underestimate the population mean).

A practical reason to avoid percentile goals is that the goal can imply a sample size requirement that is unnecessarily large. For example, you cannot measure accurately at the 95th percentile unless there are at least 20 measurements (in fact, there must be many more than 20 measurements for accurate measurement). For more details, see Section 3.1.

An exception to this is the specification of successful task completions (or any other measurement that is based on counting events), which necessarily requires a percentile goal, usually set at or near 100% (unless there are historical data that indicate an acceptable lower level for a specific test). If 10 out of 10 participants complete a task successfully, the observed completion rate is 100%, but a 90% exact binomial confidence interval for this result ranges from 74 to 100%. In other words, even perfect performance for 10 participants with this type of measure leaves open the possibility (with 90% confidence) that the true completion rate could be as low as 75%. See Section 3.2.2 for more information on computing and using this information in usability tests.

After the usability goals have been established, the next step is to collect data to determine if the product has met its goals. Representative participants perform the target tasks in the specified environment as test observers record the target measurements and identify, to the extent possible within the constraints of a more formal testing protocol, details about any usability problems that occur. The usability team conducting the test provides information about goal achievement and prioritized problems to the development team, and a decision is made regarding whether or not there is sufficient evidence that the product has met its objectives. The ideal stopping rule for measurement-based iterations is to continue testing until the product has met its goals.

When there are only a few goals, it is reasonable to expect to achieve all of them. When there are many goals (e.g., 5 objectives per task multiplied by 10 tasks, for a total of 50 objectives), it is more difficult to determine when to declare success and to stop testing. Thus, it is sometimes necessary to specify a metaobjective of the percentage of goals to achieve.

Despite the reluctance of some usability practitioners to conduct statistical tests to quantitatively assess the strength of the available evidence regarding whether or not a product has achieved a particular goal, the best practice is to conduct such tests. The best approach is to conduct multiple $t$-tests or nonparametric analogs of $t$-tests (Lewis, 1993) because this gives practitioners the level of detail that they require. There is a well-known prohibition against doing this because it can lead investigators to mistakenly accept as real some differences that are due to chance [technically, alpha ($\alpha$) inflation]. On the other hand, if this is the required level of information, it is an appropriate method (Abelson, 1995). Furthermore, the practice of avoiding alpha inflation is a concern more related to scientific hypothesis testing than to usability testing (Wickens, 1998), although usability practitioners should be aware of its existence and take it into account when interpreting their statistical results. For example, if you compare two products by conducting 50 $t$-tests with alpha set to 0.10, and only 5 (10%) of the $t$-tests are significant (have a $p$-value below 0.10), you should question whether or not to use those results as evidence of the superiority of one product over the other. On the other hand, if substantially more than 5 of the $t$-tests are significant, you can be more confident that the differences indicated are real.

In addition to (or as an alternative to) conducting multiple $t$-tests, practitioners should compute confidence intervals for their measurements. This applies to the measurements made for the purpose of establishing test criteria (such as measurements made on predecessor versions of the target product or competitive products) and to the measurements made when testing the product under development. See Section 3.2 for more details.

### 2.3.3 Measurement Test Type II: Comparison of Products

The second type of measurement test is to conduct usability tests for the purpose of direct comparison of one product with another. As long as there is only one measurement that decision makers plan to consider, a standard $t$-test (ideally, in combination with the computation of confidence intervals) will suffice for the purpose of determining which product is superior.

If decision makers care about multiple dependent measures, standard multivariate statistical procedures [such as multivariate analysis of variance (MANOVA) or discriminant analysis] are not often helpful in guiding a decision about which of two products has superior usability. The statistical reason for this is that multivariate statistical procedures depend on the computation of centroids (a weighted average of multiple dependent measures) using a least-squares linear model that maximizes the difference between the centroids of the two products (Cliff, 1987). If the directions of the measurements are inconsistent (e.g., a high task completion rate is desirable, but a high mean task completion time is not), the resulting centroids are uninterpretable for the purpose of usability comparison. In some cases it is possible to recompute variables so they have consistent directions (e.g., recomputing task completion rates as task failure rates). If this is not possible,

another approach is to convert measurements to ranks (Lewis, 1991a) or standardized (*Z*) scores (Sauro and Kindlund, 2005) for the purpose of principled combination of different types of measurements.

To help consumers compare the usability of different products, ANSI (2001) has published the Common Industry Format (CIF) for usability test reports. Originally developed at the National Institute of Standards and Technology (NIST), this test format requires measurement of effectiveness (accuracy and completeness —completion rates, errors, assists), efficiency (resources expended in relation to accuracy and completeness—task completion time), and satisfaction (freedom from discomfort, positive attitude toward use of the product—using any of a number of standardized satisfaction questionnaires). It also requires a complete description of participants and tasks.

Morse (2000) reviewed a NIST project conducted to pilot test the CIF. The purpose of the CIF is to make it easier for purchasers to compare the usability of different products. The pilot study ran into problems, such as inability to find a suitable software product for both supplier and consumer, reluctance to share information, and uncertainty about how to design a good usability study. To date, there has been little if any use (at least, no published use) of the CIF for its intended purpose.

## 2.4 Variations on a Theme: Other Types of Usability Tests

### 2.4.1 Think Aloud

In a standard, formal usability test, test participants perform tasks without necessarily speaking as they work. The defining characteristic of a TA study is the instruction to participants to talk about what they are doing as they do it (in other words, to produce verbal reports). If participants stop talking (as commonly happens when they become very engaged in a task), they are prompted to resume talking.

The most common theoretical justification for the use of TA is from the work in cognitive psychology (specifically, human problem solving) of Ericsson and Simon (1980). Responding to a review by Nisbett and Wilson (1977) that described various ways in which verbal reports were unreliable, Ericsson and Simon provided evidence that certain kinds of verbal reports could produce reliable data. They stated that reliable verbalizations are those that participants produce during task performance that do not require additional cognitive processing beyond the processing required for task performance and verbalization.

TA is not feasible when testing systems that include speech recognition (Lewis, 2008, 2011). For usability testing of other systems, the use of TA is fairly common. Dumas (2003) encouraged the use of TA because (1) TA tests are more productive for finding usability problems (van den Haak and de Jong, 2003; Virzi et al., 1993) and (2) thinking aloud does not affect user ratings or performance (Bowers and Snyder, 1990; Ohnemus and Biers, 1993; Olmsted-Hawala et al., 2010). There is some evidence in support of these statements, but the evidence is mixed.

Earlier prohibitions against the use of TA in measurement-based tests assumed that thinking aloud would cause slower task performance. Bowers and Snyder (1990), however, found no measurable task performance or preference differences between a test group that thought aloud and one that did not. Surprisingly, there are some experiments in which the investigators reported better task performance when participants were thinking aloud. Berry and Broadbent (1990) provided evidence that the process of thinking aloud invoked cognitive processes that improved rather than degraded performance, but only if people were given (1) verbal instructions on how to perform the task and (2) the requirement to justify each action aloud. Wright and Converse (1992) compared silent with TA usability testing protocols. The results indicated that the TA group committed fewer errors and completed tasks faster than the silent group, and the difference between the groups increased as a function of task difficulty.

Regarding the theoretical justification for and typical practice of TA, Boren and Ramey (2000) noted that TA practice in usability testing often does not conform to the theoretical basis most often cited for it (Ericsson and Simon, 1980). "If practitioners do not uniformly apply the same techniques in conducting thinking-aloud protocols, it becomes difficult to compare results between studies" (Boren and Ramey, 2000, p. 261). In a review of publications of TA tests and field observations of practitioners running TA tests, they reported inconsistency in explanations to participants about how to TA, practice periods, styles of reminding participants to TA, prompting intervals, and styles of intervention. They suggest that, rather than basing current practice on Ericsson and Simon, a better basis would be speech communication theory, with clearly defined communicative roles for the participant (in the role of domain expert or valued customer, making the participant the primary speaker) and the usability practitioner (the learner or listener, thus a secondary speaker).

Based on this alternative perspective for the justification of TA, Boren and Ramey (2000) provided guidance for many situations that are not relevant in a cognitive psychology experiment but are in usability tests. For example, they recommend that usability practitioners running a TA test should continually use acknowledgment tokens that do not take speakership away from the participant, such as "mm hm?" and "uh-huh?" (with the interrogative intonation) to encourage the participant to keep talking. In normal communication, silence (as recommended by the Ericsson and Simon protocols) is not a nonresponse—the speaker interprets it in a primarily negative way as indicating aloofness or condescension. They avoided providing precise statements about how frequently to provide acknowledgments or somewhat more explicit reminders (such as "And now . . . ?") because the best cues come from the participants. Practitioners need to be sensitive to these cues as they run the test.

Krahmer and Ummelen (2004) conducted an exploratory comparison of the Ericsson and Simon (E&S) versus the Boren and Ramey (B&R) TA procedures (10 participants per condition). They found that the

outcomes were similar for both procedures, with participants in both conditions saying about the same number of words, uncovering essentially the same navigation problems, and providing about equal evaluations of the quality of the website they used. The main difference was that moderators in the B&R condition made, as expected, more interventions and, perhaps as a consequence, the participants seemed less lost and completed more tasks.

Hertzum et al. (2009) compared silent task completion with strict E&S and more relaxed TA, supplemented with eye tracking and assessment of mental workload. Strict E&S TA, other than requiring more time for task completion, led to similar results as the silent condition. Relaxed TA affected participant behavior in multiple ways. Relative to silence, the TA method did not affect successful task completion rates, which tended to be high in the study. In the relaxed TA condition, participants spent more time in general distributed visual behavior, issued more navigation commands, and experienced higher mental workload.

Olmsted-Hawala et al. (2010) used a double-blind procedure to investigate the effect of different TA procedures on successful task completion, task completion times, and satisfaction. Their experimental conditions were the traditional E&S, speech-communication-based B&R, a less restrictive coaching protocol in which moderators could freely probe participants, and silence (no TA at all), with 20 participants per condition. The outcomes were similar for silence, E&S, and B&R procedures. Participants in the coaching condition successfully completed significantly more tasks and had higher satisfaction ratings. Their results for B&R differed from those reported by Krahmer and Umullen (2004): "since the test administrator in the Krahmer & Umullen study offered assistance and encouragement to the test subject during the session, we think their speech-communication protocol is more akin to the coaching condition in our study" (Olmsted-Hawala et al., 2010, p. 2387).The evidence indicates that relative to silent participation, TA can affect task performance and reported satisfaction, depending on the exact TA protocol in use. If the primary purpose of the test is problem discovery, TA appears to have advantages over completely silent task completion. If the primary purpose of the test is task performance measurement, the use of TA is somewhat more complicated. As long as all the tasks in the planned comparisons were completed under the same conditions, performance comparisons should be legitimate. It is critical, however, that practitioners using TA provide a complete description of their method, including the kind and frequency of probing.

The use of TA almost certainly prevents generalization of task performance outside the TA task, but there are many other factors that make it difficult to generalize specific task performance data collected in usability studies. For example, Cordes (2001) demonstrated that participants assume that the tasks they are asked to perform in usability tests are possible (the "I know it can be done or you wouldn't have asked me to do it" bias). Manipulations that bring this assumption into doubt can have a strong effect on quantitative usability

performance measures, such as increasing the percentage of participants who give up on a task. If uncontrolled, this bias makes performance measures from usability studies unlikely to be representative of real-world performance when users are uncertain as to whether the product they are using can support the desired tasks.

The discussion above focuses on concurrent TA, with participants talking aloud as they perform tasks. An alternative approach is to use stimulated retrospective TA, in which participants perform tasks silently, and then talk as they review the video of their task performance—an approach that avoids any influence of TA on task performance but requires twice as much time to complete data collection in a usability study. Bowers and Snyder (1990) reported similar task performance and subjective measures for concurrent and retrospective TA, but participants provided different types of information as a function of TA style, with participants in the concurrent condition tending to provide procedural information, and participants in the retrospective condition tending to give explanations and design statements. Similar findings were reported by van den Haak and de Jong (2003), along with fewer successful task completions for TA relative to silent work. Using eye tracking as a way to assess a participant's focus of attention, Guan et al. (2006) found the retrospective method to be valid and reliable, with a low risk of introducing fabrication, and with no significant effect of task complexity. Karahasanovic et al. (2009), comparing concurrent and retrospective TA with a feedback collection method (FCM) in which participants respond to probes during task performance, found that all methods were intrusive with regard to completion rates and times, but the FCM was less time consuming to analyze.

Clemmensen et al. (2009) discussed the impact of cultural differences on TA. There are several ways in which cultural differences could affect testing, such as the instructions and tasks, the participant's verbalization, how the observer "reads" the participant, and the overall relationship between participant and observer. In particular, with regard to studies that have Western observers and Eastern participants, they recommended that observers should allow sufficient time for participants to pause while thinking aloud, rely less on expressions of surprise, and be sensitive to the tendency for indirect criticism.

### 2.4.2 Multiple Simultaneous Participants

Downey (2007) described group usability testing in which multiple observers watch a number of participants individually but simultaneously perform tasks. A key benefit of the method was obtaining data from more people over a shorter period of time. She reported that the method appeared to be most effective when tasks were relatively simple and a focused discussion followed the group's completion of the tasks.

Another way to encourage participants to talk during task completion is to have them work together (Wildman, 1995), a method sometimes called constructive interaction (Nielsen, 1993). This strategy is similar to TA in its strengths and limitations, but with potentially

greater ecological validity, including less participant awareness of the observer (van den Haak and de Jong, 2005; van den Haak et al., 2006).

Hackman and Biers (1992) compared three TA methods: thinking aloud alone (Single), thinking aloud in the presence of an observer (Observer), and verbalizations occurring in a two-person team (Team). They found no significant differences in performance or subjective measures. The Team condition produced more statements of value to designers than the other two conditions, but this was probably due to the differing number of participants producing statements in the different conditions. There were three groups, with 10 participants per group for Single and Observer and 20 participants (10 two-person teams) for the Team condition. "The major result was that the team gave significantly more verbalizations of high value to designers and spent more time making high value comments. Although this can be reduced to the fact that the team spoke more overall and that there are two people talking rather than one, this finding is not trivial" (Hackman and Biers, 1992, p. 1208).

### 2.4.3  Remote Evaluation

Recent advances in the technology of collaborative software have made it easier to conduct remote software tests—tests in which the usability practitioner and the test participant are in different locations (Albert et al., 2010; Ramli and Jaafar, 2009). This can be an economical alternative to bringing one or more users into a laboratory for face-to-face user testing. A participant in a remote location can view the contents of the practitioner's screen, and in a typical system the practitioner can decide whether the participant can control the desktop. System performance is typically slower than that of a local test session.

Some of the advantages of remote testing are (1) access to participants who would otherwise be unable to participate (international, special needs, etc.), (2) the capability for participants to work in familiar surroundings, and (3) no need for either party to install or download additional software. Some of the disadvantages are (1) potential uncontrolled disruptions in the participant's workplace, (2) lack of visual feedback from the participant, and (3) the possibility of compromised security if the participant takes screen captures of confidential material. Despite these disadvantages, McFadden et al. (2002) reported data that indicated that remote testing was effective at improving product designs and that the test results were comparable to the results obtained with more traditional testing.

As described above, synchronous remote usability testing has similar time constraints as laboratory-based tests (Dumas and Salzman, 2006). More fully automated asynchronous usability testing has become available which permits more rapid testing, typically with the participant receiving information about the task and responding to questions in one window while working with the product in a different window (West and Lehman, 2006). A clear disadvantage of this type of unmoderated testing is the lack of interaction between observers and participants, but Tullis et al. (2002) reported no substantial differences between unmoderated

and laboratory testing for quantitative measurements or problem discovery. West and Lehman (2006) also reported consistency between task success and satisfaction metrics between standard and automated summative usability testing but noted that having a usability engineer observe the sessions led to the discovery of a more comprehensive set of issues. In a study focused on problem discovery, Andreasen et al. (2007) reported similar outcomes between laboratory-based and synchronous remote usability testing but found fewer problems discovered with asynchronous testing. In contrast, Bosenick et al. (2007) reported the discovery of more usability issues with remote asynchronous testing. Hopefully, further research will reveal the reasons for these discrepant outcomes when comparing asynchronous usability testing to more standard laboratory-based testing.

### 2.5  Usability Laboratories

A typical usability laboratory test suite is a set of sound-proofed rooms with a participant area and observer area separated by a one-way glass and with video cameras and microphones to capture the user experience (Marshall et al., 1990; Nielsen, 1997), possibly with an executive viewing area behind the primary observers' area. The advantages of this type of usability facility are quick setup, a place where designers can see people interacting with their products, videos to provide a historical record and backup for observers, and a professional appearance that raises awareness of usability and reassures customers about commitment to usability. In a survey of usability laboratories, Nielsen (1994) reported a median floor space of 63 m$^2$ (678 ft$^2$) for the observer room and 13 m$^2$ (144 ft$^2$) for test rooms. This type of laboratory is especially important if practitioners plan to conduct formal, summative usability tests.

If the practitioner focus is on formative, diagnostic problem discovery, this type of laboratory is not essential (although it is still convenient). "It is possible to convert a regular office temporarily into a usability laboratory, and it is possible to perform usability testing with no more equipment than a notepad" (Nielsen, 1997, p. 1561). Making an even stronger statement against the perceived requirement for formal laboratories, Landauer (1997, p. 204) stated: "Many usability practitioners have demanded greater resources and more elaborate procedures than are strictly needed for effective guidance—such as expensive usability labs rather than natural settings for test and observations, time consuming videotaping and analysis where observation and note-taking would serve as well, and large groups of participants to achieve statistical significance when qualitative naturalistic observation of task goals and situations, or of disastrous interface or functionality flaws, would be more to the point."

In addition to remote usability testing (discussed above), another alternative to a formal, fixed-location usability laboratory is a mobile laboratory (Seffah and Habieb-Mammar, 2009). Advantages of mobile usability laboratories include portability to a participant's workplace and reduced cost relative to fixed laboratories. Because the mobile usability laboratory moves to the participant, disadvantages include the need to reduce the

size of the usability testing team and complications in allowing nonteam observers to view the test.

## 2.6 Test Roles

There are several ways to categorize the roles that testers need to play in the preparation and execution of a usability test (Rubin, 1994; Dumas and Redish, 1999). Most test teams will not have a person assigned to each role, and most tests (especially informal problem discovery tests) do not require every role. The actual distribution of skills across a team might vary from these roles, but the standard roles help to organize the skills needed for effective usability testing.

### 2.6.1 Test Administrator

The test administrator is the usability test team leader. He or she designs the usability study, including the specification of the initial conditions for a test session and the codes to use for data logging. The test administrator's duties include conducting reviews with the rest of the test team, leading in the analysis of data, and putting together the final presentation or report. People in this role should have a solid understanding of the basics of usability engineering, ability to tolerate ambiguity, flexibility (knowing when to deviate from the plan), and good communication skills.

### 2.6.2 Briefer

The briefer is the person who interacts with the participants (briefing them at the start of the test, communicating with them as required during the test, and debriefing them at the end of the test sessions). On many teams, the same person takes the roles of administrator and briefer. In a TA study, the briefer has the responsibility to keep the participant talking. The briefer needs to have sufficient familiarity with the product to be able to decide what to tell participants when they ask questions. People in this role need to be comfortable interacting with people and need to be able to restrict their interactions to those that are consistent with the purposes of the test without any negative treatment of the participants.

### 2.6.3 Camera Operator

The camera operator is responsible for running the audiovisual equipment during the test. He or she must be skilled in the setup and operation of the equipment and must be able to take directions quickly when it is necessary to change the focus of the camera (e.g., from the keyboard to the user manual).

### 2.6.4 Data Recorder

The video record is useful as a data backup when things start happening quickly during the test and as a source for video examples when documenting usability problems. The primary data source for a usability study, however, is the notes that the data recorder takes during a test session. There just is not time to take notes from a more leisurely examination of the video record. Also, the camera does not necessarily catch the important action at every moment of a usability study.

For informal studies, the equipment used to record data might be nothing more than a notepad and pencil. Alternatively, the data recorder might use data-logging software to take coded notes (often time stamped, possibly synchronized with the video). Before the test begins, the data recorder needs to prepare the data-logging software with the category codes defined by the test administrator. Taking notes with data-logging software is a very demanding skill, so the test administrator does not usually assign additional tasks to the person taking this role.

### 2.6.5 Help Desk Operator

The help desk operator takes calls from the participant if the user experiences enough difficulty to place the call. The operator should have some familiarity with the call-center procedures followed by the company that has designed the product under test and must also have skills similar to those of the briefer.

### 2.6.6 Product Expert

The product expert maintains the product and offers technical guidance during the test. The product expert must have sufficient knowledge of the product to recover quickly from product failures and to help the other team members understand the system's actions during the test.

### 2.6.7 Statistician

A statistician has expertise in measurement and the statistical analysis of data. Practitioners with an educational background in experimental psychology typically have sufficient expertise to take the role of statistician for a usability test team. Informal tests rarely require the services of a statistician, but the team needs a statistician to extract the maximum amount of information from the data gathered during a formal test (especially if the purpose of the formal test was to compare two products using a battery of measurements).

## 2.7 Planning the Test

One of the first activities that a test administrator must undertake is to develop a test plan. To do this, the administrator must understand the purpose of the product, the parts of the product that are ready for test, the types of people who will use the product, what they are likely to use the product for, and in what settings.

### 2.7.1 Purpose of the Test

At the highest level, is the primary purpose of the test to identify usability problems or to gather usability measurements? The answer to this question provides guidance as to whether the most appropriate test is formal or informal, TA or silent, problem discovery or quantitative measurement. After addressing this question, the next task is to define any more specific test objectives. For example, an objective for an interactive voice response (IVR) system might be to assess whether participants can accomplish key tasks without encountering significant problems. If data are available from a previous study of a similar IVR, an alternative objective might be to determine whether participants can complete

key tasks reliably faster with the new IVR than they did with the previous IVR. Most usability tests will include several objectives.

If a key objective of the test is to compare two products, an important decision is whether the test will be within subjects or between subjects. In a within-subjects test, every participant works with both products, with half of the participants using one product first and the other half using the other product first (a technique known as *counterbalancing*). In a between-subjects study, the test groups are completely independent. In general, a within-subjects test leads to more precise measurement of product differences (requiring a smaller number of participants for equal precision, due primarily to the reduction in variability that occurs when each participant acts as his or her own control) and the opportunity to get direct subjective product comparisons from participants. Tohidi et al. (2006) reported that participants exposed to alternative design solutions (within subjects) were more likely to provide informative criticism of the designs than participants who worked with only one of the designs (between subjects). For a within-subjects test to be feasible, both products must be available and set up for use in the lab at the same time, and the amount of time needed to complete tasks with both products must not be excessive. If a within-subjects test is not possible, a between-subjects test is a perfectly valid alternative. Note that the statistical analyses appropriate for these two types of tests are different.

### 2.7.2 Participants

To determine who will participate in the test, the administrator needs to obtain or develop a user profile. A user profile is sometimes available from the marketing group, the product's functional specification, or other product planning documentation. It is important to keep in mind that the focus of a usability test is the end user of a product, not the expected product purchaser (unless the product will be purchased by end users). The most important participant characteristic is that the participant is representative of the population of end users to whom the administrator wants to generalize the results of the test. Practitioners can obtain participants from employment agencies, internal sources if the participants meet the requirements of the user profile (but avoiding internal test groups), market research firms, existing customers, colleges, newspaper ads, and user groups.

To define representativeness, it is important to specify the characteristics that members of the target population share but are not characteristic of nonmembers. The administrator must do this for the target population at large and any defined subgroups. Within group definition constraints, administrators should seek heterogeneity in the final sample to maximize the generalizability of the results (Chapanis, 1988; Landauer, 1997) and to maximize the likelihood of problem discovery. It is true that performance measurements made with a homogeneous sample will almost always have greater precision than measurements made with a heterogeneous sample, but the cost of that increased precision is limited generalizability. This raises the issue of how to define homogeneity and heterogeneity of participants. After all,

at the highest level of categorization, we are all humans, with similar general capabilities and limitations (physical and cognitive). At the other end of the spectrum, we are all individuals—no two alike.

One of the most important defining characteristics for a group in a usability test is specific relevant experience, both with the product and in the domain of interest (work experience, general product experience, specific product experience, experience with the product under test, and experience with similar products). One common categorization scheme is to consider people with less than three months' experience as novices, with more than a year of experience as expert, and those in between as intermediate (Dumas and Redish, 1999). Other individual differences that practitioners routinely track and attempt to vary are education level, age, and gender.

When acquiring participants, how can practitioners define the similarity between the participants they can acquire and the target population? An initial step is to develop a taxonomy of the variables that affect human performance (where performance should include the behaviors of indicating preference and other choice behaviors). Gawron et al. (1989) produced a human performance taxonomy during the development of a human performance expert system. They reviewed existing taxonomies and filled in some missing pieces. They structured the taxonomy as having three top levels: environment, subject (person), and task. The resulting taxonomy took up 12 pages in their paper and covered many areas that would normally not concern a usability practitioner working in the field of computer system usability (e.g., ambient vapor pressure, gravity, acceleration). Some of the key human variables in the Gawron et al. (1989) taxonomy that could affect human performance with computer systems are:

- *Physical characteristics*
  - Age
  - Agility
  - Handedness
  - Voice
  - Fatigue
  - Gender
  - Body and body part size
- *Mental state*
  - Attention span
  - Use of drugs (both prescription and illicit)
  - Long-term memory (includes previous experience)
  - Short-term memory
  - Personality traits
  - Work schedule
- *Senses*
  - Auditory acuity
  - Tone perception
  - Tactual
  - Visual accommodation

- Visual acuity
- Color perception

These variables can guide practitioners as they attempt to describe how participants and target populations are similar or different. The Gawron et al. (1989) taxonomy, however, does not provide much detail with regard to some individual differences that other researchers have hypothesized to affect human performance or preference with respect to the use of computer systems: personality traits and computer-specific experience.

Aykin and Aykin (1991) performed a comprehensive review of the published studies to that date that involved individual differences in human–computer interaction (HCI). Table 1 lists the individual differences that they found in published HCI studies, the method used to measure the individual difference, and whether there was any indication from the literature that manipulation of that individual difference led to a crossed interaction.

In statistical terminology, an interaction occurs whenever an experimental treatment has a different magnitude of effect depending on the level of a different, independent experimental treatment. A crossed interaction occurs when the magnitudes have different signs, indicating reversed directions of effects. As an example of an uncrossed interaction, consider the effect of turning off the lights on the typing throughput of blind and sighted typists. The performance of the sighted typists would probably be worse, but the presence or absence of light should not affect the performance of the blind typists. As an extreme example of a crossed interaction, consider the effect of language on task completion for people fluent only in French or English. When reading French text, French speakers would outperform English speakers, and vice versa.

For any of these individual differences, the lack of evidence for crossed interactions could be due to a paucity of research involving the individual difference or could reflect the probability that individual differences will not typically cause crossed interactions in HCI. In general, a change made to support a problem experienced by a person with a particular individual difference will either help other users or simply not affect their performance.

For example, John Black (personal communication, 1988) cited the difficulty that field-dependent users had working with one-line editors at the time (decades ago) when that was the typical user interface to a mainframe computer. Switching to full-screen editing resulted in a performance improvement for both field-dependent and field-independent users—an uncrossed interaction because both types of users improved, with the performance of field-dependent users becoming equal to (thus improving more than) that of field-independent users. Landauer (1997) cites another example of this, in which Greene et al. (1986) found that young people with high scores on logical reasoning tests could master database query languages such as SQL with little training, but older or less able people could hardly ever master these languages. They also determined that an alternative way of forming queries, selecting rows from a truth table, allowed almost everyone to make correct specification of queries, independent of their abilities. Because this redesign improved the performance of less able users without diminishing the performance of the more able, it was an uncrossed interaction. In a more recent study, Palmquist and Kim (2000) found that field dependence affected the search performance of novices using a Web browser (with field-independent users searching more efficiently) but did not affect the performance of more experienced users.

If there is a reason to suspect that an individual difference will lead to a crossed interaction as a function of interface design, it could make sense to invest the time (which can be considerable) to categorize users

**Table 1  Results of Aykin and Aykin (1991) Review of Individual Differences in HCI**

| Individual Difference | Measurement Method | Crossed Interactions |
|---|---|---|
| Level of experience | Various methods | No |
| Jungian personality types | Myers–Briggs type of indicator | No |
| Field dependence/ independence | Embedded figures test | Yes; field-dependent participants preferred organized sequential item number search mode, but field-independent subjects preferred the less organized keyword search mode (Fowler et al., 1985) |
| Locus of control | Levenson test | No |
| Imagery | Individual differences questionnaire | No |
| Spatial ability | VZ-2 | No |
| Type A/type B personality | Jenkins activity survey | No |
| Ambiguity tolerance | Ambiguity tolerance scale | No |
| Gender | Unspecified | No |
| Age | Unspecified | No |
| Other (reading speed and comprehension, intelligence, mathematical ability) | Unspecified | No |

according to these dimensions. Another situation in which it could make sense to invest the time in categorization by individual difference would be if there were reasons to believe that a change in interface would greatly help one or more groups without adversely affecting other groups. (This is a strategy that one can employ when developing hypotheses about ways to improve user interfaces.) It always makes sense to keep track of user characteristics when categorization is easy (e.g., age or gender). Another potential use of these types of variables is as covariates (used to reduce estimates of variability) in advanced statistical analyses (Cliff, 1987).

Aykin and Aykin (1991) reported effects of users' levels of experience but did not report any crossed interactions related to this individual difference. They did report that interface differences tended to affect the performance of novices but had little effect on the performance of experts. It appears that behavioral differences related to user interfaces (Aykin and Aykin, 1991) and cognitive style (Palmquist and Kim, 2000) tend to fade with practice. Nonetheless, user experience has been one of the few individual differences to receive considerable attention in HCI research (Fisher, 1991; Mayer, 1997; Miller et al., 1997; Smith et al., 1999). According to Mayer (1997), relative to novices, experts have (1) better knowledge of syntax, (2) an integrated conceptual model of the system, (3) more categories for more types of routines, and (4) higher level plans.

Fisher (1991) emphasized the importance of discriminating between computer experience (which he placed on a novice–experienced dimension) and domain expertise (which he placed on a naive–expert dimension). LaLomia and Sidowski (1990) reviewed the scales and questionnaires developed to assess computer satisfaction, literacy, and aptitudes. None of the instruments they surveyed specifically addressed measurement of computer experience. Miller et al. (1997) published the Windows Computer Experience Questionnaire (WCEQ), an instrument specifically designed to measure a person's experience with Windows 3.1. The questionnaire took about 5 min to complete and was reliable (coefficient $\alpha = 0.74$; test–retest correlation = 0.97). They found that their questionnaire was sensitive to three experiential factors: general Windows experience, advanced Windows experience, and instruction. Arning and Ziefle (2008) published an 18-item computer expertise questionnaire for older adults (the CE) which assesses both theoretical computer knowledge and practical computer knowledge and takes about 20 min to complete.

Smith et al. (1999) distinguished between subjective and objective computer experience. The paper was relatively theoretical and "challenges researchers to devise a reliable and valid measure" (p. 239) for subjective computer experience, but did not offer one. One user characteristic not addressed in any of the literature cited is one that becomes very important when designing products for international use: cultural characteristics. For example, in adapting an interface for use by members of another country, it is extremely important that all text be translated accurately. It is also important to be sensitive to the possibility that these types of individual differences might be more likely than others to result in crossed interactions.

For comparison studies, having multiple groups (e.g., males and females or experts and novices) allows the assessment of potential interactions that might otherwise go unnoticed. Ultimately, the decision for one or multiple groups must be based on expert judgment and a few guidelines. For example, practitioners should consider sampling from different groups if they have reason to believe:

- There are potential and important differences among groups on key measures (Dickens, 1987).
- There are potential interactions as a function of group (Aykin and Aykin, 1991).
- The variability of key measures differs as a function of the group.
- The cost of sampling differs significantly from group to group.

Gordon and Langmaid (1988) recommended the following approach to defining groups:

1. Write down all the important variables.
2. If necessary, prioritize the list.
3. Design an ideal sample.
4. Apply common sense to collapse cells.

For example, suppose that a practitioner starts with 24 cells, based on the factorial combination of six demographic locations, two levels of experience, and the two levels of gender. The practitioner should ask himself or herself whether there is a high likelihood of learning anything new and important after completing the first few cells or whether additional testing would be wasteful. Can one learn just as much from having one or a few cells that are homogeneous within cells and heterogeneous between cells with respect to an important variable but are heterogeneous within cells with regard to other, less important variables? For example, a practitioner might plan to (1) include equal numbers of males and females over and under 40 years of age in each cell, (2) have separate cells for novice and experienced users, and (3) drop intermediate users from the test. The resulting design requires testing only two cells (groups), but a design that did not combine genders and age groups in the cells would have required eight cells.

The final issue is the number of participants to include in the test. According to Dumas and Redish (1999), typical usability tests have 6–12 participants divided among two to three subgroups. For any given test, the required sample size depends on the number of subgroups, available resources (time/money), and purpose of the test (e.g., precise measurement or problem discovery). It also depends on whether a study is single shot (needing a larger sample size) or iterative (needing a smaller sample size per iteration, building up the total sample size over iterations). For more detailed treatment of this topic, see Section 3.1.

### 2.7.3 Test Task Scenarios

As with participants, the most important consideration for test tasks is that they are representative of the types of tasks that real users will perform with the product. For any product, there will be a core set of tasks that anyone using the product will perform. People who use barbecue grills use them to cook. People who use desktop speech dictation products use them to produce text. For usability tests, these are the most important tasks to test.

After defining these core tasks, the next step is to list any more peripheral tasks that the test should cover. If a barbecue grill has an external burner for heating pans, it might make sense to include a task that requires participants to work with that burner. If in addition to the basic vocabulary in a speech dictation system the program allows users to enable additional special topic vocabularies such as cooking or sports, it might make sense to devise a task that requires participants to activate and use one of these topics. Practitioners should avoid frivolous or humorous tasks because what is humorous to one person might be offensive or annoying to another.

From the list of test tasks, create scenarios of use (with specific goals) that require participants to perform the identified tasks. Critical tasks can appear in more than one scenario. For repeated tasks, vary the task details to increase the generalizability of the results. When testing relatively complex systems, some scenarios should stay within specific parts of the system (e.g., typing and formatting a document) and others should require the use of different parts of the system (e.g., creating a figure using a spreadsheet program, adding it to the document, attaching the document to a note, and sending it to a specified recipient).

The complete specification of a scenario should include several items. It is important to document (but not to share with the participant) the required initial conditions so it will be easy to determine before a test session starts if the system is ready and the required ending conditions that define successful task completion (Howard, 2008; Howard and Howard, 2009). The written description of the scenario (presented to the participant) should state what the participant is trying to achieve and why (the motivation), keeping the description of the scenario as short as possible to keep the test session moving quickly. The scenario should end with an instruction for the action the participant should take upon finishing the task (to make it easier to measure task completion times). The descriptions of the scenario's tasks should not typically provide step-by-step instructions on how to complete the task but should include details (e.g., actual names and data) rather than general statements. For tasks in which users work with highly personalized data (email, calendar, financial), scenarios constructed with a participant's own real data can increase the validity of the study (Genov et al., 2009).

The order in which participants complete scenarios should reflect the way in which users would typically work and with the importance of the scenario, with important scenarios done first unless there are other less important scenarios that produce outputs that the important scenario requires as an initial condition. Not all participants need to receive the same scenarios, especially if there are different groups under study. The tasks performed by administrators of a Web system that manages subscriptions will be different from the tasks performed by users who are requesting subscriptions.

Here are some examples of scenarios:

- Frank Smith's business telephone number has changed to (896) 555-1234. Please change the appropriate address book entry so you have this new phone number available when you need it. When you have finished, please say "I'm done."
- You've just found out that you need to cancel a car reservation that you made for next Wednesday. Please call the system that you used to make the reservation (1-888-555-1234) and cancel it. When you have finished, please hang up the phone and say "I'm done."

Bailey et al. (2009) have described stopping a task after the first step as a means for assessing a large number of tasks in a relatively short period of time. Over a number of website studies, they found that if the first click of a task was correct, the likelihood of final task success was 0.87, whereas if the first click was incorrect, the likelihood of final success was 0.46. The more tasks covered in a usability test, the greater the likelihood of discovery of usability problems (Lindgaard and Chattratichart, 2007).

### 2.7.4 Procedure

The test plan should include a description of the procedures to follow when conducting a test session. Most test sessions include an introduction, task performance, posttask activities, and debriefing.

A common structure for the introduction is for the briefer (review Section 2.6) to start with the purpose of the test, emphasizing that its goal is to improve the product, not to test the participant. Participation is voluntary, and the participant can stop at any time without penalty. The briefer should inform the participant that all test results will be confidential. The participant should be aware of any planned audio or video recording. Finally, the briefer should provide any special instructions (e.g., TA instructions) and answer any other questions that the participant might have.

The participant should then complete any preliminary questionnaires and forms, such as a background questionnaire, an informed consent form (including consent for any recording, if applicable), and, if necessary, a confidential disclosure form. If the participant will be using a workstation, the briefer should help the participant make any necessary adjustments (unless, of course, the purpose of the test is to evaluate workstation adjustability). Finally, the participant should complete any prerequisite training. This can be especially important if the goal of the study is to investigate usability after some period of use (ease of use) rather than immediate usability (ease of learning).

The procedure section should indicate the order in which participants will complete task scenarios. For

each participant, start with the first task scenario assigned and complete additional scenarios until the participant finishes (or runs out of time). The procedure section should specify when and how to interact with participants, according to the type of study. This section should also indicate when it is permissible to provide assistance to participants if they encounter difficulties in task performance.

Normally, practitioners should avoid offering assistance unless the participant is visibly distressed. When participants initially request help at a given step in a task, refer them to documentation or other supporting materials if available. If that doesn't help, provide the minimal assistance required to keep the participant moving forward in the task, note the assistance, and score the task as failed. When participants ask questions, try to avoid direct answers, instead turning their attention back to the task and encouraging them to take whatever action seems right at that time. When asking questions of participants, it is important to avoid biasing the participant's response. Try to avoid the use of loaded adjectives and adverbs in questions (Dumas and Redish, 1999). Instead of asking if a task was easy, ask the participant to describe what it was like performing the task. Give a short satisfaction questionnaire (such as the ASQ; see Section 3.3 for details) at the end of each scenario.

After participants have finished the assigned scenarios, it is common to have them complete a final questionnaire, usually a standard questionnaire and any additional items required to cover other test- or product-specific issues. For standardized questionnaires, ISO lists the SUMI (Software Usability Measurement Inventory) (Kirakowski and Corbett, 1993; Kirakowski, 1996) and PSSUQ (Post-Study System Usability Questionnaire) (Lewis, 1995, 2002). In addition to the SUMI and PSSUQ, ANSI lists the QUIS (Questionnaire for User Interaction Satisfaction) (Chin et al., 1988) and SUS (System Usability Scale) (Brooke, 1996) as widely used questionnaires. After completing the final questionnaire, the briefer should debrief the participant. Toward the end of debriefing, the briefer should tell the participant that the test session has turned up several opportunities for product improvement (this is almost always true) and thank the participant for his or her contribution to product improvement. Finally, the briefer should discuss any questions that the participant has about the test session and then take care of any remaining activities, such as completing time cards. If any deception has been employed in the test (which is rare but can happen legitimately when conducting certain types of simulations), the briefer must inform the participant.

### 2.7.5   Pilot Testing

Practitioners should always plan for a pilot test before running a usability test. A usability test is a designed artifact and like any other designed artifact needs at least some usability testing to find problems in the test procedures and materials. A common strategy is to have an initial walkthrough with a member of the usability test team or some other convenient participant. After making the appropriate adjustments, the next pilot participant should be a more representative participant. If there

are no changes made to the design of the usability test after running this participant, the second pilot participant can become the first real participant (but this is rare). Pilot testing should continue until the test procedures and materials have become stable.

### 2.7.6   Number of Iterations

It is better to run one usability test than not to run any at all. On the other hand, "usability testing is most powerful and most effective when implemented as part of an iterative product development process" (Rubin, 1994, p. 30). Ideally, usability testing should begin early and occur repeatedly throughout the development cycle. When development cycles are short, it is a common practice to run, at a minimum, exploratory usability tests on prototypes at the beginning of a project, to run a usability test on an early version of the product during the later part of functional testing, and then to run another during system testing. Once the final version of the product is available, some organizations run an additional usability test focused on the measurement of usability performance benchmarks. At this stage of development, it is too late to apply information about any problems discovered during the usability test to the soon-to-be-released version of the product, but the information can be useful as early input to a follow-on product if the organization plans to develop another version of the product.

### 2.7.7   Ethical Treatment of Test Participants

Usability testing always involves human participants, so usability practitioners must be aware of professional practices in the ethical treatment of test participants. Practitioners with professional education in experimental psychology are usually familiar with the guidelines of the American Psychology Association (APA; see http://www.apa.org/ethics/), and those with training in human factors engineering are usually familiar with the guidelines of the Human Factors and Ergonomics Society (HFES) (see http://www.hfes.org/About/Code.html). It is particularly important (Dumas, 2003) to be aware of the concepts of informed consent (participants are aware of what will happen during the test, agree to participate, and can leave the test at any time without penalty) and minimal risk (participating in the test does not place participants at any greater risk of harm or discomfort than situations normally encountered in daily life). Most usability tests are consistent with guidelines for informed consent and minimal risk. Only the test administrator should be able to match a participant's name and data, and the names of test participants should be confidential. Anyone interacting with a participant in a usability test has a responsibility to treat the participant with respect.

Usability practitioners rarely use deception in usability tests. One technique in which there is potential use of deception is the WOZ method (originally, the OZ Paradigm) (Kelley, 1985; see also http://www.musicman.net/oz.html). In a test using the WOZ method, a human (the Wizard) plays the part of the system, remotely controlling what the participant sees happen in response to the participant's manipulations. This method

is particularly effective in early tests of speech recognition IVR systems because all the Wizard needs is a script and a phone (Sadowski, 2001). Often, there is no compelling reason to deceive participants, so they know that the system they are working with is remotely controlled by another person for the purpose of early evaluation. If there is a compelling need for deception (e.g., to manage the participant's expectations and encourage natural behaviors), this deception must be revealed to the participant during debriefing.

## 2.8  Reporting Results

There are two broad classes of usability test results, problem reports and quantitative measurements. It is possible for a test report to contain one type exclusively (e.g., the ANSI Common Industry Format has no provision for reporting problems, which led the National Institute of Standards and Technology to investigate a similar standard for formative test reports; see Theofanos and Quesenbery, 2005), but most usability test reports will contain both types of results. Høegh et al. (2006) reported that usability reports can have a strong impact on developers' understanding of specific usability problems, especially if the developers have also observed usability test sessions. Of particular interest to the developers was the list of specific usability problems and redesign proposals, consistent with the results of Capra (2007) and Nørgaard and Hornbæk (2009).

### 2.8.1  Describing Usability Problems

"We broadly define a usability defect as: Anything in the product that prevents a target user from achieving a target task with reasonable effort and within a reasonable time.... Finding usability problems is relatively easy. However, it is much harder to agree on their importance, their causes and the changes that should be made to eliminate them (the fixes)" (Marshall et al., 1990, p. 245).

The best way to describe usability problems depends on the purpose of the descriptions. For usability practitioners, the goal should be to describe problems in such a way that the description leads logically to one or more potential interventions (recommendations) that will help designers and developers improve the system under evaluation (Høegh et al., 2006; Hornbæk, 2010). Ideally, the problem description should also include some indication of the importance of fixing the problem (most often referred to as problem severity). For more scientific investigations, there can be value in higher levels of problem description (Keenan et al., 1999), but developers rarely care about these levels of description. They just want to know what they need to do to make things better while also managing the cost (both monetary and time) of interventions (Gray and Salzman, 1998).

The problem description scheme of Lewis and Norman (1986) has both scientific and practical merit because their problem description categories indicate, at least roughly, an appropriate intervention. They stated (p. 413) that "although we do not believe it possible to design systems in which people do not make errors, we do believe that much can be done to minimize the incidence of error, to maximize the discovery of the error, and to make it easier to recover from the error." They separated errors into mistakes (errors due to incorrect intention) and slips (errors due to appropriate intention but incorrect action), further breaking slips down into mode errors (which indicate a need for better feedback or elimination of the mode), capture errors (which indicate a need for better feedback), and description errors (which indicate a need for better design consistency). In one study using this type of problem categorization, Prümper et al. (1992) found that expertise did not affect the raw number of errors made by participants in their study, but experts handled errors much more quickly than novices. The types of errors that experts made were different from those made by novices, with experts' errors occurring primarily at the level of slips rather than mistakes (knowledge errors).

Using an approach similar to that of Lewis and Norman (1986), Rasmussen (1986) described three levels of errors: skill based, rule based, and knowledge based. Other classification schemes include Structured Usability Problem Extraction, or SUPEX (Cockton and Lavery, 1999), the User Action Framework, or UAF (Andre et al., 2000), and the Classification of Usability Problems (CUP) scheme (Vilbergsdóttir et al., 2006). The UAF requires a series of decisions, starting with an interaction cycle (planning, physical actions, assessment) based on the work of Norman (1986). Most classifications require four or five decisions, with interrater reliability [as measured with kappa ($\kappa$)] highest at the first step ($\kappa = 0.978$) but remaining high through the fourth and fifth steps ($\kappa > 0.7$).

Whether any of these classification schemes will see widespread use by usability practitioners is still unknown. For example, the CUP scheme requires some training for inexperienced evaluators to effectively use the scheme, even though a simplified version may be useful for developers and usability practitioners (Vilbergsdóttir et al, 2006). There is considerable pressure on practitioners to produce results and recommendations as quickly as possible. Even if these classification schemes see little use by practitioners, effective problem classification is a very important problem to solve as usability researchers strive to compare and improve usability testing methods.

### 2.8.2  Crafting Design Recommendations from Problem Descriptions

The development of recommendations from problem descriptions is a craft rather than a rote procedure. A well-written problem description will often strongly imply an intervention, but it is also often the case that there might be several ways to attack a problem. It can be helpful for practitioners to discuss problems and potential interventions with the other members of their team and to get input from other stakeholders as necessary (especially, the developers of the product). This is especially important if the practitioner has observed problems but is uncertain as to the appropriate level of description of the problem.

For example, suppose that you have written a problem description about a missing Help button in a

software application. This could be a problem with the overall design of the software or might be a problem isolated to one screen. You might be able to determine this by inspecting other screens in the software, but it could be faster to check with one of the developers.

The first recommendations to consider should be for interventions that will have the widest impact on the product. "Global changes affect everything and need to be considered first" (Rubin, 1994, p. 285). After addressing global problems, continue working through the problem list until there is at least one recommendation for each problem. For each problem, start with interventions that would eliminate the problem, then follow, if necessary, with other less drastic (less expensive, more likely to be implemented) interventions that would reduce the severity of the remaining usability problem. When different interventions involve different trade-offs, it is important to communicate this clearly in the recommendations. This approach can lead to two tiers of recommendations: those that will happen for the version of the product currently under development (short-term) and those that will happen for a future version of the product (long-term).

Molich et al. (2007) used results from CUE-4 to develop guidelines for making usability recommendations useful and usable. By their assessment, only 14 of 84 studied comments (17%) were both useful and usable. To address the weaknesses observed in the recommendations, they concluded:

- Communicate clearly at the conceptual level.
- Ensure that recommendations improve overall usability.
- Be aware of business or technical constraints.
- Solve the whole problem, not just a special case.

Nørgaard and Hornbæk (2009) conducted an exploratory study in which three developers assessed 40 usability findings presented using five feedback formats. The developers rated redesign proposals, multimedia presentations, and screen dumps as useful inputs, problem lists second, and scenarios as least helpful. "Problem lists seem best suited for communicating simple and uncontroversial usability problems for which no contextual information is needed" (p. 64). The preferred feedback formats provided strong contextual information. These results suggest that problem lists can be useful, but it is important to provide sufficient contextual information, if not possible through verbal description, then through associated redesign proposals, screen dumps, and multimedia presentations.

### 2.8.3   Prioritizing Problems

Because usability tests can reveal more problems than there are resources to address, it is important to have some means for prioritizing problems, keeping in mind that design process considerations (stage of development and cost-effectiveness) can also influence the specific usability changes made to a product (Hertzum, 2006). There are two approaches to prioritization that have appeared in the usability testing literature: (1) judgment

driven (Virzi, 1992) and (2) data driven (Lewis et al., 1990; Rubin, 1994; Dumas and Redish, 1999). The bases for judgment-driven prioritizations are the ratings of stakeholders in the project (such as usability practitioners and developers). The bases for data-driven prioritizations are the data associated with the problems, such as frequency, impact, ease of correction, and likelihood of usage of the portion of the product that was in use when the problem occurred. Of these, the most common measurements are frequency and impact (sometimes referred to as severity, although, strictly speaking, severity should include the effect of all of the types of data considered for prioritization). In a study of the two approaches to prioritization, Hassenzahl (2000) found a lack of correspondence between data-driven and judgment-driven severity estimates. This suggests that the preferred approach should be data driven.

The usual method for measuring the frequency of occurrence of a problem is to divide the number of occurrences within participants by the number of participants. A common method (Rubin, 1994; Dumas and Redish, 1999) for assessing the impact of a problem is to assign impact scores according to whether the problem (1) prevents task completion, (2) causes a significant delay or frustration, (3) has a relatively minor effect on task performance, or (4) is a suggestion. This is similar to the scheme of Lewis et al. (1990), in which the impact levels were (1) scenario failure or irretrievable data loss (e.g., the participant required assistance to get past the problem or caused the participant to believe the scenario to be properly completed when it was not), (2) considerable recovery effort (recovery took more than 1 min or the participant repeatedly experienced the problem within a scenario), (3) minor recovery effort (the problem occurred only once within a scenario with recovery time at or under 1 min), or (4) inefficiency (a problem not meeting any of the other criteria).

When considering multiple types of data in a prioritization process, it is necessary to combine the data in some way. A graphical approach is to create a problem grid with frequency on one axis and impact on the other. High-frequency, high-impact problems would receive treatment before low-frequency, low-impact problems. The relative treatment of high-frequency, low-impact problems and low-frequency, high-impact problems depends on practitioner judgment.

An alternative approach is to combine the data arithmetically. Rubin (1994) described a procedure for combining four levels of impact (using the criteria described above with 4 assigned to the most serious level) with four levels of frequency (4: frequency $\geq$ 90%; 3: 51–89%; 2: 11–50%; 1: $\leq$ 10%) by adding the scores. For example, if a problem had an observed frequency of occurrence of 80% and had a minor effect on performance, its priority would be 5 (a frequency rating of 3 plus an impact rating of 2). With this approach, priority scores can range from a low of 2 to a high of 8. If information is available about the likelihood that a user would work with the part of the product that enables the problem, this information would be used to adjust the frequency rating. Continuing the example, if the expectation is that only 10% of users would encounter the

problem, the priority would be 3 (a frequency rating of 1 for the 10% × 80%, or an 8% likelihood of occurrence plus an impact rating of 2).

A similar strategy is to multiply the observed percentage frequency of occurrence by the impact score. The range of priorities depends on the values assigned to each impact level. Assigning 10 to the most serious impact level leads to a maximum priority (severity) score of 1000 (which can optionally be divided by 10 to create a scale that ranges from 1 to 100). Appropriate values for the remaining three impact categories depend on practitioner judgment, but a reasonable set is 5, 3, and 1. Using those values, the problem with an observed frequency of occurrence of 80% and a minor effect on performance would have a priority of 24 (80 × 3/10). It is possible to extend this method to account for the likelihood of use using the same procedure as that described by Rubin (1994), which in the example resulted in modifying the frequency measurement from 80 to 8%. Another way to extend the method is to categorize the likelihood of use with a set of categories such as very high likelihood (assigned a score of 10), high likelihood (assigned a score of 5), moderate likelihood (assigned a score of 3), and low likelihood (assigned a score of 1) and multiply all three scores to get the final priority (severity) score (then optionally divide by 100 to create a scale that ranges from 1 to 100). Continuing the previous example with the assumption that the task in which the problem occurred has a high likelihood of occurrence, the problem's priority would be 12 (5 × 240/100). In most cases, applying the different data-driven prioritization schemes to the same set of problems should result in a very similar prioritization (but there has been no research published on this topic).

### 2.8.4 Working with Quantitative Measurements

The most common use of quantitative measurements is to characterize performance and preference variables by computing means, standard deviations, and ideally confidence intervals. Practitioners use these results to compare observed to target measurements when targets are available. When targets are not available, the results can still be informative, for example, for use as future target measurements or as relatively gross diagnostic indicators.

The failure to meet targets is an obvious diagnostic cue. A less obvious cue is an unusually large standard deviation. Landauer (1997) describes a case in which the times to record an order were highly variable. The cause for the excessive variability was that a required phone number was sometimes, but not always, available, which turned out to be an easy problem to fix. Because the means and standard deviations of time scores tend to correlate, one way to detect an unusually large variance is to compute the coefficient of variation by dividing the standard deviation by the mean (Jeff Sauro, personal communication, April 26, 2004) or the normalized performance ratio by dividing the mean by the standard deviation (Moffat, 1990). Large coefficients of variation (or, correspondingly, small normalized

performance ratios) are potentially indicative of the presence of usability problems.

## 3  STATISTICAL TOPICS

This section covers statistical topics in usability testing, including sample size estimation for problem discovery and measurement tests (both comparative and parameter estimation), confidence intervals based on $t$-scores and binomial confidence intervals, and standardized usability questionnaires. This chapter contains a considerable amount of information about statistical topics because statistical methods do not typically receive much attention in chapters on usability testing, and properly practiced, these techniques can be very valuable. On the other hand, practitioners should keep in mind that the most important factors that lead to successful usability evaluation are the appropriate selection of participants and tasks. No statistical analysis can repair a study in which you watch the wrong people doing the wrong activities.

### 3.1  Sample Size Estimation

The purpose of this section is to discuss the principles of sample size estimation for three types of usability test: population parameter estimation, comparative (also referred to as experimental), and problem discovery (also referred to as diagnostic, observational, or formative). This section assumes some knowledge of introductory applied statistics, so if you're not comfortable with terms such as *mean, variance, standard deviation*, $p$, $t$, and $Z$, refer to an introductory statistics text such as Walpole (1976) for definitions of these and other fundamental terms.

Sample size estimation requires a blend of mathematics and judgment. The computations are straightforward, and it is possible to make reasoned judgments (e.g., judgments about expected costs and precision requirements) for those values that the mathematics cannot determine.

#### 3.1.1  Sample Size Estimation for Parameter Estimation and Comparative Studies

Traditional sample size estimation for population parameter estimation and comparative studies depends on having an estimate of the variance of the dependent measure(s) of interest and an idea of how precise (the magnitude of the critical difference and the statistical confidence level) the measurement must be (Walpole, 1976). Once you have that, the rest is mathematical mechanics (typically, using the formula for the $t$-statistic).

You can (1) get an estimate of variance from previous studies using the same method (same or similar tasks and measures), (2) run a quick pilot study to get the estimate (e.g., piloting with four participants should suffice to provide an initial estimate of variability), or (3) set the critical difference you are trying to detect to some fraction of the standard deviation (Diamond, 1981). (See the following examples for more details about these different methods.)

Certainly, people prefer precise measurement to imprecise measurement, but all other things being equal, the more precise a measurement is, the more it will cost, and running more participants than necessary is wasteful of resources (Kraemer and Thiemann, 1987). The process of carrying out sample size estimation can lead usability practitioners and their management to a realistic determination of how much precision they really need to make their required decisions.

Alreck and Settle (1985) recommend using a "what if" approach to help decision makers determine their required precision. Start by asking the decision maker what would happen if the average value from the study was off the true value by 1%. Usually, the response would be that a difference that small would not matter. Then ask what would happen if the measurement were off by 5%. Continue until you determine the magnitude of the critical difference. Then start the process again, this time pinning down the required level of statistical confidence. Note that statistically unsophisticated decision makers are likely to start out by expecting 100% confidence (which is only possible by sampling every unit in the population). Presenting them with the sample sizes required to achieve different levels of confidence can help them settle in on a more realistic confidence level.

***Example 1: Parameter Estimation Given Estimate of Variability and Realistic Criteria*** This example illustrates the process of computing the sample size requirement for the estimation of a population parameter given an existing estimate of variability and realistic measurement criteria. For speech recognition applications, the recognition accuracy is an important value to track due to the adverse effects misrecognitions have on product usability. Thus, part of the process of evaluating the usability of a speech recognition product is estimating its accuracy. For this example, suppose that:

- Recognition variability (variance) from a previous similar evaluation: 6.35
- Critical difference ($d$): 2.5%
- Desired level of confidence: 90%.

The appropriate procedure for estimating a population parameter is to construct a confidence interval (Bradley, 1976). To determine the upper and lower limits of a confidence interval, add to and subtract the following from the observed mean:

$$d = \text{SEM} \times t_{\text{crit}} \tag{1}$$

where SEM is the standard error of the mean (the standard deviation, $S$, divided by the square root of the sample size, $n$) and $t_{\text{crit}}$ is the $t$-value associated with the desired level of confidence (found in a $t$-table, available in most statistics texts). Setting the critical difference to 2.5 is the same as saying that the value of SEM $\times$ $t_{\text{crit}}$ should be equal to 2.5. In other words, you do not want the upper or lower bound of the confidence interval to be more than 2.5 percentage points away from the observed mean, for a confidence interval width equal to 5.0.

Calculating the SEM depends on knowing the sample size, and the value of $t_{\text{crit}}$ also depends on the sample size, but you do not know the sample size yet. Iterate using the following method.

1. Start with the $Z$-score for the desired level of confidence in place of $t_{\text{crit}}$. For 90% confidence, this is 1.645. (By the way, if you actually *know* the true variability for the measurement rather than just having an estimate, you are done at this point because it is appropriate to use the $Z$-score rather than a $t$-score. However, you almost never know the true variability but must work with estimates.)

2. Algebraic manipulations based on the formula SEM $\times Z = d$ results in $n = Z^2 S^2/d^2$, which for this example is $n = (1.645^2)(6.35)/2.5^2$, which equals 2.7. Always round sample size estimates up to the next whole number, so this initial estimate is 3.

3. Now you need to adjust the estimate by replacing the $Z$-score with the $t$-score for a sample size of 3. For this estimate, the degrees of freedom (df) to use when looking up the value in a $t$ table is $n - 1$, or 2. This is important because the value of $Z$ will always be smaller than the appropriate value of $t$, making the initial estimate smaller than it should be. For this example, $t_{\text{crit}}$ is 2.92.

4. Recalculating for $n$ using 2.92 in place of 1.645 produces 8.66, which rounds up to 9.

5. Because the appropriate value of $t_{\text{crit}}$ is now a little smaller than 2.92 (because the estimated sample size is now larger, with $9 - 1$, or 8, degrees of freedom), recalculate $n$ again, using $t_{\text{crit}}$ equal to 1.86. The new value for $n$ is 3.515, which rounds up to 4.

6. Stop iterating when you get the same value for $n$ on two iterations or you begin cycling between two values for $n$, in which case you should choose the larger value. Table 2 shows the full set of iterations for this example, which ends by estimating the appropriate sample size as 5. Note that there is nothing in these computations that makes reference to the size of the population. Unless the size of the sample is a significant percentage of the total population under study (which is rare but correctable using a finite population correction), the size of the population is irrelevant. Alreck and Settle (1985) explain this with a soup-tasting analogy. Suppose that you are cooking soup in a one-quart saucepan and want to test if it is hot enough. You would stir it thoroughly, then taste a teaspoonful. If it were a two-quart saucepan, you would follow the same procedure—stir thoroughly, then taste a teaspoonful.

Diamond (1981) points out that you can usually get by with an initial estimate and one iteration because most researchers do not mind having a sample size that

**Table 2  Full Set of Iterations for Example 1**

|                 | Initial | 1     | 2     | 3     | 4     | 5     |
|-----------------|---------|-------|-------|-------|-------|-------|
| $t_{crit}$      | 1.645   | 2.92  | 1.86  | 2.353 | 2.015 | 2.132 |
| $t_{crit}^2$    | 2.71    | 8.53  | 3.46  | 5.54  | 4.06  | 4.55  |
| $S^2$           | 6.35    | 6.35  | 6.35  | 6.35  | 6.35  | 6.35  |
| $d$             | 2.5     | 2.5   | 2.5   | 2.5   | 2.5   | 2.5   |
| Estimated $n$   | 2.749   | 8.663 | 3.515 | 5.625 | 4.125 | 4.618 |
| Rounded up      | 3       | 9     | 4     | 6     | 5     | 5     |
| df              | 2       | 8     | 3     | 5     | 4     | 4     |

is a little larger than necessary. If the cost of each sample is high, though, it makes sense to iterate until reaching one of the stopping criteria. Note that the initial estimate establishes the lower bound for the sample size (3 in this example), and the first iteration establishes the upper bound (9 in this example).

**Example 2: Parameter Estimation Given Estimate of Variability and Unrealistic Criteria**  The measurement criteria in Example 1 were reasonable—90% confidence that the interval (limited to a total length of 5%) contains the true mean. This example shows what happens when the measurement criteria are less realistic, illustrating the potential cost associated with high confidence and high measurement precision. Suppose that the measurement criteria for the situation described in Example 1 were less realistic, with:

- Recognition variability from a previous similar evaluation: 6.35
- Critical difference ($d$): 0.5%
- Desired level of confidence: 99%

In that case, the initial $Z$-score would be 2.576, and the initial estimate of $n$ would be

$$n = \frac{(2.576^2)(6.35)}{0.5^2} = 168.549 \qquad (2)$$

which rounds up to 169. Recalculating $n$ with $t_{crit}$ equal to 2.605 ($t$ with 168 degrees of freedom) results in $n$ equal to 172.37, which rounds up to 173. (Rather than continuing to iterate, note that the final value for the sample size must lie between 169 and 173.) There might be some industrial environments in which usability investigators would consider 169–173 participants a reasonable and practical sample size, but they are rare. (On the other hand, collecting data from this number of participants or more in a mailed survey is common.)

**Example 3: Parameter Estimation Given No Estimate of Variability**  For both Examples 1 and 2, it does not matter if the estimate of variability came from a previous study or a quick pilot study. Suppose, however, that you do not have any idea what the measurement variability is, and it is too expensive to run a pilot study to get an initial estimate. This example

illustrates a technique (from Diamond, 1981) for getting around this problem. To do this, though, you need to give up the definition of the critical difference ($d$) in terms of the variable of interest and replace it with a definition in terms of a fraction of the standard deviation.

In this example, the measurement variance is unknown. To get started, the testers have decided that with 90% confidence they do not want $d$ to exceed half the value of the standard deviation. The measurement criteria are:

- Recognition variability from a previous similar evaluation: N/A
- Critical difference ($d$): 0.5$S$
- Desired level of confidence: 90%

The initial sample size estimate is

$$n = \frac{(1.645^2)(S^2)}{(0.5S^2)} = \frac{1.645^2}{0.5^2} = 10.824 \qquad (3)$$

which rounds up to 11. The result of the first iteration, replacing 1.645 with $t_{crit}$ for 10 degrees of freedom (1.812), results in a sample size estimation of 13.13, which rounds up to 14. The appropriate sample size is therefore somewhere between 11 and 14, with the final estimate determined by completing the full set of iterations.

**Example 4: Comparing a Parameter to a Criterion**  For an example comparing a measured parameter to a criterion value, suppose that you have a product requirement that installation should take no more than 30 min. In a preliminary evaluation, participants needed an average of 45 min to complete installation. Development has fixed a number of usability problems found in that preliminary study, so you are ready to measure installation time again using the following measurement criteria:

- Performance variability from the previous evaluation: 10.0
- Critical difference ($d$): 3 min
- Desired level of confidence: 90%

The interpretation of these measurement criteria is that you want to be 90% confident that you can detect

**Table 3 Full Set of Iterations for Example 4**

|  | Initial | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $t_{crit}$ | 1.645 | 2.353 | 1.943 | 2.132 | 2.015 |
| $t_{crit}{}^2$ | 2.706 | 5.537 | 3.775 | 4.545 | 4.060 |
| $s^2$ | 10 | 10 | 10 | 10 | 10 |
| $d$ | 3 | 3 | 3 | 3 | 3 |
| $d^2$ | 9 | 9 | 9 | 9 | 9 |
| Estimated $n$ | 3.007 | 6.152 | 4.195 | 5.050 | 4.511 |
| Rounded up | 4 | 7 | 5 | 6 | 5 |
| df | 3 | 6 | 4 | 5 | 4 |

a difference as small as 3 min between the mean of the data gathered in the test and the criterion you are trying to beat. In other words, the installation will pass if the observed mean time is 27 min or less, because the sample size should guarantee an upper limit to the confidence interval that is no more than 3 min above the mean (as long as the observed variance is less than or equal to the initial estimate of the variance). The procedure for determining the sample size in this situation is the same as that of Example 1, shown in Table 3. The outcome of these iterations is a sample size requirement of 6 because the sample size estimates begin cycling between 5 and 6. Because you only care if you beat the criterion, you could perform a one-sided evaluation (Sauro and Lewis, 2012). For the same measurement criteria (but one sided), the initial value of $t_{crit}$ would be 1.282 and the recommended minimum sample size would be 4.

### Example 5: Sample Size for a Paired t-Test

When you obtain two comparable measurements from each participant in a test (a within-subjects design), you can assess the results using a paired $t$-test. Another name for a *paired t-test* is a *difference score t-test*, because the measurements of concern are the mean and standard deviation of the set of difference scores rather than the raw scores. Suppose that you plan to obtain recognition accuracy scores from participants who have dictated test texts into your product under development and a competitor's product [following all the appropriate experimental design procedures such as counterbalancing the order of presentation of products to participants; see a text such as Myers (1979) for guidance in experimental design], using the following criteria:

- Difference score variability from a previous evaluation: 5.0
- Critical difference ($d$): 2%
- Desired level of confidence: 90%

This situation is similar to that of Example 4 because the typical goal of a difference score $t$-test is to determine if the average difference between scores is statistically significantly different from zero. Thus, the usability criterion in this case is zero, and you want to be 90% confident that if the true difference between system accuracies is 2% or more, you will be able to detect it because the confidence interval for the difference scores will not contain zero. Table 4 shows the iterations for this situation, leading to a sample size estimate of 6.

### Example 6: Sample Size for a Two-Groups t-Test

Up to this point, the examples have all involved one group of scores and have been amenable to similar treatment. If you have a situation in which you plan to compare scores from two independent groups, things get a little more complicated. For one thing, you now have two sample sizes to consider, one for each group.

To simplify things in this example, assume that the groups are essentially equal (especially with regard to performance variability), which should be the case if the groups contain participants from a single population who have received random assignment to treatment conditions. In this case it is reasonable to believe that the sample size for both groups will be equal, which simplifies things. For this situation, the formula for the initial estimate of the sample size for each group is

$$n = \frac{2Z^2 S^2}{d^2} \qquad (4)$$

**Table 4 Full Set of Iterations for Example 5**

|  | Initial | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $t_{crit}$ | 1.645 | 2.353 | 1.943 | 2.132 | 2.015 |
| $t_{crit}{}^2$ | 2.706 | 5.537 | 3.775 | 4.545 | 4.060 |
| $s^2$ | 5 | 5 | 5 | 5 | 5 |
| $d$ | 2 | 2 | 2 | 2 | 2 |
| $d^2$ | 4 | 4 | 4 | 4 | 4 |
| Estimated $n$ | 3.383 | 6.921 | 4.719 | 5.682 | 5.075 |
| Rounded up | 4 | 7 | 5 | 6 | 6 |
| df | 3 | 6 | 4 | 5 | 5 |

**Table 5  Full Set of Iterations for Example 6**

|                | Initial | 1     | 2     | 3     |
| -------------- | ------- | ----- | ----- | ----- |
| $t_{crit}$     | 1.645   | 1.943 | 1.833 | 1.86  |
| $t_{crit}^2$   | 2.706   | 3.775 | 3.360 | 3.460 |
| $s^2$          | 5       | 5     | 5     | 5     |
| $d$            | 2       | 2     | 2     | 2     |
| $d^2$          | 4       | 4     | 4     | 4     |
| Estimated $n$  | 6.765   | 9.438 | 8.400 | 8.649 |
| Rounded up     | 7       | 10    | 9     | 9     |
| df             | 6       | 9     | 8     | 8     |

Note that this is similar to the formula presented in Example 1, with the numerator multiplied by 2. After getting the initial estimate, begin iterating using the appropriate value for $t_{crit}$ in place of $Z$. For example, suppose that you needed to conduct the experiment described in Example 5 with independent groups of participants, keeping the measurement criteria the same:

- Estimate of variability from a previous evaluation: 5.0
- Critical difference ($d$): 2%
- Desired level of confidence: 90%

In that case, as shown in Table 5, iterations would converge on a sample size of nine participants per group, for a total sample size of 18.

This illustrates the well-known measurement efficiency of experiments that produce difference scores (within-subjects designs) relative to experiments involving independent groups (between-subjects designs). For the same measurement precision, the estimated sample size for Example 5 was six participants, one-third the sample size requirement estimated for this example.

Doing this type of analysis gets more complicated if you have reason to believe that the groups are different, especially with regard to variability of performance. In that case you would want to have a larger sample size for the group with greater performance variability in an attempt to obtain more equal precision of measurement for each group. Advanced market research texts (such as Brown, 1980) provide sample size formulas for these situations.

***Example 7: Making Power Explicit in the Sample Size Formula***  The power of a procedure is not an issue when estimating the value of a parameter, but it is an issue when testing a hypothesis (as in Example 6). In traditional hypothesis testing, there is a null ($H_0$) and an alternative ($H_a$) hypothesis. The typical null hypothesis is that there is no difference between groups, and the typical alternative hypothesis is that the difference is greater than zero. When the alternative hypothesis is that the difference is nonzero, the test is two-tailed because you can reject the null hypothesis with either a sufficiently positive or a sufficiently negative outcome. If you have reason to believe that you can predict the direction of the outcome, or if an outcome in only one

**Table 6  Possible Outcomes of a Hypothesis Test**

|                                      | Reality          |                |
| ------------------------------------ | ---------------- | -------------- |
| Decision                             | $H_0$ Is True    | $H_0$ Is False |
| Insufficient evidence to reject $H_0$ | Fail to reject $H_0$ | Type II error  |
| Sufficient evidence to reject $H_0$  | Type I error     | Reject $H_0$   |

direction is meaningful, you can construct an alternative hypothesis that considers only a sufficiently positive or a sufficiently negative outcome (a one-tailed test). For more information, see an introductory statistics text (such as Walpole, 1976).

When you test a hypothesis (e.g., that the difference in recognition accuracy between two competitive dictation products is nonzero), there are two ways to make a correct decision and two ways to be wrong, as shown in Table 6. Strictly speaking, you never accept the null hypothesis, because the failure to acquire sufficient evidence to reject the null hypothesis could be due to (1) no significant difference between groups or (2) a sample size too small to detect an existing difference. Rather than accepting the null hypothesis, you fail to reject it.

Returning to Table 6, the two ways to be right are (1) to fail to reject the null hypothesis ($H_0$) when it is true or (2) to reject the null hypothesis when it is false. The two ways to be wrong are (1) to reject the null hypothesis when it is true (type I error) or (2) to fail to reject the null hypothesis when it is false (type II error). Table 7 shows the relationship between these concepts and their corresponding statistical testing terms.

The formula presented in Example 6 for an initial sample size estimate was

$$n = \frac{2Z^2 S^2}{d^2} \tag{5}$$

In Example 6, the $Z$-score was set for 90% confidence (which means that $\alpha = 0.10$). To take power into account in this formula, you need to add another $Z$-score to the formula, the $Z$-score associated with the desired power of the test (as defined in Table 7). Thus, the formula becomes

$$n = \frac{2(Z_\alpha + Z_\beta)^2 S^2}{d^2} \tag{6}$$

Note that you should always use the one-sided value for $z_\beta$, regardless of whether you are conducting a

**Table 7  Statistical Testing Terms**

| Statistical Concept                          | Testing Term |
| -------------------------------------------- | ------------ |
| Acceptable probability of a type I error     | $\alpha$     |
| Acceptable probability of a type II error    | $\beta$      |
| Confidence                                   | $1 - \alpha$ |
| Power                                        | $1 - \beta$  |

**Table 8  Full Set of Iterations for Example 7**

|               | Initial | 1     | 2     | 3     |
|---------------|---------|-------|-------|-------|
| $t(\alpha)$   | 1.645   | 1.753 | 1.740 | 1.746 |
| $t(\beta)$    | 0.842   | 0.866 | 0.863 | 0.865 |
| $t$(total)    | 2.487   | 2.619 | 2.603 | 2.611 |
| $t$(total)$^2$| 6.18    | 6.86  | 6.78  | 6.81  |
| $S^2$         | 5       | 5     | 5     | 5     |
| $d$           | 2       | 2     | 2     | 2     |
| $d^2$         | 4       | 4     | 4     | 4     |
| Estimated $n$ | 15.5    | 17.2  | 16.9  | 17.04 |
| Rounded up    | 16      | 18    | 17    | 18    |
| df            | 15      | 17    | 16    | 17    |

one- or two-sided test (Diamond, 1981). So, what was the value for power in Example 6? When beta ($\beta$) equals 0.5 (in other words, when the power is 50%), the one-sided value of $z_\beta$ is zero, so $z_\beta$ disappears from the formula. Thus, in this example the implicit power was 50%. Suppose that you want to increase the power of the test to 80%, reducing $\beta$ to 0.2.

- Estimate of variability from a previous evaluation: 5.0
- Critical difference ($d$): 2
- Desired level of confidence: 90% (two-sided $Z_\alpha = 1.645$)
- Desired power: 80% (one-sided $Z_\beta = 0.842$)

With this change, the iterations indicate a sample size of 18 participants per group, for a total sample size of 36, as shown in Table 8. To achieve the stated goal for power results in a considerably larger sample size. Note that the stated power of a test is relative to the critical difference—the smallest effect worth finding. Either increasing the value of the critical difference or reducing the power of a test will result in a smaller required sample size.

### Appropriate Statistical Criteria for Industrial Testing

In scientific publishing, the usual criterion for statistical significance is to set the permissible type I error ($\alpha$) equal to 0.05. This is equivalent to seeking to have 95% confidence that the effect is real rather than random and is focused on controlling the type I error (the likelihood that you decide that an effect is real when it is random). There is no corresponding scientific recommendation for the type II error ($\beta$, the likelihood that you will conclude an effect is random when it is real), although some suggest setting it to 0.20 (Diamond, 1981). The rationale behind the emphasis on controlling the type I error is that it is better to delay the introduction of good information into the scientific database (a type II error) than to let erroneous information in (a type I error).

In industrial evaluation, the appropriate values for type I and II errors depend on the demands of the situation, whether the cost of a type I or II error would be more damaging to the organization. Because we are often resource constrained, especially with regard to

making timely decisions to compete in dynamic marketplaces, this chapter has used measurement criteria (such as 90% confidence rather than 95% confidence and fairly large values for $d$) that seek a greater balance between type I and II errors than is typical in work designed to result in scientific publications. Nielsen (1997) has suggested that 80% confidence is appropriate for practical development purposes. For an excellent discussion of this topic for usability researchers, see Wickens (1998). For other technical issues and perspectives, see Landauer (1997).

Another way to look at the issue is to ask the question, "Am I typically interested in small high-variability effects or large low-variability effects?" The correct answer depends on the situation, but in most usability testing, the emphasis is on the detection of large low-variability effects (either large performance effects or frequently occurring problems). You should not need a large sample to verify the existence of large low-variability effects. Some writers equate sample size with population coverage, but this is not true. A small sample size drawn from the right population provides better coverage than a large sample size drawn from the wrong population. The statistics involved in computing confidence intervals from small samples compensate for the potentially smaller variance in the small sample by forcing the confidence interval to be wider than that for a larger sample (specifically, the value of $t$ is greater when samples are smaller).

Coming from a different tradition than usability research, many market research texts provide rules of thumb recommending large sample sizes. For example, Aaker and Day (1986) recommend a minimum of 100 per group, with 20–50 for subgroups. For national surveys with many subgroup analyses, the typical total sample size is 2500 (Sudman, 1976). These rules of thumb do not make any formal contact with statistical theory and may in fact be excessive, depending on the goals of the study. Other market researchers (e.g., Banks, 1965, p. 252) do promote a careful evaluation of the goals of a study:

> It is urged that instead of a policy of setting uniform requirements for type I and II errors, regardless of the economic consequences of the various decisions to be made from experimental data, a much more flexible approach be adopted. After all, if a researcher sets himself a policy of always choosing the apparently most effective of a group of alternative treatments on the basis of data from unbiased surveys or experiments and pursues this policy consistently, he will find that in the long run he will be better off than if he chose any other policy. This fact would hold even if none of the differences involved were statistically significant according to our usual standards or even at probability levels of 20 or 30 percent.

Finally, Alreck and Settle (1985) provide an excellent summary of the factors indicating appropriate use of large and small samples. Use a large sample size when:

1. Decisions based on the data will have very serious or costly consequences.

2. The sponsors (decision makers) demand a high level of confidence.

3. The important measures have high variance.

4. Analyses will require dividing the total sample into small subsamples.

5. Increasing the sample size has a negligible effect on the cost and timing of the study.

6. Time and resources are available to cover the cost of data collection.

Use a small sample size when:

1. The data will determine few major commitments or decisions.

2. The sponsors (decision makers) require only rough estimates.

3. The important measures have low variance.

4. Analyses will use the entire sample or just a few relatively large subsamples.

5. Costs increase dramatically with sample size.

6. Budget constraints or time limitations limit the amount of data you can collect.

**Tips on Reducing Variance**  Because measurement variance is such an important factor in sample size estimation for these types of studies, it generally makes sense to attempt to manage variance (although in some situations, such management is out of a practitioner's control). Here are some ways to reduce variance:

- Make sure that participants understand what they are supposed to do in the study. Unless potential participant confusion is part of the evaluation (and it sometimes is), it can only add to measurement variance.

- One way to accomplish this is through practice trials that allow participants to get used to the experimental situation without unduly revealing study-relevant information.

- If appropriate, use expert rather than novice participants. Almost by definition, expertise implies reduced performance variability (increased automaticity) (Mayer, 1997). With regard to reducing variance, the farther up the learning curve, the better.

- A corollary of this is that if you need to include both expert and novice users, you should be able to get equal measurement precision for both groups with unequal sample sizes (fewer experts required than novices—which is good, because experts are typically harder than novices to recruit as participants).

- If appropriate, study simple rather than complex tasks.

- Use data transformations for measurements that typically exhibit correlations between means and variances or standard deviations. For example, frequency counts often have proportional means and variances (treated with the square-root transformation), and time scores often have proportional means and standard deviations (treated with the logarithmic transformation) (Myers, 1979; Sauro and Lewis, 2010).

- For comparative studies, use within-subjects designs rather than between-subjects designs whenever possible.

- Keep user groups as homogeneous as possible (but although this reduces variability, it can, simultaneously, pose a threat to a study's external validity if the test group is more homogenous than the population under study) (Campbell and Stanley, 1963).

Keep in mind that it is reasonable to use these tips only when their use does not adversely affect the validity and generalizability of the study. Having a valid and generalizable study is far more important than reducing variability.

**Tips for Estimating Unknown Variance**  Parasuraman (1986) described a method for estimating variability if you have an idea about the largest and smallest values for a population of measurements but do not have the information you need to actually calculate the variability. Estimate the standard deviation (the square root of the variability) by dividing the difference between the largest and smallest values by 6. This technique assumes that the population distribution is normal and then takes advantage of the fact that 99% of a normal distribution will lie in the range of plus or minus three standard deviations of the mean.

Nielsen (1997) surveyed 36 published usability studies and found that the mean standard deviation for measures of expert performance was 33% of the mean value of the usability measure (in other words, if the mean completion time was 100 s, the mean standard deviation was about 33 s). For novice user learning the mean standard deviation was 46%, and for measures of error rates the value was 59%.

Churchill (1991) provided a list of typical variances for data obtained from rating scales. Because the number of items in the scale affects the possible variance (with more items leading to more variance), the table takes the number of items into account. For five-point scales, the typical variance is 1.2–2.0; for seven-point scales it is 2.4–4.0; and for 10-point scales it is 3.0–7.0. Because data obtained using rating scales tends to have a more uniform than normal distribution, he advises using a number nearer the high end of the listed range when estimating sample sizes.

Measurement theorists who agree with Steven's (1951) principle of invariance might yell "foul" at this point because they believe that it is not permissible to calculate averages or variances from rating scale data. There is considerable controversy on this point (see, e.g., Lord, 1953; Nunnally, 1978; Harris, 1985). Data reported by Lewis (1993) indicate that taking averages and conducting $t$-tests on multipoint rating data provides far more interpretable and consistent results than the alternative of taking medians and conducting Mann–Whitney $U$-tests. When you make claims about the meaning of the outcomes of your statistical tests you do have to be careful not to act as if rating scale data are

interval data rather than ordinal data. An average rating of 4 might be better than an average rating of 2, but you cannot claim that it is twice as good (a ratio claim), nor can you claim that the difference between 4 and 2 is equal to the difference between 4 and 6 (an interval claim).

### 3.1.2 Sample Size Estimation for Problem Discovery (Formative) Studies

"Having collected data from a few test subjects—and initially a few are all you need—you are ready for a revision of the text" (Al-Awar et al., 1981, p. 34). "This research does not mean that all of the *possible* problems with a product appear with 5 or 10 participants, but most of the problems that are going to show up with one sample of tasks and one group of participants will occur early" (Dumas, 2003, p. 1098).

Although these types of general guidelines have been helpful, it is possible to use more precise methods to estimate sample size requirements for problem discovery usability tests (Turner et al., 2006). Estimating sample sizes for tests that have the primary purpose of discovering the problems in an interface depends on having an estimate of $p$, characterized as the average likelihood of problem occurrence or, alternatively, the problem discovery rate. As with comparative studies, this estimate can come from previous studies using the same method and similar system under evaluation or can come from a pilot study. For standard scenario-based usability studies, the literature contains large-sample examples that show $p$ ranging from 0.08 to 0.46 (Hwang and Salvendy, 2007, 2009, 2010; Lewis, 1994). For heuristic evaluations, the reported value of $p$ from large-sample studies ranges from 0.08 to 0.60 (Hwang and Salvendy, 2007, 2009, 2010; Nielsen and Molich, 1990). The well-known (and often misused and maligned) guideline that five participants are enough to discover 80% of problems available for discovery is true only when $p$ equals .275. As the reported ranges of $p$ indicate, there will be many studies for which this guideline (or any similar guideline) will not apply, making it important for usability practitioners to obtain estimates of $p$ for their usability studies.

When estimating $p$ from a small sample, it is important to adjust its initially estimated value because a small-sample estimate of $p$ (e.g., fewer than 20 participants) has a bias that results in potentially substantial overestimation of its value (Hertzum and Jacobsen, 2003). A series of Monte Carlo experiments (Lewis, 2001) have demonstrated that a formula combining Good–Turing discounting with a deflation procedure provides a reasonably accurate adjustment of initial estimates of $p$ ($p_{est}$), even when the sample size for that initial estimate has as few as two participants (preferably four participants, though, because the variability of estimates of $p$ is greater for smaller samples) (Lewis, 2001; Faulkner, 2003). This formula for the adjustment of $p$ is

$$P_{adj} = \frac{1}{2}\left[\left(p_{est} - \frac{1}{n}\right)\left(1 - \frac{1}{n}\right)\right] + \frac{1}{2}\left(\frac{p_{est}}{1 + GT_{adj}}\right) \tag{7}$$

where $GT_{adj}$ is the Good–Turing adjustment to probability space (which is the proportion of the number of problems that occurred once divided by the total number of different discovered problems). The $p_{est}/(1 + GT_{adj})$ component in the equation produces the Good–Turing adjusted estimate of $p$ by dividing the observed, unadjusted estimate of $p$ ($p_{est}$) by the Good–Turing adjustment to probability space. The $(p_{est} - 1/n)(1 - 1/n)$ component in the equation produces the deflated estimate of $p$ from the observed, unadjusted estimate of $p$ and $n$ (the sample size used to estimate $p$). The reason for averaging these two different estimates is that the Good–Turing estimator tends to overestimate the true value of $p$, and deflation tends to underestimate it. For more details and experimental data supporting the use of this formula for estimates of $p$ based on sample sizes from 2 to 10 participants, see Lewis (2001).

***Adjusting the Initial Estimate of p***    Because this is a procedure not yet in common use by practitioners, this section contains a detailed illustration of the steps used to adjust an initial estimate of $p$. To start with, organize the problem discovery data in a table (e.g., Table 9) that shows which participants experienced which problems. With four participants and eight observed problems, there are 32 cells in the table. The total number of problem occurrences is 16, so the initial estimate of $p$ ($p_{est}$) is 0.50 (16/32). Note that averaging the proportion of problem occurrence across participants or across problems also equals 0.50.

To apply the Good–Turing adjustment, count the number of problems that occurred with only one participant. In Table 9 this happened for three problems (problems 4, 6, and 8) out of the eight unique problems listed in the table. Thus, the value of $GT_{adj}$ is 0.375 ($\frac{3}{8}$), and the value of $p_{est}/(1 + GT_{adj})$ is 0.36 (0.5/1.375).

**Table 9  Hypothetical Results for a Problem–Discovery Usability Study**

| Participant | Problem | | | | | | | | Count | Proportion |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | | |
| 1 | × | | × | | × | | × | × | 5 | 0.63 |
| 2 | × | × | | | × | | × | | 4 | 0.50 |
| 3 | × | | × | × | × | | | | 4 | 0.50 |
| 4 | × | × | | | | × | | | 3 | 0.38 |
| Count | 4 | 2 | 2 | 1 | 3 | 1 | 2 | 1 | 16 | |
| Proportion | 1.00 | 0.50 | 0.50 | 0.25 | 0.75 | 0.25 | 0.50 | 0.25 | | 0.50 |

To apply the deflation adjustment, start by computing $1/n$, which in Table 9 is 0.25 ($\frac{1}{4}$). The value of ($p_{est}$ − $1/n$)(1 − $1/n$) is 0.19 [(0.25)(0.75)].

The average of the two adjustments produces $p_{adj}$, which in this example equals 0.28 ((0.36 + 0.19)/2). In this example, the adjusted estimate of $p$ is almost half of the initial estimate.

***Using the Adjusted Estimate of p***  Once you have an appropriate (adjusted) estimate for $p$, you can use the formula $1 − (1 − p)^n$ [derivable from both the binomial probability formula (Lewis, 1982, 1994) and the Poisson probability formula (Nielsen and Landauer, 1993)] for various values of $n$ from, say, 1 to 20, to generate the curve of diminishing returns expected as a function of sample size. It is possible to get even more sophisticated, taking into account the fixed and variable costs of the evaluation (especially the variable costs associated with the study of additional participants) to estimate when running an additional participant will result in costs that exceed the value of the additional problems discovered (Lewis, 1994).

The Monte Carlo experiments reported in Lewis (2001) demonstrated that an effective strategy for planning the sample size for a usability study is first to establish a problem discovery goal (e.g., 90% or 95%). Run the first two participants and, based on those results, calculate the adjusted value of $p$ using equation (7). This provides an early indication of the probable sample size required, which might estimate the final sample size exactly or, more likely, underestimate by one or two participants (but will provide an early estimate of the required sample size). Collect data from two more participants (for a total of four). Recalculate

the adjusted estimate of $p$ using equation (7) and project the required sample size using $1−(1−p)^n$. The estimated sample size requirement based on data from four participants will generally be highly accurate, allowing accurate planning for the remainder of the study. Practitioners should do this even if they have calculated a preliminary estimate of the required sample size from an adjusted value for $p$ obtained from a previous study.

Figure 1 shows the discovery rates predicted for problems of differing likelihoods of observation during a usability study. Several independent studies have verified that these types of predictions fit observed data very closely for both usability and heuristic evaluations (Lewis, 1994; Nielsen and Landauer, 1993; Nielsen and Molich, 1990; Virzi, 1990, 1992; Wright and Monk, 1991). Furthermore, the predictions work both for predicting the discovery of individual problems with a given probability of detection and for modeling the discovery of members of sets of problems with a given mean probability of detection (Lewis, 1994). For usability studies, the sample size is the number of participants. For heuristic evaluations, the sample size is the number of evaluators.

Table 10 shows problem detection sample size requirements as a function of problem detection probability and the cumulative likelihood of detecting the problem at least once during the study. The sample size required for detecting the problem twice during a study appears in parentheses. To use this information to establish a usability sample size, you need to determine three things:

1. What is the average likelihood of problem detection probability ($p$)? This plays a role
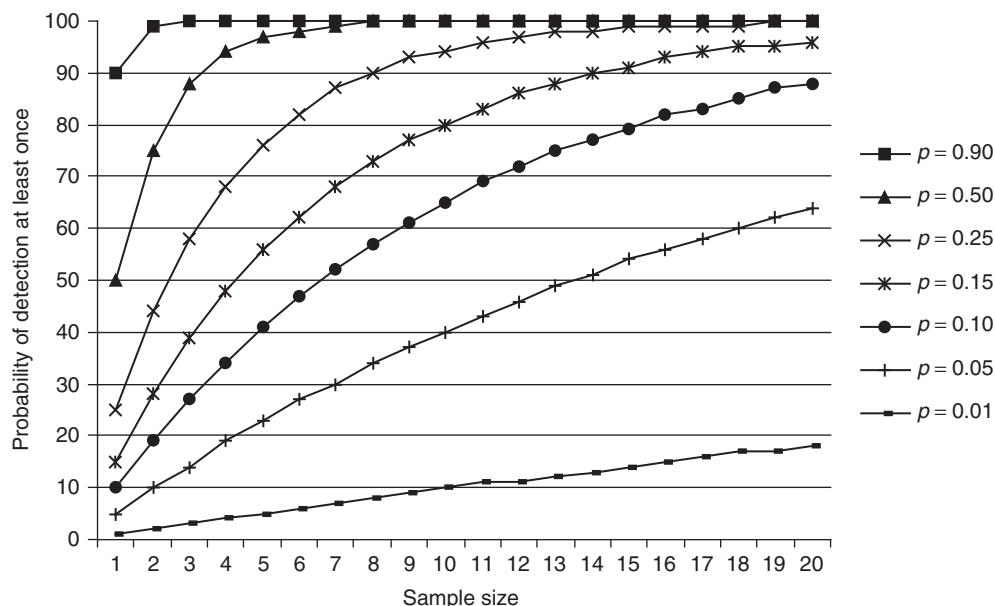


**Figure 1**  Predicted discovery as a function of problem likelihood.

**Table 10  Sample Size Requirements for Problem Discovery (Formative) Studies**

| Problem Occurrence Probability | Cumulative Likelihood of Detecting the Problem at Least Once (Twice) | | | | | |
|---|---|---|---|---|---|---|
| | 0.50 | 0.75 | 0.85 | 0.90 | 0.95 | 0.99 |
| 0.01 | 69 (168) | 138 (269) | 189 (337) | 230 (388) | 299 (473) | 459 (662) |
| 0.05 | 14 (34) | 28 (53) | 37 (67) | 45 (77) | 59 (93) | 90 (130) |
| 0.10 | 7 (17) | 14 (27) | 19 (33) | 22 (38) | 29 (46) | 44 (64) |
| 0.15 | 5 (11) | 9 (18) | 12 (22) | 15 (25) | 19 (30) | 29 (42) |
| 0.25 | 3 (7) | 5 (10) | 7 (13) | 9 (15) | 11 (18) | 17 (24) |
| 0.50 | 1 (3) | 2 (5) | 3 (6) | 4 (7) | 5 (8) | 7 (11) |
| 0.90 | 1 (2) | 1 (2) | 1 (3) | 1 (3) | 2 (3) | 2 (4) |

similar to the role of variance in the previous examples. If you do not know this value (from previous studies or a pilot study), you need to decide on the lowest problem detection probability that you want to (or have the resources to) tackle. The smaller this number, the larger the required sample size.

2.  What proportion of the problems that exist at that level do you need (or have the resources) to discover during the study (in other words, the cumulative likelihood of problem detection)? The larger this number, the larger the required sample size.

3.  Are you willing to take single occurrences of problems seriously or must problems appear at least twice before receiving consideration? Requiring two occurrences results in a larger sample size.

For values of $p$ or problem discovery goals that are outside tabled values, you can use the following formula [derived algebraically from Goal $= 1-(1-p)^n$] to compute the sample size required for a given problem discovery goal (taking single occurrences of problems seriously) and value of $p$:

$$n = \frac{\log(1 - \text{goal})}{\log(1 - p)} \qquad (8)$$

In the example from Table 9, the adjusted value of $p$ was 0.28. Suppose that the practitioner decided that the appropriate problem discovery goal was to find 97% of the discoverable problems. The computed value of $n$ is 10.6 ($\log(0.03)/\log(0.72)$, or $-1.522/-0.143$). The practitioner can either round the sample size up to 11 or adjust the problem discovery goal down to 96.3%-[$1-(1-0.28)^{10}$].

Lewis (1994) created a return-on-investment (ROI) model to investigate appropriate cumulative problem detection goals. It turned out that the appropriate goal depended on the average problem detection probability in the evaluation, the same value that has a key role in determining the sample size. The model indicated that if the expected value of $p$ was small (say, around 0.10), practitioners should plan to discover about 86% of the problems. If the expected value of $p$ was larger

(say, around 0.25 or 0.50), practitioners should plan to discover about 98% of the problems. For expected values of $p$ between 0.10 and 0.25, practitioners should interpolate between 87 and 97% to determine an appropriate goal for the percentage of problems to discover.

The cost of an undiscovered problem had a strong effect on the magnitude of the maximum ROI, but, contrary to expectation, it had only a minor effect on sample size at maximum ROI (Lewis, 1994). Usability practitioners should be aware of these costs in their settings and their effect on ROI (Boehm, 1981), but these costs have relatively little effect on the appropriate sample size for a usability study.

In summary, there is compelling evidence that the law of diminishing returns, based on the cumulative binomial probability formula, applies to problem discovery studies. To use this formula to determine an appropriate sample size, practitioners must form an idea about the expected value of $p$ (the average likelihood of problem detection) for the study and the percentage of problems that the study should uncover. Practitioners can use the ROI model from Lewis (1994) or their own ROI formulas to estimate an appropriate goal for the percentage of problems to discover and can examine data from their own or published usability studies to get an initial estimate of $p$ (which published studies to date indicate can range at least from 0.08 to 0.60). With these two estimates, practitioners can use Table 10 (or, for computations outside tabled values, the appropriate equations) to estimate appropriate sample sizes for their usability studies.

It is interesting to speculate that a new product that has not yet undergone any usability evaluation is likely to have a higher $p$ than an established product that has gone through several development iterations (including usability testing). This suggests that it is easier (takes fewer participants) to improve a completely new product than to improve an existing product (as long as that existing product has benefited from previous usability evaluation). This is related to the idea that usability testing is a hill-climbing procedure, in which the results of a usability test are applied to a product to push its usability up the hill. The higher up the hill you go, the more difficult it becomes to go higher, because you have already weeded out the problems that were easy to find and fix.

Practitioners who wait to see a problem at least twice before giving it serious consideration can see from Table 10 the sample size implications of this strategy. Certainly, all other things being equal, it is more important to correct a problem that occurs frequently than one that occurs infrequently. However, it is unrealistic to assume that the frequency of detection of a problem is the only criterion to consider in the analysis of usability problems. The best strategy is to consider problem frequency and other problem data (such as severity and likelihood of use) simultaneously to determine which problems are most important to correct rather than establishing a cutoff rule such as "fix every problem that appears two or more times."

Note that, in contrast to the results reported by Virzi (1992), the results reported by Lewis (1994) did not indicate any consistent relationship between problem frequency and impact (severity). It is possible that this difference was due to the difference in the methods used to assess severity [judgment driven in Virzi (1992); data driven in Lewis (1994)]. Thus, the safest strategy is for practitioners to assume independence of frequency and impact until further research resolves the discrepancy between the outcomes of these studies.

It is important for practitioners to consider the risks as well as the gains when using small samples for usability studies. Although the diminishing returns for inclusion of additional participants strongly suggest that the most efficient approach is to run a small sample (especially if $p$ is high, if the study will be iterative, and if undiscovered problems will not have dangerous or expensive outcomes), human factors engineers and other usability practitioners must not become complacent regarding the risk of failing to detect low-frequency but important problems.

One could argue that the true number of possible usability problems in any interface is essentially infinite, with an essentially infinite number of problems with nonzero probabilities that are extremely close to zero. For the purposes of determining sample size, the $p$ we are really dealing with is the $p$ that represents the number of discovered problems divided by the number of discoverable problems, where the definition of a discoverable problem is vague but almost certainly constrained by details of the experimental setting, such as the studied scenarios and tasks and the skill of the observer(s). Despite this vagueness and some recent criticism of the use of $p$ to model problem discovery (Caulton, 2001; Schmettow, 2008, 2009; Woolrych and Cockton, 2001), these techniques seem to work reasonably well in practice (Lewis, 2006; Turner et al., 2006).

***Examples of Sample Size Estimation for Problem Discovery (Formative) Studies*** This section contains several examples illustrating the use of Table 10 as an aid in selecting an appropriate sample size for a problem discovery study.

A. Given the following problem discovery criteria:
- Detect problems with an average probability of 0.25
- Minimum number of detections required: 1
- Planned proportion to discover: 0.90

The appropriate sample size is nine participants.

B. Given the same discovery criteria, except that the practitioner requires problems to be detected twice before receiving serious attention:
- Detect problems with an average probability of 0.25
- Minimum number of detections required: 2
- Planned proportion to discover: 0.90

The appropriate sample size would be 15 participants.

C. Returning to requiring a single detection, but increasing the planned proportion to discover to 0.99:
- Detect problems with an average probability of 0.25
- Minimum number of detections required: 1
- Planned proportion to discover: 0.99

The appropriate sample size would be 17 participants.

D. Given the following extremely stringent discovery criteria:
- Detect problems with an average probability of 0.01
- Minimum number of detections required: 1
- Planned proportion to discover: 0.99

The sample size required would be 459 participants (an unrealistic requirement in most settings, implying unrealistic study goals).

Note that there is no requirement to run the entire planned sample through the usability study before reporting clear problems to development and getting those problems fixed before continuing. These required sample sizes are total sample sizes, not sample sizes per iteration. The following testing strategy promotes efficient iterative problem discovery studies and is similar to strategies published by a number of usability specialists (Bailey et al., 1992; Fu et al., 2002; Jeffries and Desurvire, 1992; Kantner and Rosenbaum, 1997; Macleod et al., 1997; Nielsen, 1993; Rosenbaum, 1989).

1. Start with an expert (heuristic) evaluation or one-participant pilot study to uncover the obvious problems. Correct as many of these problems as possible before starting the iterative cycles with step 2. List all unresolved problems and carry them to step 2.

2. Watch a small sample of participants (e.g., three or four) use the system. Record all observed usability problems. Calculate an adjusted estimate of $p$ based on these results and reestimate the required sample size.

3. Redesign based on the problems discovered. Focus on fixing high-frequency and high-impact problems. Fix as many of the remaining problems as possible. Record any outstanding problems so they can remain open for all following iterations.

4.  Continue iterating until you have reached your sample size goal (or must stop for any other reason, such as running out of time).

5.  Record any outstanding problems remaining at the end of testing and carry them over to the next product for which they are applicable.

This strategy blends the benefits of large and small sample studies. During each iteration, you observe only three or four participants before redesigning the system. Therefore, you can quickly identify and correct the most frequent problems (which means that you waste less time watching the next set of participants encounter problems that you already know about). With five iterations, for example, the total sample size would be 15–20 participants. With several iterations you will identify and correct many less frequent problems because you record and track the uncorrected problems through all iterations.

Note that using this sort of iterative procedure affects estimates of $p$ as you go along. The value of $p$ in the system you end with should generally be lower than the $p$ you started with (as long as the process of fixing problems does not create as many other problems). For this reason it is a good idea to recompute the adjusted value of $p$ after each iteration.

### Evaluating Sample Size Effectiveness Given Fixed $n$

Suppose that you know you have time to run only a limited number of participants, are willing to treat a single occurrence of a problem seriously, and want to determine what you can expect to get out of a problem discovery study with that number of participants. If that number were 6, for example, examination of Table 10 indicates:

-   You are almost certain to detect problems that have a 0.90 likelihood of occurrence (it only takes two participants to have a 99% cumulative likelihood of seeing the problem at least once).

-   You are almost certain (between 95 and 99% likely) to detect problems that have a 0.50 likelihood of occurrence (for this likelihood of occurrence, the sample size required at 95% is 5 and at 99% is 7).

-   You have a reasonable chance (about 80% likely) of detecting problems that have a 0.25 likelihood of occurrence (for this likelihood of occurrence, the required sample size at 75% is 5 and at 85% is 7).

-   You have a little better than even odds of detecting problems that have a 0.15 likelihood of occurrence (the required sample size at 50% is 5).

-   You have a little less than even odds of detecting problems that have a 0.10 likelihood of occurrence (the required sample size at 50% is 7).

-   You are not likely to detect many of the problems that have a likelihood of occurrence of 0.05 or 0.01 (for these likelihoods of occurrence, the

sample sizes required at 50% are 14 and 69 respectively).

This analysis illustrates that although a problem discovery study with a sample size of 6 participants will typically not discover problems with very low likelihoods of occurrence, the study is almost certainly worth conducting.

Applying this procedure to a number of different sample sizes produces Table 11. The cells in Table 11 are the probability of having a problem with a specified occurrence probability happen at least once during a usability study with the given sample size $(1-(1-p)^n)$. Practitioners who are uncomfortable with sample size estimation procedures that implicitly assume a fixed number of problems available for discovery (Hornbæk, 2010) or are concerned with unmodeled variability of an averaged estimate of $p$ (Caulton, 2001; Schmettow, 2008; Woolrych and Cockton, 2001) can use Table 11 to plan their formative usability studies without those limitations.

### Estimating the Number of Problems Available for Discovery

Another approach to assessing sample size effectiveness is to estimate the number of undiscovered problems. Returning to the situation illustrated in Table 9, the adjusted estimate of $p$ is 0.28 with four participants and eight unique problems. The estimated proportion of problems discovered with those four participants is 0.73 $[1-(1-0.28)^4]$. If eight problems are about 73% of the total number of problems available for discovery, the total number of problems available for discovery (given the constraints of the testing situation) is about 11 (8/0.73). Thus, there appear to be about three undiscovered problems. With an estimate of only three undiscovered problems, the sample size of 4 is approaching adequacy.

Contrast this with the MACERR study described in Lewis (2001), which had an estimated value of $p$ of 0.16 with 15 participants and 145 unique problems. For this study, the estimated proportion of discovered problems at the end of the test was 0.927 $[1-(1-0.16)^{15}]$. The estimate of the total number of problems available for discovery was about 156 (145/0.927). With about 11 problems remaining available for discovery, it might be wise to run a few more participants.

On the other hand, with an estimated 92.7% of problems available for discovery extracted from the problem discovery space defined by the test conditions, it might make more sense to make changes to the test conditions (in particular, to make reasonable changes to the tasks) to create additional opportunities for problem discovery. This is one of many areas in which practitioners need to exercise professional judgment using the available tables and formulas to guide that judgment.

Note the use of the phrase "problems available for discovery." A given set of tasks and participants defines a pool of potentially discoverable usability problems from the set of all possible usability problems. Even within that restricted pool there will always be uncertainty regarding the "true" number of usability

**Table 11 Likelihood of Discovering Problems of Probability *p* at Least Once in a Study with Sample Size *n***

| *p* | N | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 |
| 0.05 | 0.10 | 0.14 | 0.19 | 0.23 | 0.26 | 0.30 | 0.34 |
| 0.10 | 0.19 | 0.27 | 0.34 | 0.41 | 0.47 | 0.52 | 0.57 |
| 0.15 | 0.28 | 0.39 | 0.48 | 0.56 | 0.62 | 0.68 | 0.73 |
| 0.25 | 0.44 | 0.58 | 0.68 | 0.76 | 0.82 | 0.87 | 0.90 |
| 0.50 | 0.75 | 0.88 | 0.94 | 0.97 | 0.98 | 0.99 | 1.00 |
| 0.90 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| *p* | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 0.01 | 0.09 | 0.10 | 0.10 | 0.11 | 0.12 | 0.13 | 0.14 |
| 0.05 | 0.37 | 0.40 | 0.43 | 0.46 | 0.49 | 0.51 | 0.54 |
| 0.10 | 0.61 | 0.65 | 0.69 | 0.72 | 0.75 | 0.77 | 0.79 |
| 0.15 | 0.77 | 0.80 | 0.83 | 0.86 | 0.88 | 0.90 | 0.91 |
| 0.25 | 0.92 | 0.94 | 0.96 | 0.97 | 0.98 | 0.98 | 0.99 |
| 0.50 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.90 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| *p* | 16 | 17 | 18 | 19 | 20 | 25 | 30 |
| 0.01 | 0.15 | 0.16 | 0.17 | 0.17 | 0.18 | 0.22 | 0.26 |
| 0.05 | 0.56 | 0.58 | 0.60 | 0.62 | 0.64 | 0.72 | 0.79 |
| 0.10 | 0.81 | 0.83 | 0.85 | 0.86 | 0.88 | 0.93 | 0.96 |
| 0.15 | 0.93 | 0.94 | 0.95 | 0.95 | 0.96 | 0.98 | 0.99 |
| 0.25 | 0.99 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.50 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.90 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

problems (Hornbæk, 2010). The technique described in this section is a way to estimate, not to guarantee, the likely number of discoverable problems.

***Some Tips on Managing p*** Because *p* (the average likelihood of problem discovery) is such an important factor in sample size estimation for usability tests, it generally makes sense to attempt to manage it (although in some situations such management is out of a practitioner's control). Here are some ways to increase *p*:

- Use highly skilled observers for usability studies.
- Use multiple observers rather than a single observer (Hertzum and Jacobsen, 2003).
- Focus evaluation on new products with newly designed interfaces rather than older, more refined interfaces.
- Study less skilled participants in usability studies (as long as they are appropriate participants).
- Make the user sample as heterogeneous as possible, within the bounds of the population to which you plan to generalize the results.
- Make the task sample as heterogeneous as possible.
- Emphasize complex rather than simple tasks.
- For heuristic evaluations, use examiners with usability and application-domain expertise (double experts) (Nielsen, 1992).

- For heuristic evaluations, if you must make a trade-off between having a single evaluator spend a lot of time examining an interface versus having more examiners spend less time each examining an interface, choose the latter option (Dumas et al., 1995; Virzi, 1997).

Note that some of the tips for increasing *p* are the opposite of those that reduce measurement variability.

### 3.1.3 Sample Sizes for Nontraditional Areas of Usability Evaluation

Nontraditional areas of usability evaluation include activities such as the evaluation of visual design and marketing materials. As with traditional areas of evaluation, the first step is to determine if the evaluation is comparative/parameter estimation or problem discovery.

Part of the problem with nontraditional areas is that there is less information regarding the values of the variables needed to estimate sample sizes. Another issue is whether these areas are focused inherently on detecting more subtle effects than is the norm in usability testing, which has a focus on large low-variability effects (and correspondingly small sample size requirements). Determining this requires the involvement of someone with domain expertise in these nontraditional areas. It seems, however, that even these nontraditional areas would benefit from focusing on the discovery of large

low-variability effects. Only if there were a business case which held that investment in a study to detect small, highly variable effects would ultimately pay for itself should you conduct such a study.

For example, in *The Survey Research Handbook*, Alreck and Settle (1985) point out that the reason that survey samples rarely contain fewer than several hundred respondents is due to the cost structure of surveys. The fixed costs of the survey include activities such as determining information requirements, identifying survey topics, selecting a data collection method, writing questions, choosing scales, composing the questionnaire, and so on. For this type of research, the additional or marginal cost of including hundreds of additional respondents can be very small relative to the fixed costs. Contrast this with the cost (or feasibility) of adding participants to a usability study in which there might be as little as a week or two between the availability of testable software and the deadline for affecting the product, with resources limiting the observation of participants to one at a time and the test scenarios requiring two days to complete. The potentially high cost of observing participants in usability tests is one reason why usability researchers have devoted considerable attention to sample size estimation, despite some assertions that sample size estimation is relatively unimportant (Wixon, 2003).

"Since the numbers don't know where they came from, they always behave just the same way, regardless" (Lord, 1953, p. 751). What potentially differs for nontraditional areas of usability evaluation is not the behavior of numbers or statistical procedures, but the researchers' goals and economic realities.

### 3.2   Confidence Intervals

A major trend in modern statistical evaluation has been a reduced focus on hypothesis testing and a move toward more informative analyses such as effect sizes and confidence intervals (Landauer, 1997). For most applied usability work, confidence intervals are more useful than effect sizes because they have the same units of measurement as the variables from which they are computed. Even when confidence intervals are very wide, they can still be informative, so practitioners should routinely report confidence intervals for their measurements (Sauro, 2006). Although 95% confidence is a commonly used level, confidence as low as 80% will often be appropriate for applied usability measurements (Nielsen, 1997).

### 3.2.1   Intervals Based on *t*-Scores

Formulas for the computation of confidence intervals based on *t*-scores are algebraically equivalent to those used to estimate required sample sizes for measurement-based usability tests, but isolate the critical difference (*d*) instead of the sample size (*n*):

$$d = \text{SEM} \times t_{\text{crit}} \qquad (9)$$

where SEM is the standard error of the mean (the standard deviation, *S*, divided by the square root of

the sample size, *n*) and $t_{\text{crit}}$ is the *t*-value associated with the desired level of confidence (found in a *t*-table, available in most statistics texts). (Practitioners who are concerned about departures from normality can perform a logarithmic transformation on their raw data before computing the confidence interval, then transform the data back to report the mean and confidence interval limits.)

For example, suppose a task in a usability test with seven participants has an average completion time of 5.4 min with a standard deviation of 2.2 min. The SEM is 0.83 ($2.2/7^{1/2}$). For 90% confidence and 6 ($n-1$) degrees of freedom, the tabled value of *t* is 1.943. The computed value of *d* is 1.6 [(0.83)(1.943)], so the 90% confidence interval is 5.4 ± 1.6 min.

As a second example, suppose that the results of a within-subjects test of the time required for two installation procedures showed that the mean of the difference scores (version A minus version B) was 2 min with a standard deviation of 2 min for a sample size of eight participants. The SEM is 0.71 ($2/8^{1/2}$). For 95% confidence and 7 ($n-1$) degrees of freedom, the tabled value of *t* is 2.365. The computed value of *d* is 1.7 [(0.71)(2.365)], so the 95% confidence interval is 2.0 ± 1.7 min (ranging from 0.3 to 3.7 min). Because the confidence interval does not contain a zero, this interval indicates that with $\alpha$ of 0.05 (where $\alpha$ is 1 minus the confidence expressed as a proportion rather than a percentage) you should reject the null hypothesis of no difference. The evidence indicates that version A takes longer than version B. The major advantage of a confidence interval over a significance test is that you also know with 95% confidence that the magnitude of the difference is probably no less than 0.3 min and no greater than 3.7 min. If the versions are otherwise equal, version B is the clear winner. If the cost of version B is greater than the cost of version A (e.g., due to a need to license a new technology for version B), the decision about which version to implement is more difficult but is certainly aided by having an estimate of the upper and lower limits of the difference between the two versions.

### 3.2.2   Binomial Confidence Intervals

As discussed above, confidence intervals constructed around a mean can be very useful. Many usability measurements, however, are proportions or percentages computed from count data rather than means. For example, the maximum-likelihood estimate of a usability defect rate for a specific problem is the proportion computed by dividing the number of participants who experience the problem divided by the total number of participants ($x/n$). There are other, potentially more accurate ways to estimate completion rates (Lewis and Sauro, 2006), such as the Laplace method [adding 1 to the numerator and denominator before computing the percentage, in other words, $(x + 1)/(n + 1)$], but for practical usability measurement, improving the point estimate of a percentage is less important than computing a binomial confidence interval.

The statistical term for a study designed to estimate proportions is a *binomial experiment*, because a given problem either will or will not occur for each trial

(participant) in the experiment. For example, a participant either will or will not install an option correctly. The point estimate of the defect rate is the observed proportion of failures (*p*). However, the likelihood is very small that the point estimate from a study is exactly the same as the true percentage of failures, especially if the sample size is small (Walpole, 1976). To compensate for this, you can calculate interval estimates that have a known likelihood of containing the true proportion (Steele and Torrie, 1960). You can use these binomial confidence intervals to describe the proportion of usability defects effectively, often with only a small sample (Lewis, 1996a). Cordes and Lentz (1986) and Lewis (1996a) provided BASIC programs for the computation of exact binomial confidence intervals. There are similar programs available at the website of the Southwest Oncology Group Statistical Center (SOGSC, 2004), the GraphPad website (GraphPad, 2004), and the Measuring Usability website (Sauro, 2004).

Some programs (Cordes and Lentz, 1986; Lewis, 1996a; SOGSC, 2004) produce binomial confidence intervals that always contain the exact binomial confidence interval. Other programs (GraphPad, 2004; Sauro, 2004) also produce a new type of interval called *approximate* binomial confidence intervals (Agresti and Coull, 1998). Exact and approximate binomial confidence intervals differ in a number of ways. An exact binomial confidence interval guarantees that the actual confidence is equal to or greater than the nominal confidence. An approximate interval guarantees that the average of the actual confidence in the long run will be equal to the nominal confidence, but for any specific test, the actual confidence could be lower than the nominal confidence. On the other hand, approximate binomial confidence intervals tend to be narrower than exact intervals. When sample sizes are large (*n* > 100), the two types of intervals are virtually indistinguishable. When sample sizes are small, though, there can be a considerable difference in the width of the intervals, especially when the observed proportion is close to 0 or 1. The exact interval often has an actual confidence closer to 99% when the nominal confidence is 95%, making it too conservative.

Monte Carlo studies that have compared exact and approximate binomial confidence intervals using standard statistical distributions (Agresti and Coull, 1998) and data from usability studies (Sauro and Lewis, 2005) generally support the use of approximate rather than exact binomial confidence intervals. When the actual confidence of an approximate binomial confidence interval is below the nominal level, the actual level tends to be close to the nominal level. For example, Agresti and Coull (1998, p. 125) found that the actual level for 95% approximate binomial confidence intervals using the adjusted-Wald method was never less than 89%:

In forming a 95% confidence interval, is it better to use an approach that guarantees that the actual coverage probabilities are *at least* .95 yet typically achieves coverage probabilities of about .98 or .99, or an approach giving narrower intervals for which the actual coverage probability could be less than .95 but

is usually quite *close* to .95? For most applications, we would prefer the latter.

This conclusion, that using approximate binomial confidence intervals will tend to produce superior decisions relative to the use of exact intervals, seems to apply to usability test data (Sauro and Lewis, 2005). If, however, it is critical for a specific test to achieve or exceed the nominal level of confidence, then it is reasonable to use an exact binomial confidence interval.

When using binomial confidence intervals, note that if the failure rate is fairly high, you do not need a very large sample to acquire convincing evidence of failure. In the first evaluation of a wordless graphic instruction (Lewis and Pallo, 1991), 9 of 11 installations (82%) were incorrect. The exact 90% binomial confidence interval for this outcome ranged from 0.53 to 0.97. This interval allowed us to argue that, without intervention, the failure rate for installation would be at least 53% (and more likely closer to the observed 82%).

This suggests that a reasonable strategy for binomial experiments is to start with a small sample size and record the number of failures. From these results, compute a confidence interval. If the lower limit of the confidence interval indicates an unacceptably high failure rate, stop testing. Otherwise, continue testing and evaluating in increments until you reach a specified level of precision or you reach the maximum sample size allowed for the study.

This method can rapidly demonstrate with a small sample that a usability defect is unacceptably high if the criterion is low and the true defect rate is high. Although the confidence interval will be wide (50 percentage points in the graphic symbols example), the lower limit of the interval may be clearly unacceptable. When the true defect rate is low or the criterion is high, this procedure may not work without a large sample size. The decision to continue sampling or to stop the study should be determined by a reasonable business case that balances the cost of continued data collection against the potential cost of allowing defects to go uncorrected.

You cannot use this procedure with small samples to prove that a success rate is acceptably high. With small samples, even if the defect percentage observed is zero or close to 0%, the interval will be wide, so it will probably include defect percentages that are unacceptable. For example, suppose that you have run five participants through a task and all five have completed the task successfully. The 90% confidence interval on the percentage of defects for these results ranges from 0 to 45%, with a 45% defect rate almost certainly unacceptable. If you had 50 out of 50 successful task completions, the 90% binomial confidence interval would range from 0 to 6%, which would indicate a greater likelihood of the true defect rate being close to 0%. The moral of the story is that it is relatively easy to prove (requires a small sample) that a product is unacceptable, but it is difficult to prove (requires a large sample) that a product is acceptable.

## 3.3 Standardized Usability Questionnaires

Standardized satisfaction measures offer many advantages to the usability practitioner. Specifically,

standardized measurements provide objectivity, replicability, quantification, economy, communication, and scientific generalization (Nunnally, 1978). Comparisons of the reliability of standardized versus ad hoc (home-grown) usability questionnaires consistently favor the use of standardized instruments (Hornbæk, 2006; Hornbæk and Law, 2007; Sauro and Lewis, 2009). The first published standardized usability questionnaires appeared in the late 1980s (Chin et al., 1988; Kirakowski and Dillon, 1988). Questionnaires focused on the measurement of computer satisfaction preceded these questionnaires (e.g., the Gallagher Value of MIS Reports Scale and the Hatcher and Diebert Computer Acceptance Scale) (see LaLomia and Sidowski, 1990, for a review), but these questionnaires were not applicable to scenario-based usability tests.

The most widely used standardized usability questionnaires are the QUIS (Chin et al., 1988), the SUMI (Kirakowski and Corbett, 1993; Kirakowski, 1996), the PSSUQ (Lewis, 1992, 1995, 2002), and the SUS (Brooke, 1996). The most common application of these questionnaires is at the end of a test (after completing a series of test scenarios). The ASQ (Lewis, 1991b) is a short three-item questionnaire designed for administration immediately following the completion of a test scenario. The ASQ takes less than a minute to complete. The longer standard questionnaires typically have completion times of less than 10 min (Dumas, 2003).

The primary measures of standardized questionnaire quality are reliability (consistency of measurement) and validity (measurement of the intended attribute) (Nunnally, 1978). There are several ways to assess reliability, including test–retest and split-half reliability. The most common method for the assessment of reliability is coefficient $\alpha$, a measurement of internal consistency. Coefficient $\alpha$ can range from 0 (no reliability) to 1 (perfect reliability). Measures that can affect a person's future, such as IQ tests or college entrance exams, should have a minimum reliability of 0.90 (preferably, reliability greater than 0.95). For other research or evaluation, measurement reliability in the range of 0.70–0.80 is acceptable (Nunnally, 1978; Landauer, 1997).

A questionnaire's validity is the extent to which it measures what it claims to measure. Researchers commonly use the Pearson correlation coefficient to assess criterion-related validity (the relationship between the measure of interest and a different concurrent or predictive measure). These correlations do not have to be large to provide evidence of validity. For example, personnel selection instruments with validities as low as 0.30 or 0.40 can be large enough to justify their use (Nunnally, 1978). Another approach to validity is content validity, typically assessed through the use of factor analysis (which also helps questionnaire developers discover or confirm clusters of related items that can form reasonable subscales).

Regarding the appropriate number of scale steps, it is true that more scale steps are better than fewer scales steps, but with rapidly diminishing returns. The reliability of individual items is a monotonically increasing function of the number of steps (Nunnally, 1978). As the number of scale steps increase from 2 to 20, the increase in reliability is very rapid at first but tends to level off at about 7. After 11 steps there is little gain in reliability from increasing the number. The number of steps in an item is very important for measurements based on a single item but is less important when computing measurements over a number of items (as in the computation of an overall or subscale score).

### 3.3.1 QUIS

The QUIS (Shneiderman, 1987; Chin et al., 1988; see also http://lap.umd.edu/QUIS/) is a product of the Human–Computer Interaction Lab at the University of Maryland. Its use requires the purchase of a license. Chin et al. (1988) evaluated several early versions of the QUIS (Versions 3–5). They reported an overall reliability (coefficient $\alpha$) of 0.94 but did not report any subscale reliability.

The QUIS is currently at Version 7. This version includes demographic questions, an overall measure of system satisfaction, and 11 specific interface factors. The QUIS is available in two lengths, short (26 items) and long (71 items). The items are nine-point scales anchored with opposing adjective phrases (such as "confusing" and "clear" for the item "messages which appear on screen").

### 3.3.2 CUSI and SUMI

The Human Factors Research Group (HFRG) at University College Cork published their first standardized questionnaire, the Computer Usability Satisfaction Inventory (CUSI), in 1988 (Kirakowski and Dillon, 1988). The CUSI was a 22-item questionnaire containing two subscales: affect and competence. Its overall reliability was 0.94, with 0.91 for affect and 0.89 for competence.

The HFRG replaced the CUSI with the SUMI (Kirakowski and Corbett, 1993; Kirakowski, 1996), a questionnaire that has six subscales: global, efficiency, affect, helpfulness, control, and learnability. Its 50 items are statements (such as "The instructions and prompts are helpful") to which participants indicate that they agree, are undecided, or disagree. The SUMI has undergone a significant amount of psychometric development and evaluation to arrive at its current form. The results of studies that included significant main effects of system, SUMI scales, and their interaction support its validity (McSweeney, 1992; Wiethoff et al., 1992). The reported reliabilities of the six subscales (measured with coefficient $\alpha$) are:

- Global: 0.92
- Efficiency: 0.81
- Affect: 0.85
- Helpfulness: 0.83
- Control: 0.71
- Learnability: 0.82

One of the greatest strengths of the SUMI is the database of results that is available for the construction of interpretive norms. This makes it possible for practitioners to compare their results with those of similar products and tasks [as long as there are similar products and tasks in the database; Cavallin et al. (2007) reported a significant effect of tasks on SUMI scores]. Another strength is that the SUMI is available in different

languages (such as UK English, American English, Italian, Spanish, French, German, Dutch, Greek, and Swedish). Like the QUIS, practitioners planning to use SUMI must purchase a license for its use (which includes questionnaires and scoring software). For an additional fee, a trained psychometrician at the HFRG will score the results and produce a report.

### 3.3.3  SUS

Usability practitioners at Digital Equipment Corporation (DEC) developed the SUS in the mid-1980s (Dumas, 2003). The 10 five-point items of the SUS provide a unidimensional (no subscales) usability measurement that ranges from 0 to 100. In the first published account of the SUS, Brooke (1996) stated that the SUS was robust, reliable, and valid but did not publish the specific reliability or validity measurements. With regard to validity, "it correlates well with other subjective measures of usability (e.g., the general usability subscale of the SUMI)" (Brooke, 1996, p. 194). According to Brooke (1996, p. 194), "the only prerequisite for its use is that any published report should acknowledge the source of the measure." The standard SUS consists of the following 10 items (odd-numbered items worded positively; even-numbered items worded negatively):

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

To use the SUS, present the items to participants as five-point scales numbered from 1 (anchored with "Strongly disagree") to 5 (anchored with "Strongly agree"). If a participant fails to respond to an item, assign it a 3 (the center of the rating scale). After completion, determine each item's score contribution, which will range from 0 to 4. For positively worded items (1, 3, 5, 7, and 9), the score contribution is the scale position minus 1. For negatively worded items (2, 4, 6, 8, and 10), it is 5 minus the scale position. To get the overall SUS score, multiply the sum of the item score contributions by 2.5. Thus, SUS scores range from 0 to 100 in 2.5-point increments.

Since its initial publication, research on the SUS has led to some proposed changes in the original wording of the items. Finstad (2006) and Bangor et al. (2008)

recommend the use of the word "awkward" rather than "cumbersome" in item 8. The original SUS items refer to "system," but substituting the word "product" or the use of the actual product name in place of "system" seems to have no effect on SUS scores (Lewis and Sauro, 2009), but, of course, substitutions should be consistent across the items.

An early assessment of the SUS indicated reliability (assessed using coefficient $\alpha$) of 0.85 (Lucey, 1991). More recent estimates of SUS reliability indicate the reliability of the SUS is somewhat higher (0.91 from Bangor et al., 2008; 0.92 from Lewis and Sauro, 2009). Tullis and Stetson (2004) indirectly provided additional evidence of SUS reliability when they found that of five methods for assessing satisfaction with usability, the SUS was the quickest to converge on the "correct" conclusion regarding the usability of two websites as a function of sample size, where "correct" meant a significant $t$-test consistent with the decision reached using the total sample size.

In addition to being highly reliable, recent studies have shown evidence of the validity of the SUS. Bangor et al. (2008) reported that the SUS was sensitive to differences among types of interfaces and as a function of changes made to a product and showed concurrent validity (a significant correlation of 0.806 between the SUS and a single seven-point rating of user friendliness). Lewis and Sauro (2009) also found the SUS to be sensitive.

Another recent finding is that the SUS, long assumed to be a unidimensional measure, actually appears to have two components (Borsci et al., 2009; Lewis and Sauro, 2009), with items 1, 2, 3, 5, 6, 7, 8, and 9 aligning with a factor named "Usable" (coefficient $\alpha = 0.91$) and items 4 and 10 aligning with "Learnable" (coefficient $\alpha = 0.70$). Practitioners who use the SUS can continue doing so, but in addition to working with the overall SUS score, they can easily decompose it into its Usable and Learnable components, extracting additional information from their data with very little effort.

### 3.3.4  PSSUQ and CSUQ

The PSSUQ is a questionnaire designed for the purpose of assessing users' perceived satisfaction with their computer systems. It has its origin in an internal IBM project called SUMS (System Usability MetricS), headed by Suzanne Henry in the late 1980s. A team of human factors engineers and usability specialists working on SUMS created a pool of seven-point scale items based on the work of Whiteside et al. (1988) and from that pool selected 18 items to use in the first version of the PSSUQ (Lewis, 1992). Each item was worded positively, with the scale anchors "strongly agree" at the first scale position (1) and "strongly disagree" at the last scale position (7). A "not applicable" (NA) choice and a comment area were available for each item [see Lewis (1995) for examples of the appearance of the items].

The development of the Computer System Usability Questionnaire (CSUQ) followed the development of the first version of the PSSUQ. Its items are identical to those of the PSSUQ except that their wording is

appropriate for use in field settings or surveys rather than in a scenario-based usability test, making it, essentially, an alternative form of the PSSUQ. For a discussion of CSUQ research and comparison of the PSSUQ and CSUQ items, see Lewis (1995).

An unrelated series of IBM investigations into customer perception of usability revealed a common set of five usability characteristics associated with usability by several different user groups (Doug Antonelli, personal communication, January 5, 1991). The 18-item version of the PSSUQ addressed four of these five characteristics (quick completion of work, ease of learning, high-quality documentation and online information, and functional adequacy) but did not address the fifth (rapid acquisition of productivity). The second version of the PSSUQ (Lewis, 1995) included an additional item to address this characteristic, bringing the total number of items up to 19.

Lewis (2002) conducted a psychometric evaluation of the PSSUQ using data from several years of usability studies (primarily studies of speech dictation systems, but including studies of other types of applications). The results of a factor analysis on these data were consistent with earlier factor analyses (Lewis, 1992, 1995) used to define three PSSUQ subscales: system usefulness (SysUse), information quality (InfoQual), and interface quality (IntQual). Estimates of reliability were also consistent with those of earlier studies. Analyses of variance indicated that variables such as the specific study, developer, state of development, type of product, and type of evaluation significantly affected PSSUQ scores. Other variables, such as gender and completeness of responses to the questionnaire, did not. Norms derived from the new data correlated strongly with norms derived from earlier studies.

Significant correlation analyses indicated scale validity (Lewis, 1995). For a sample of 22 participants who completed all PSSUQ and ASQ items in a usability study (Lewis et al., 1990), the overall PSSUQ score correlated highly with the sum of the ASQ ratings that participants gave after completing each scenario [$r(20) = 0.80, p = 0.0001$]. The overall PSSUQ score correlated significantly with the percentage of successful scenario completions [$r(29) = -0.40, p = 0.026$]. SysUse [$r(36) = -0.40, p = 0.006$] and IntQual [$r(35) = -0.29, p = 0.08$] also correlated with the percentage of successful scenario completions.

One potential criticism of the PSSUQ has been that some items seemed redundant and that this redundancy might inflate estimates of reliability. Lewis (2002) investigated the effect of removing three items from the second version of the PSSUQ (items 3, 5, and 13). With these items removed, the reliability of the overall PSSUQ score (using coefficient $\alpha$) was 0.94 (remaining very high), and the reliabilities of the three subscales were:

- SysUse: 0.90
- InfoQual: 0.91
- IntQual: 0.83

All of the reliabilities exceeded 0.80, indicating sufficient reliability to be valuable as usability measurements

(Anastasi, 1976; Landauer, 1997). Thus, the third (and current) version of the PSSUQ has 16 seven-point scale items (see Table 12 for the items and their normative scores).

Note that the scale construction is such that lower scores are better than higher scores and that the means of the items and scales all fall below the scale midpoint of 4. With the exception of item 7 ("The system gave error messages that clearly told me how to fix problems"), the upper limits of the confidence intervals are below 4. This shows that practitioners should not use the scale midpoint exclusively as a reference from which they would judge participants' perceptions of usability. Rather, they should also use the norms shown in Table 12 (and comparison with these norms is probably more meaningful than comparison with the scale midpoint).

The way that item 7 stands out from the others indicates:

- It should not surprise practitioners if they find this in their own data.
- It is a difficult task to provide usable error messages throughout a product.
- It may well be worth the effort to focus on providing usable error messages.
- If practitioners find the mean for this item to be equal to or less than the mean of the other items in InfoQual (assuming that they are in line with the norms), they have been successful in creating better-than-average error messages.

The consistent pattern of relatively poor ratings for InfoQual versus IntQual [seen across all the studies; for details and complete normative data, see Lewis (2002)] suggests that practitioners who find this pattern in their data should not conclude that they have poor documentation or a great interface. Suppose, however, that this pattern appeared in the first iteration of a usability evaluation and the developers decided to emphasize improvement to the quality of their information. Any subsequent decline in the difference between InfoQual and IntQual would be evidence of a successful intervention.

Another potential criticism of the PSSUQ is that the items do not follow the typical convention of varying the tone of the items so that half of the items elicit agreement and the other half elicit disagreement (Swamy, 2007). The rationale for the decision to align the items consistently was to make it as easy as possible for participants to complete the questionnaire. With consistent item alignment, the proper way to mark responses on the items is clearer, potentially reducing response errors due to participant confusion. Also, the use of negatively worded items can produce a number of undesirable effects (Barnette, 2000; Ibrahim, 2001; Sauro and Lewis, 2011), including problems with internal consistency and factor structure. The setting in which balancing the tone of the items is likely to be of greatest value is when participants do not have a high degree of motivation for providing reasonable and honest responses (e.g., in clinical and educational settings).

**Table 12 PSSUQ Version 3 Items, Scales, and Normative Scores (99% Confidence Intervals)[a]**

| | | Norm (99% CI) | | |
| --- | --- | --- | --- | --- |
| Item/Scale | Item Text/Scale Scoring Rule | Lower Limit | Mean | Upper Limit |
| Q1 | Overall, I am satisfied with how easy it is to use this system. | 2.60 | 2.85 | 3.09 |
| Q2 | It was simple to use this system. | 2.45 | 2.69 | 2.93 |
| Q3 | I was able to complete the tasks and scenarios quickly using this system. | 2.86 | 3.16 | 3.45 |
| Q4 | I felt comfortable using this system. | 2.40 | 2.66 | 2.91 |
| Q5 | It was easy to learn to use this system. | 2.07 | 2.27 | 2.48 |
| Q6 | I believe I could become productive quickly using this system. | 2.54 | 2.86 | 3.17 |
| Q7 | The system gave error messages that clearly told me how to fix problems. | 3.36 | 3.70 | 4.05 |
| Q8 | Whenever I made a mistake using the system, I could recover easily and quickly. | 2.93 | 3.21 | 3.49 |
| Q9 | The information (such as on-line help, on-screen messages and other documentation) provided with this system was clear. | 2.65 | 2.96 | 3.27 |
| Q10 | It was easy to find the information I needed. | 2.79 | 3.09 | 3.38 |
| Q11 | The information was effective in helping me complete the tasks and scenarios. | 2.46 | 2.74 | 3.01 |
| Q12 | The organization of information on the system screens was clear. | 2.41 | 2.66 | 2.92 |
| Q13 | The interface[b] of this system was pleasant. | 2.06 | 2.28 | 2.49 |
| Q14 | I liked using the interface of this system. | 2.18 | 2.42 | 2.66 |
| Q15 | This system has all the functions and capabilities I expect it to have. | 2.51 | 2.79 | 3.07 |
| Q16 | Overall, I am satisfied with this system. | 2.55 | 2.82 | 3.09 |
| *SysUse* | Average items 1–6. | 2.57 | 2.80 | 3.02 |
| *InfoQual* | Average items 7–12. | 2.79 | 3.02 | 3.24 |
| *IntQual* | Average items 13–15. | 2.28 | 2.49 | 2.71 |
| *Overall* | Average items 1–16. | 2.62 | 2.82 | 3.02 |

*Source*: Lewis (2002).
[a] SysUse, system usefulness; InfoQual, information quality; IntQual, interface quality; CI, confidence interval. Scores can range from 1 (strongly agree) to 7 (strongly disagree), with lower scores better than higher scores.
[b] The ''interface'' includes those items that you use to interact with the system. For example, some components of the interface are the keyboard, the mouse, the microphone, and the screens (including their graphics and language).

Obtaining reasonable and honest responses is rarely a problem in most usability testing settings.

Additional key findings and conclusions from Lewis (2002) were:

- There was no evidence of response styles (especially, no evidence of extreme response style) in the PSSUQ data.
- Because there is a possibility of extreme response and acquiescence response styles in cross-cultural research (Baumgartner and Steenkamp, 2001; Clarke, 2001; Grimm and Church, 1999; van de Vijver and Leung, 2001), practitioners should avoid using questionnaires for cross-cultural comparison unless that use has been validated. Other types of group comparisons with the PSSUQ are valid because any effect of response style should cancel out across experimental conditions.
- Scale scores from incomplete PSSUQs were indistinguishable from those computed from complete PSSUQs. These data do not provide information concerning how many items a participant might ignore and still produce reliable

scale scores. They do suggest that, in practice, participants typically complete enough items to produce reliable scale scores.

The similarity of psychometric properties across the various versions of the PSSUQ, despite the passage of time and differences in the types of systems studied, provides evidence of significant generalizability for the questionnaire, supporting its use by practitioners for measuring participant satisfaction with the usability of tested systems. Due to its generalizability, practitioners can confidently use the PSSUQ when evaluating different types of products and at different times during the development process. The PSSUQ can be especially useful in competitive evaluations (for an example, see Lewis, 1996b) or when tracking changes in usability as a function of design changes made during development. Practitioners and researchers are free to use the PSSUQ and CSUQ (no license fees), but anyone using them should cite the source.

### 3.3.5 ASQ

The ASQ (Lewis, 1991b, 1995) is an extremely short questionnaire (three seven-point scale items using

the same format as the PSSUQ). The items address three important aspects of user satisfaction with system usability: ease of task completion ("Overall, I am satisfied with the ease of completing the tasks in this scenario"), time to complete a task ("Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario"), and adequacy of support information ["Overall, I am satisfied with the support information (on-line help, messages, documentation) when completing tasks"]. The overall ASQ score is the average of responses to these three items.

Because the questionnaire is short, it takes very little time for participants to complete, an important practical consideration for usability studies. Measurements of ASQ reliability (using coefficient $\alpha$) have ranged from 0.90 to 0.96 (Lewis, 1995). A significant correlation between ASQ scores and successful scenario completion [$r(46) = -0.40$, $p < 0.01$] in Lewis et al. (1990; analysis reported in Lewis, 1995) provided evidence of concurrent validity. Like the PSSUQ and CSUQ, the ASQ is available for free use by practitioners and researchers, but anyone using the ASQ should cite the source.

### 3.3.6   One-Question Posttask Usability Questionnaires

Sauro and Dumas (2009) compared three one-question rating types in a study with 26 participants, 5 tasks, and 2 software applications. The types were a Likert scale, a usability magnitude estimation (UME) judgment, and a subjective mental effort question (SMEQ). The SMEQ was a 150-point online slider scale with anchors at various points (e.g., "Not at all hard to do" at 0; "Tremendously hard to do" at about 113). The Likert type was an item that stated "Overall, this task was:" followed by seven radio buttons anchored on the left with "Very Easy" and on the right with "Very Difficult." All three types successfully distinguished between the applications, but the Likert and SMEQ types were more sensitive with small sample sizes and were easy to learn and quick to execute. For paper-based questionnaires, the Likert type would be the most effective. For online questionnaires, the Likert and SMEQ are about equally effective.

## 4   WRAPPING UP

### 4.1   Getting More Information about Usability Testing

This chapter has provided fundamental and some advanced information about usability testing, but there is only so much that you can cover in a single chapter. For additional chapter-length treatments of the basics of usability testing, see Nielsen (1997), Dumas (2003), and Dumas and Salzman (2006). There are also three well-known books devoted to the topic of usability testing: Dumas and Redish (1999), Rubin (1994), and Barnum (2002).

Dumas and Redish (1999) was the first of these book-length treatments of usability testing, making the content and references somewhat dated. The 1999 copyright date is a bit misleading, as the body of the book has not changed since its 1993 edition. The 1999 edition does include a new preface and some updated reading recommendations and provides excellent coverage of the fundamentals of usability testing.

Like Dumas and Redish (1999), the content and references of Rubin (1994) are out of date. It, too, covers the fundamentals of usability testing (which have not really changed for 25 years) very well and contains many useful samples of a variety of testing-related forms and documents.

Barnum (2002) is more recent but, at the time of writing this chapter, is almost 10 years old. It has a companion website (www.ablongman.com/barnum/) that includes sample reports and usability laboratory resources.

The most recent book on usability testing is the second edition of Rubin's *Handbook of Usability Testing*, coauthored with Dana Chisnell (Rubin and Chisnell, 2008). It also has a companion website (www.wiley.com/go/usabilitytesting).

Tullis and Albert's (2008) *Measuring the User Experience: Collecting, Analyzing and Presenting Usability Metrics* is a book-length treatment of usability measurement, with a companion website at www.measuringux.com.

Sauro and Lewis (2012) is a book-length treatment of statistical methods for usability testing and other user research applications.

For late-breaking developments in usability research and practice, there are a number of annual conferences that have usability evaluation as a significant portion of their content. Companies making a sincere effort in the professional development of their usability practitioners should ensure that their personnel have access to the proceedings of these conferences and should support attendance at one or more of these conferences at least every few years. These major conferences are:

- Usability Professionals Association (www.upassoc.org/)
- Human–Computer Interaction International (www.hci-international.org/)
- ACM Special Interest Group in Computer–Human Interaction (www.acm.org/sigchi/)
- Human Factors and Ergonomics Society (hfes.org/)
- INTERACT (held every two years; see, e.g., www.interact2005.org/)

### 4.2   Usability Testing: Yesterday, Today, and Tomorrow

It seems clear that usability testing (both summative and formative) is here to stay and that its general form will remain similar to the forms that emerged in the late 1970s and early 1980s. The last 30 years have seen the introduction of more usability evaluation techniques and some consensus (and some continuing debate) on the conditions under which to use the various techniques, either alone or in combination (Al-Wabil and Al-Khalifa, 2009; Hornbæk, 2010; Jarrett et al., 2009).

In the last 20 years, usability researchers have made significant progress in the areas of standardized usability questionnaires and sample size estimation for formative usability tests. More recently, there has been significant advancement in large-sample remote usability testing (Albert et al., 2010). Also, given its emerging focus on commercial self-service, it is reasonable to anticipate standardized usability questionnaires for the Internet (Bargas-Avila et al., 2009; Lascu and Clow, 2008), with extensions to address Internet-specific factors such as trust and other elements of customer satisfaction from the marketing research literature (Safar and Turner, 2005).

As we look to the future, usability practitioners should monitor the continuing research taking place in the scientific study of usability (Gillan and Bias, 2001). Of particular interest are the various proposed extensions of usability beyond effectiveness, efficiency, and satisfaction to include factors such as hedonics, aesthetics, safety, and flexibility (Bevan, 2009; Bødker and Sundblad, 2008; Hornbæk, 2006; Sonderegger and Sauer, 2010). If we expand the definition of usability in these ways, then do we risk obscuring the fundamental concept of usability?

In the meantime, practitioners will continue to perform usability tests, exercising professional judgment as required. Usability testing is not a perfect usability evaluation method in the sense that it does not guarantee the discovery of all possible usability problems, but it does not have to be perfect to be useful and effective. It is, however, important to understand its strengths, limitations, and current leading practices to ensure its proper (most effective) use.

## REFERENCES

Aaker, D. A., and Day, G. S. (1986), *Marketing Research*, Wiley, New York.

Abelson, R. P. (1995), *Statistics as Principled Argument*, Lawrence Erlbaum Associates, Mahwah, NJ.

Agresti, A., and Coull, B. (1998), "Approximate Is Better Than 'Exact' for Interval Estimation of Binomial Proportions," *The American Statistician*, Vol. 52, pp. 119–126.

Al-Awar, J., Chapanis, A., and Ford, R. (1981), "Tutorials for the First-Time Computer User," *IEEE Transactions on Professional Communication*, Vol. 24, pp. 30–37.

Albert, W., Tullis, T., and Tedesco, D. (2010), *Beyond the Usability Lab: Conducting Large-Scale User Experience Studies*, Morgan-Kaufmann, Burlington, MA.

Alonso-Ríos, D., Vázquez-García, A., Mosqueira-Rey, E., and Moret-Bonillo. (2010), "Usability: A Critical Analysis and a Taxonomy," *International Journal of Human-Computer Interaction*, Vol. 26, No. 1, pp. 53–74.

Alreck, P. L., and Settle, R. B. (1985), *The Survey Research Handbook*, Richard D. Irwin, Homewood, IL.

Al-Wabil, A., and Al-Khalifa, H. (2009), "A Framework for Integrating Usability Methods: The Mawhiba Web Portal Case Study," in *Proceedings of Current Trends in Information Technology 2009*, IEEE, Dubai, UAE, pp. 1–6.

American National Standards Institute (ANSI) (2001), *Common Industry Format for Usability Test Reports*, ANSI-NCITS 354-2001, ANSI, Washington, DC.

Anastasi, A. (1976), *Psychological Testing*, Macmillan, New York.

Andre, T. S., Belz, S. M, McCreary, F. A., and Hartson, H. R. (2000), "Testing a Framework for Reliable Classification of Usability Problems," in *Proceedings of the IEA 2000/HFES 2000 Congress*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 573–576.

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O., and Stage, J. (2007), "What Happened to Remote Usability Testing? An Empirical Study of Three Methods," in *Proceedings of CHI 2007*, Association for Computing Machinery, San Jose, CA, pp. 1405–1414.

Arning, K., and Ziefle, M. (2008), "Development and Validation of a Computer Expertise Questionnaire for Older Adults," *Behaviour & Information Technology*, Vol. 27, No. 1, pp. 89–93.

Aykin, N. M., and Aykin, T. (1991), "Individual Differences in Human–Computer Interaction," *Computers and Industrial Engineering*, Vol. 20, pp. 373–379.

Baecker, R. M. (2008), "Themes in the Early History of HCI—Some Unanswered Questions," *Interactions*, Vol. 15, No. 2, pp. 22–27.

Bailey, G. (1993), "Iterative Methodology and Designer Training in Human–Computer Interface Design," in *INTERCHI '93 Conference Proceedings*, Association for Computing Machinery, New York, pp. 198–205.

Bailey, R. W., Allan, R. W., and Raiello, P. (1992), "Usability Testing vs. Heuristic Evaluation: A Head to Head Comparison," in *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 409–413.

Bailey, R. W., Wolfson, C. A., Nall, J., and Koyani, S. (2009), "Performance-Based Usability Testing: Metrics That Have the Greatest Impact for Improving a System's Usability" in *Human Centered Design, HCII 2009*, M. Kurosu, Ed., Springer-Verlag, Heidelberg, Germany, pp. 3–12.

Bangor, A., Kortum, P. T., and Miller, J. T. (2008), "An Empirical Evaluation of the System Usability Scale," *International Journal of Human-Computer Interaction*, Vol. 24, pp. 574–594.

Banks, S. (1965), *Experimentation in Marketing*, McGraw-Hill, New York.

Bargas-Avila, J. A., Lötscher, J., Orsini, S., and Opwis, K. (2009), "Intranet Satisfaction Questionnaire: Development and Validation of a Questionnaire to Measure User Satisfaction with the Intranet," *Computers in Human Behavior*, Vol. 25, pp. 1241–1250.

Barnette, J. J. (2000), "Effects of Stem and Likert Response Option Reversals on Survey Internal Consistency: If You Feel the Need, There Is a Better Alternative to Using Those Negatively Worded Stems," *Educational and Psychological Measurement*, Vol. 60, pp. 361–370.

Barnum, C. (2002), *Usability Testing and Research*, Longman, New York.

Baumgartner, H., and Steenkamp, J. B. E. M. (2001), "Response Styles in Marketing Research: A Cross-National Investigation," *Journal of Marketing Research*, Vol. 38, pp. 143–156.

Bennett, J. L. (1979), "The Commercial Impact of Usability in Interactive Systems," *Infotech State of the Art Report: Man/Computer Communication*, Vol. 2, pp. 289–297.

Berry, D. C., and Broadbent, D. E. (1990), "The Role of Instruction and Verbalization in Improving Performance on Complex Search Tasks," *Behaviour & Information Technology*, Vol. 9, pp. 175–190.

Bevan, N. (2009), "Extending Quality in Use to Provide a Framework for Usability Measurement," in *Human Centered Design, HCII 2009*, M. Kurosu, Ed., Springer-Verlag, Heidelberg, Germany, pp. 13–22.

Bevan, N., Kirakowski, J., and Maissel, J. (1991), "What Is Usability?" in *Human Aspects in Computing, Design and Use of Interactive Systems and Work with Terminals, Proceedings of the 4th International Conference on Human–Computer Interaction*, H. J. Bullinger, Ed., Elsevier Science, Stuttgart, Germany, pp. 651–655.

Bias, R. G., and Mayhew, D. J. (1994), *Cost-Justifying Usability*, Academic, Boston.

Blalock, H. M. (1972), *Social Statistics*, McGraw-Hill, New York.

Bødker, S., and Sundblad, Y. (2008), "Usability and Interaction Design—New Challenges for the Scandinavian Tradition," *Behaviour & Information Technology*, Vol. 27, No. 4, pp. 293–300.

Boehm, B. W. (1981), *Software Engineering Economics*, Prentice-Hall, Englewood Cliffs, NJ.

Boren, T., and Ramey, J. (2000), "Thinking Aloud: Reconciling Theory and Practice," *IEEE Transactions on Professional Communications*, Vol. 43, pp. 261–278.

Borsci, S., Federici, S., and Lauriola, M. (2009), "On the Dimensionality of the System Usability Scale: A Test of Alternative Measurement Models," *Cognitive Processes*, Vol. 10, pp. 193–197.

Bosenick, T., Kehr, S., Kühn, M., and Nufer, S. (2007), "Remote Usability Tests—An Extension of the Usability Toolbox for Online-Shops," in *Universal Access in HCI, Part I, HCII 2007*, C. Stephanidis, Ed., Springer-Verlag, Heidelberg, Germany, pp. 392–398.

Bowers, V., and Snyder, H. (1990), "Concurrent versus Retrospective Verbal Protocols for Comparing Window Usability," in *Proceedings of the Human Factors Society 34th Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 1270–1274.

Bradley, J. V. (1976), *Probability; Decision; Statistics*, Prentice-Hall, Englewood Cliffs, NJ.

Brooke, J. (1996), "SUS: A 'Quick and Dirty' Usability Scale," in *Usability Evaluation in Industry*, P. Jordan, B. Thomas, and B. Weerdmeester, Eds., Taylor & Francis, London, pp. 189–194.

Brown, F. E. (1980), *Marketing Research: A Structure for Decision Making*, Addison-Wesley, Reading, MA.

Campbell, D. T., and Stanley, J. C. (1963), *Experimental and Quasi-Experimental Designs for Research*, Rand McNally, Chicago.

Capra, M. G. (2007), "Comparing Usability Problem Identification and Description by Practitioners and Students," in *Proceedings of the Human Factors and Ergonomics Society 51st Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 474–478.

Caulton, D. A. (2001), "Relaxing the Homogeneity Assumption in Usability Testing," *Behaviour & Information Technology*, Vol. 20, pp. 1–7.

Cavallin, H., Martin, W. M., and Heylighen, A. (2007), "How Relative Absolute Can Be: SUMI and the Impact of the Nature of the Task in Measuring Perceived Software Usability," *Artificial Intelligence and Society*, Vol. 22, pp. 227–235.

Chapanis, A. (1981), "Evaluating Ease of Use," unpublished manuscript prepared for IBM, Boca Raton, FL, available from J. R. Lewis.

Chapanis, A. (1988), "Some Generalizations about Generalization," *Human Factors*, Vol. 30, pp. 253–267.

Chin, J. P., Diehl, V. A., and Norman, K. L. (1988), "Development of an Instrument Measuring User Satisfaction of the Human–Computer Interface," in *CHI '88 Conference Proceedings: Human Factors in Computing Systems*, E. Soloway, D. Frye, and S. B. Sheppard, Eds., Association for Computing Machinery, Washington, DC, pp. 213–218.

Churchill, G. A., Jr. (1991), *Marketing Research: Methodological Foundations*, Dryden, Ft. Worth, TX.

Clarke, I. (2001), "Extreme Response Style in Cross-Cultural Research," *International Marketing Review*, Vol. 18, pp. 301–324.

Clemmensen, T., Hertzum, M., Hornbæk, K., Shi, Q., and Yammiyavar, P. (2009), "Cultural Cognition in Usability Evaluation," *Interacting with Computers*, Vol. 21, pp. 212–220.

Cliff, N. (1987), *Analyzing Multivariate Data*, Harcourt Brace Jovanovich, San Diego, CA.

Cockton, G., and Lavery, D. (1999), "A Framework for Usability Problem Extraction," in *Human Computer Interaction, INTERACT '99*, M. A. Sasse and C. Johnson, Eds., IOS Press, Amsterdam, pp. 344–352.

Cordes, R. E. (1993), "The Effects of Running Fewer Subjects on Time-on-Task Measures," *International Journal of Human–Computer Interaction*, Vol. 5, pp. 393–403.

Cordes, R. E. (2001), "Task-Selection Bias: A Case for User-Defined Tasks," *International Journal of Human–Computer Interaction*, Vol. 13, pp. 411–419.

Cordes, R. E., and Lentz, J. L. (1986), *Usability-Claims Evaluation Using a Binomial Test*, Technical Report 82.0267, International Business Machines Corporation, Tucson, AZ.

Desurvire, H. W., Kondziela, J. M., and Atwood, M. E. (1992), "What Is Gained and Lost When Using Evaluation Methods Other Than Empirical Testing," in *Proceedings of HCI '92: People and Computers VII*, A. Monk, D. Diaper, and M. D. Harrison, Eds., Cambridge University Press, New York, pp. 89–102.

Diamond, W. J. (1981), *Practical Experiment Designs for Engineers and Scientists*, Lifetime Learning Publications, Belmont, CA.

Dickens, J. (1987), "The Fresh Cream Cakes Market: The Use of Qualitative Research as Part of a Consumer Research Programme," in *Applied Marketing and Social Research*, U. Bradley, Ed., Wiley, New York, pp. 23–68.

Dow, S., MacIntyre, B., Lee, J., Oczbek, C., Bolter, J. D., and Gandy, M. (2005), "Wizard of Oz Support throughout an Iterative Design Process," *Pervasive Computing*, Vol. 4, No. 4, pp. 18–26.

Downey, L. L. (2007), "Group Usability Testing: Evolution in Usability Techniques," *Journal of Usability Studies*, Vol. 2, No. 3, pp. 133–144.

Dumas, J. S. (2003), "User-Based Evaluations," in *The Human–Computer Interaction Handbook*, J. A. Jacko and A. Sears, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1093–1117.

Dumas, J. (2007), "The Great Leap Forward: The Birth of the Usability Profession (1988–1993)," *Journal of Usability Studies*, Vol. 2, No. 2, pp. 54–60.

Dumas, J., and Redish, J. C. (1999), *A Practical Guide to Usability Testing*, Intellect, Portland, OR.

Dumas, J. S., and Salzman, M. C. (2006), "Usability Assessment Methods," in *Reviews of Human Factors and Ergonomics*, Vol. 2, R. C. Williges, Ed., Human Factors and Ergonomics Society, Santa Monica, CA, pp. 109–140.

Dumas, J., Sorce, J., and Virzi, R. (1995), "Expert Reviews: How Many Experts Is Enough?" in *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 228–232.

Ericsson, K. A., and Simon, H. A. (1980), "Verbal Reports as Data," *Psychological Review*, Vol. 87, pp. 215–251.

Faulkner, L. (2003), "Beyond the Five-User Assumption: Benefits of Increased Sample Sizes in Usability Testing," *Behavior Research Methods, Instruments, and Computers*, Vol. 35, pp. 379–383.

Finstad, K. (2006), "The System Usability Scale and Non-Native English Speakers," *Journal of Usability Studies*, Vol. 1, No. 4, pp. 185–188.

Fisher, J. (1991), "Defining the Novice User," *Behaviour & Information Technology*, Vol. 10, pp. 437–441.

Fowler, C. J. H., Macaulay, L. A., and Fowler, J. F. (1985), "The Relationship between Cognitive Style and Dialogue Style: An Exploratory Study," in *People and Computers: Designing the Interface*, P. Johnson and S. Cook, Eds., Cambridge University Press, Cambridge, pp. 186–198.

Fu, L., Salvendy, G., and Turley, L. (2002), "Effectiveness of User Testing and Heuristic Evaluation as a Function of Performance Classification," *Behaviour & Information Technology*, Vol. 21, pp. 137–143.

Gawron, V. J., Drury, C. G., Czaja, S. J., and Wilkins, D. M. (1989), "A Taxonomy of Independent Variables Affecting Human Performance," *International Journal of Man–Machine Studies*, Vol. 31, pp. 643–672.

Genov, A., Keavney, M., and Zazelenchuk, T. (2009), "Usability Testing with Real Data," *Journal of Usability Studies*, Vol. 4, No. 2, pp. 85–92.

Gillan, D., and Bias, R. (2001), "Usability Science I: Foundations," *International Journal of Human-Computer Interaction*, Vol. 13, No. 4, pp. 351–372.

Gordon, W., and Langmaid, R. (1988), *Qualitative Market Research: A Practitioner's and Buyer's Guide*, Gower, Aldershot, Hampshire, England.

Gould, J. D. (1988), "How to Design Usable Systems," in *Handbook of Human–Computer Interaction*, M. Helander, Ed., North-Holland, Amsterdam, pp. 757–789.

Gould, J. D., and Boies, S. J. (1983), "Human Factors Challenges in Creating a Principal Support Office System: The Speech Filing System Approach," *ACM Transactions on Information Systems*, Vol. 1, pp. 273–298.

Gould, J. D., and Lewis, C. (1984), *Designing for Usability: Key Principles and What Designers Think*, Technical Report RC-10317, International Business Machines Corporation, Yorktown Heights, NY.

Gould, J. D., Boies, S. J., Levy, S., Richards, J. T., and Schoonard, J. (1987), "The 1984 Olympic Message System: A Test of Behavioral Principles of System Design," *Communications of the ACM*, Vol. 30, pp. 758–769.

GraphPad. (2004), "Confidence Interval of a Proportion or Count," available: http://graphpad.com/quickcalcs/ConfInter-val1.cfm.

Gray, W. D., and Salzman, M. C. (1998), "Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods," *Human–Computer Interaction*, Vol. 13, pp. 203–261.

Greene, S. L., Gomez, L. M., and Devlin, S. J. (1986), "A Cognitive Analysis of Database Query Production," in *Proceedings of the 30th Annual Meeting of the Human Factors Society*, Human Factors Society, Santa Monica, CA, pp. 9–13.

Grimm, S. D., and Church, A. T. (1999), "A Cross-Cultural Study of Response Biases in Personality Measures," *Journal of Research in Personality*, Vol. 33, pp. 415–441.

Guan, Z., Lee, S., Cuddihy, E., and Ramey, J. (2006), "The Validity of the Stimulated Retrospective Think-Aloud Method as Measured by Eye Tracking," in *Proceedings of CHI 2006*, Association for Computing Machinery, Montreal, Quebec, pp. 1253–1262.

Hackman, G. S., and Biers, D. W. (1992), "Team Usability Testing: Are Two Heads Better Than One?" in *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1205–1209.

Harris, R. J. (1985), *A Primer of Multivariate Statistics*, Academic, Orlando, FL.

Hassenzahl, M. (2000), "Prioritizing Usability Problems: Data Driven and Judgement Driven Severity Estimates," *Behaviour & Information Technology*, Vol. 19, pp. 29–42.

Hertzum, M. (2006), "Problem Prioritization in Usability Evaluation: From Severity Assessments to Impact on Design," *International Journal of Human-Computer Interaction*, Vol. 21, No. 2, pp. 125–146.

Hertzum, M. (2010), "Images of Usability," *International Journal of Human-Computer Interaction*, Vol. 26, No. 6, pp. 567–600.

Hertzum, M., and Jacobsen, N. J. (2003), "The Evaluator Effect: A Chilling Fact about Usability Evaluation Methods," *International Journal of Human–Computer Interaction*, Vol. 15, pp. 183–204.

Hertzum, M., Hansen, K. D., and Andersen, H. H. K. (2009), "Scrutinising Usability Evaluation: Does Thinking Aloud Affect Behaviour and Mental Workload?" *Behaviour & Information Technology*, Vol. 28, No. 2, pp. 165–181.

Høegh, R. T., and Jensen, J. J. (2008), "A Case Study of Three Software Projects: Can Software Developers Anticipate the Usability Problems in Their Software?" *Behaviour & Information Technology*, Vol. 27, No. 4, pp. 307–312.

Høegh, R. T., Nielsen, C. M., Overgaard, M., Pedersen, M. B., and Stage, J. (2006), "The Impact of Usability Reports and User Test Observations on Developers' Understanding of Usability Data: An Exploratory Study," *International Journal of Human-Computer Interaction*, Vol. 21, No. 2, 173–196.

Holleran, P. A. (1991), "A Methodological Note on Pitfalls in Usability Testing," *Behaviour &Information Technology*, Vol. 10, pp. 345–357.

Hornbæk, K. (2006), "Current Practice in Measuring Usability: Challenges to Usability Studies and Research," *International Journal of Human-Computer Studies*, Vol. 64, No. 2, pp. 79–102.

Hornbæk, K. (2010), "Dogmas in the Assessment of Usability Evaluation Methods," *Behaviour & Information Technology*, Vol. 29, No. 1, pp. 97–111.

Hornbæk, K., and Frøkjær, E. (2008a), "A Study of the Evaluator Effect in Usability Testing," *Human-Computer Interaction*, Vol. 23, pp. 251−277.

Hornbæk, K., and Frøkjær, E. (2008b), "Comparison of Techniques for Matching of Usability Problem Descriptions," *Interacting with Computers*, Vol. 20, pp. 505−514.

Hornbæk, K., and Law, E. L. (2007), "Meta-Analysis of Correlations among Usability Measures," in *Proceedings of CHI 2007*, Association for Computing Machinery, San Jose, CA, pp. 617−626.

Howard, T., and Howard, W. (2009), "Unexpected Complexity in User Testing of Information Products," in *Proceedings of the Professional Communication Conference*, Institute of Electrical and Electronics Engineers, Waikiki, HI, pp. 1−5.

Howard, T. W. (2008), "Unexpected Complexity in a Traditional Usability Study," *Journal of Usability Studies*, Vol. 3, No. 4, pp. 189−205.

Hwang, W., and Salvendy, G. (2007), "What Makes Evaluators to Find More Usability Problems?: A Meta-Analysis for Individual Detection Rates," in *Human-Computer Interaction, Part I, HCII 2007*, J. Jacko, Ed., Springer-Verlag, Heidelberg, Germany, pp. 499−507.

Hwang, W., and Salvendy, G. (2009), "Integration of Usability Evaluation Studies via a Novel Meta-Analytic Approach: What Are Significant Attributes for Effective Evaluation," *International Journal of Human-Computer Interaction*, Vol. 25, No. 4, pp. 282−306.

Hwang, W., and Salvendy, G. (2010), "Number of People Required for Usability Evaluation: The 10+2 Rule," *Communications of the ACM*, Vol. 53, No. 5, pp. 130−133.

Ibrahim, A. M. (2001), "Differential Responding to Positive and Negative Items: The Case of a Negative Item in a Questionnaire for Course and Faculty Evaluation," *Psychological Reports*, Vol. 88, pp. 497−500.

Illmensee, T., and Muff, A. (2009), "5 Users Every Friday: A Case Study in Applied Research," in *Proceedings of the 2009 Agile Conference*, Institute of Electrical and Electronics Engineers, Chicago, IL, pp. 404−409.

International Organization for Standardization (ISO) (1998), *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)*, Part 11, *Guidance on Usability*, ISO 9241-11:1998(E), ISO, Geneva, Switzerland:

Jarrett, C., Quesenbery, W., Roddis, I., Allen, S., and Stirling, V. (2009), "Using Measurements from Usability Testing, Search Log Analysis and Web Traffic Analysis to Inform Development of a Complex Web Site Used for Complex Tasks," in *Human Centered Design, HCII 2009*, M. Kurosu, Ed., Springer-Verlag, Heidelberg, Germany, pp. 729−738.

Jeffries, R., and Desurvire, H. (1992), "Usability Testing vs. Heuristic Evaluation: Was There a Contest?" *SIGCHI Bulletin*, Vol. 24, pp. 39−41.

Jokela, T., Koivumaa, J., Pirkola, J., Salminen, P., and Kantola, N. (2006), "Methods for Quantitative Usability Requirements: A Case Study on the Development of the User Interface of a Mobile Phone," *Personal and Ubiquitous Computing*, Vol. 10, pp. 345−355.

Kantner, L., and Rosenbaum, S. (1997), "Usability Studies of WWW Sites: Heuristic Evaluation vs. Laboratory Testing," in *Proceedings of SIGDOC 1997*, Association for Computing Machinery, Salt Lake City, UT, pp. 153−160.

Karahasanovic, A., Hinkel, U. N., Sjøberg, D. I. K., and Thomas, R. (2009), "Comparing of Feedback-Collection and Think-Aloud Methods in Program Comprehension Studies," *Behaviour & Information Technology*, Vol. 28, No. 2, pp. 139−164.

Karat, C. (1997), "Cost-Justifying Usability Engineering in the Software Life Cycle," in *Handbook of Human−Computer Interaction*, 2nd ed., M. Helander, T. K. Landauer, and P. Prabhu, Eds., Elsevier, Amsterdam, The Netherlands, pp. 767−778.

Karat, J. (1997), "User-Centered Software Evaluation Methodologies," in *Handbook of Human−Computer Interaction*, 2nd ed., M. Helander, T. K. Landauer, and P. Prabhu, Eds., Elsevier, Amsterdam, pp. 689−704.

Keenan, S. L., Hartson, H. R., Kafura, D. G., and Schulman, R. S. (1999), "The Usability Problem Taxonomy: A Framework for Classification and Analysis," *Empirical Software Engineering*, Vol. 1, pp. 71−104.

Kelley, J. F. (1984), "An Iterative Design Methodology for User-Friendly Natural Language Office Information Applications," *ACM Transactions on Information Systems*, Vol. 2, pp. 26−41.

Kelley, J. F. (1985), "CAL: A Natural Language Program Developed with the OZ Paradigm: Implications for Supercomputing Systems," in *Proceedings of the First International Conference on Supercomputing Systems*, Association for Computing Machinery, New York, pp. 238−248.

Kennedy, P. J. (1982), "Development and Testing of the Operator Training Package for a Small Computer System," in *Proceedings of the Human Factors Society 26th Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 715−717.

Kessner, M., Wood, J., Dillon, R. F., and West, R. L. (2001), "On the Reliability of Usability Testing," in *Conference on Human Factors in Computing Systems: CHI 2001 Extended Abstracts*, J. Jacko and A. Sears, Eds., Association for Computing Machinery, Seattle, WA, pp. 97−98.

Kirakowski, J. (1996), "The Software Usability Measurement Inventory: Background and Usage," in *Usability Evaluation in Industry*, P. Jordan, B. Thomas, and B. Weerdmeester, Eds., Taylor & Francis, London, pp. 169−178 (see also http://www.ucc.ie/hfrg/questionnaires/sumi/index.html).

Kirakowski, J., and Corbett, M. (1993), "SUMI: The Software Usability Measurement Inventory," *British Journal of Educational Technology*, Vol. 24, pp. 210−212.

Kirakowski, J., and Dillon, A. (1988), *The Computer User Satisfaction Inventory (CUSI): Manual and Scoring Key*, Human Factors Research Group, University College of Cork, Cork, Ireland.

Kraemer, H. C., and Thiemann, S. (1987), *How Many Subjects? Statistical Power Analysis in Research*, Sage Publications, Newbury Park, CA.

Krahmer, E., and Ummelen, N. (2004), "Thinking about Thinking Aloud: A Comparison of Two Verbal Protocols for Usability Testing," *IEEE Transactions on Professional Communication*, Vol. 47, No. 2, pp. 105−117.

LaLomia, M. J., and Sidowski, J. B. (1990), "Measurements of Computer Satisfaction, Literacy, and Aptitudes: A Review," *International Journal of Human−Computer Interaction*, Vol. 2, pp. 231−253.

Landauer, T. K. (1997), "Behavioral Research Methods in Human–Computer Interaction," in *Handbook of Human–Computer Interaction*, 2nd ed., M. Helander, T. K. Landauer, and P. Prabhu, Eds., Elsevier, Amsterdam, pp. 203–227.

Lascu, D., and Clow, K. E. (2008), "Web Site Interaction Satisfaction: Scale Development Consideration," *Journal of Internet Commerce*, Vol. 7, No. 3, pp. 359–378.

Law, E. L., Hvannberg, E. T., Cockton, G., Palanque, P., Scapin, D., Springett, M., Stary, C., and Vanderdonckt, J. (2005), "Towards the Maturation of IT Usability Evaluation (MAUSE)," In *Proceedings of INTERACT 2005*, M. F. Costabile and F. Paterno, Eds., IFIP, Rome, Italy, pp. 1134–1137.

Lewis, C., and Norman, D. (1986), "Designing for Error," in *User Centered System Design: New Perspectives on Human–Computer Interaction*, D. A. Norman and S. W. Draper, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 411–432.

Lewis, J. R. (1982), "Testing Small System Customer Set-up," in *Proceedings of the Human Factors Society 26th Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 718–720.

Lewis, J. R. (1991a), "A Rank-Based Method for the Usability Comparison of Competing Products," in *Proceedings of the Human Factors Society 35th Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 1312–1316.

Lewis, J. R. (1991b), "Psychometric Evaluation of an After-Scenario Questionnaire for Computer Usability Studies: The ASQ," *SIGCHI Bulletin*, Vol. 23, pp. 78–81.

Lewis, J. R. (1992), "Psychometric Evaluation of the Post-Study System Usability Questionnaire: The PSSUQ," in *Proceedings of the Human Factors Society 36th Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 1259–1263.

Lewis, J. R. (1993), "Multipoint Scales: Mean and Median Differences and Observed Significance Levels," *International Journal of Human–Computer Interaction*, Vol. 5, pp. 382–392.

Lewis, J. R. (1994), "Sample Sizes for Usability Studies: Additional Considerations," *Human Factors*, Vol. 36, pp. 368–378.

Lewis, J. R. (1995), "IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use," *International Journal of Human–Computer Interaction*, Vol. 7, pp. 57–78.

Lewis, J. R. (1996a), "Binomial Confidence Intervals for Small Sample Usability Studies," in *Advances in Applied Ergonomics: Proceedings of the 1st International Conference on Applied Ergonomics, ICAE '96*, G. Salvendy and A. Ozok, Eds., USA Publishing, Istanbul, Turkey, pp. 732–737.

Lewis, J. R. (1996b), "Reaping the Benefits of Modern Usability Evaluation: The Simon Story," In *Advances in Applied Ergonomics: Proceedings of the 1st International Conference on Applied Ergonomics, ICAE '96*, G. Salvendy and A. Ozok, Eds., USA Publishing, Istanbul, Turkey, pp. 752–757.

Lewis, J. R. (2001), "Evaluation of Procedures for Adjusting Problem-Discovery Rates Estimated from Small Samples," *International Journal of Human–Computer Interaction*, Vol. 13, pp. 445–479.

Lewis, J. R. (2002), "Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies," *International Journal of Human–Computer Interaction*, Vol. 14, pp. 463–488.

Lewis, J. R. (2006), "Sample Sizes for Usability Tests: Mostly Math, Not Magic," *Interactions*, Vol. 13, No. 6, pp. 29–33. See corrected formula in *Interactions*, Vol. 14, No. 1, p. 4.

Lewis, J. R. (2008), "Usability Evaluation of a Speech Recognition IVR," in *Measuring the User Experience, Chapter 10: Case Studies*, T. Tullis and B. Albert, Eds., Morgan-Kaufmann, Amsterdam, Netherlands, pp. 244–252.

Lewis, J. R. (2011), *Practical Speech User Interface Design*, Taylor & Francis, Boca Raton, FL.

Lewis, J. R., and Pallo, S. (1991), *Evaluation of Graphic Symbols for Phone and Line*, Technical Report 54.572, International Business Machines Corporation, Boca Raton, FL.

Lewis, J. R., and Sauro, J. (2006), "When 100% Really Isn't 100%: Improving the Accuracy of Small-Sample Estimates of Completion Rates," *Journal of Usability Studies*, Vol. 3, No. 1, pp. 136–150.

Lewis, J. R., and Sauro, J. (2009), "The Factor Structure of the System Usability Scale," in *Human Centered Design, HCII 2009*, M. Kurosu, Ed., Springer-Verlag, Heidelberg, Germany, pp. 94–103.

Lewis, J. R., Henry, S. C., and Mack, R. L. (1990), "Integrated Office Software Benchmarks: A Case Study," in *Proceedings of the 3rd IFIP Conference on Human–Computer Interaction, INTERACT '90*, D. Diaper et al., Eds., Elsevier Science, Cambridge, pp. 337–343.

Lilienfeld, S. O., Wood, J. M., and Garb, H. N. (2000), "The Scientific Status of Projective Techniques," *Psychological Science in the Public Interest*, Vol. 1, pp. 27–66.

Lindgaard, G. (2006), "Notions of Thoroughness, Efficiency, and Validity: Are They Valid in HCI Practice?" *International Journal of Industrial Ergonomics*, Vol. 36, pp. 1069–1074.

Lindgaard, G., and Chattratichart, J. (2007), "Usability Testing: What Have We Overlooked?" in *Proceedings of CHI 2007*, Association for Computing Machinery, San Jose, CA, pp. 1415–1424.

Lord, F. M. (1953), "On the Statistical Treatment of Football Numbers," *American Psychologist*, Vol. 8, pp. 750–751.

Lucey, N. M. (1991), "More Than Meets the I: User-Satisfaction of Computer Systems," Unpublished thesis for Diploma in Applied Psychology, University College Cork, Cork, Ireland.

Macefield, R. (2007), "Usability Studies and the Hawthorne Effect," *Journal of Usability Studies*, Vol. 2, No. 3, pp. 145–154.

MacKenzie, I. S., and Read, J. C. (2007), "Using Paper Mockups for Evaluating Soft Keyboard Layouts," in *Proceedings of the 2007 Conference of the Center for Advanced Studies on Collaborative Research*, Association for Computing Machinery, Richmond Hill, Canada, pp. 98–108.

Macleod, M., Bowden, R., Bevan, N., and Curson, I. (1997), "The MUSiC Performance Measurement Method," *Behaviour & Information Technology*, Vol. 16, pp. 279–293.

Mao, J., Vredenburg, K., Smith, P. W., and Carey, T. (2005), "The State of User-Centered Design Practice," *Communications of the ACM*, Vol. 48, No. 3, pp. 105–109.

Marshall, C., Brendan, M., and Prail, A. (1990), "Usability of Product X: Lessons from a Real Product," *Behaviour & Information Technology*, Vol. 9, pp. 243–253.

Mayer, R. E. (1997), "From Novice to Expert," In *Handbook of Human–Computer Interaction*, 2nd ed., M. G. Helander, T. K. Landauer, and P. V. Prabhu, Eds., Elsevier, Amsterdam, pp. 781–795.

McCurdy, M., Connors, C., Pyrzak, G., Kanefsky, B., and Vera, A. (2006), "Breaking the Fidelity Barrier: An Examination of Our Current Characterization of Prototypes and an Example of a Mixed-Fidelity Success," in *Proceedings of CHI 2006*, Association for Computing Machinery, Montreal, Canada, pp. 1233–1242.

McFadden, E., Hager, D. R., Elie, C. J., and Blackwell, J. M. (2002), "Remote Usability Evaluation: Overview and Case Studies," *International Journal of Human–Computer Interaction*, Vol. 14, pp. 489–502.

McSweeney, R. (1992), "SUMI: A Psychometric Approach to Software Evaluation," unpublished M.A. (Qual.) thesis in applied psychology, University College of Cork, Cork, Ireland.

Miller, L. A., Stanney, K. M., and Wooten, W. (1997), "Development and Evaluation of the Windows Computer Experience Questionnaire (WCEQ)," *International Journal of Human–Computer Interaction*, Vol. 9, pp. 201–212.

Moffat, B. (1990), "Normalized Performance Ratio: A Measure of the Degree to Which a Man–Machine Interface Accomplishes Its Operational Objective," *International Journal of Man–Machine Studies*, Vol. 32, pp. 21–108.

Molich, R., and Dumas, J. S. (2008), "Comparative Usability Evaluation (CUE-4)," *Behaviour & Information Technology*, Vol. 27, No. 3, pp. 263–281.

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D., and Kirakowski, J. (1998), "Comparative Evaluation of Usability Tests," in *Usability Professionals Association Annual Conference Proceedings*, Usability Professionals Association, Washington, DC, pp. 189–200.

Molich, R., Ede, M. R., Kaasgaard, K., and Karyukin, B. (2004), "Comparative Usability Evaluation," *Behaviour & Information Technology*, Vol. 23, pp. 65–74.

Molich, R., Jeffries, R., and Dumas, J. S. (2007), "Making Usability Recommendations Useful and Usable," *Journal of Usability Studies*, Vol. 2, No. 4, pp. 162–179.

Morse, E. L. (2000), "The IUSR Project and the Common Industry Reporting Format," in *Proceedings of the ACM Conference on Universal Usability*, Association for Computing Machinery, Arlington, VA, pp. 155–156.

Myers, J. L. (1979), *Fundamentals of Experimental Design*, Allyn & Bacon, Boston.

Nielsen, J. (1992), "Finding Usability Problems through Heuristic Evaluation," in *CHI '92 ACM Conference Proceedings*, Monterey, CA, pp. 373–380.

Nielsen, J. (1993), *Usability Engineering*, Academic, San Diego, CA.

Nielsen, J. (1994), "Usability Laboratories," *Behaviour & Information Technology*, Vol. 13, pp. 3–8.

Nielsen, J. (1997), "Usability Testing," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York.

Nielsen, J., and Landauer, T. K. (1993), "A Mathematical Model of the Finding of Usability Problems," in *Proceedings of the ACM INTERCHI'93 Conference*, Association for Computing Machinery, Amsterdam. pp. 206–213.

Nielsen, J., and Molich, R. (1990), "Heuristic Evaluation of User Interfaces," in *Conference Proceedings on Human Factors in Computing Systems, CHI '90*, Association for Computing Machinery, New York, pp. 249–256.

Nisbett, R. E., and Wilson, T. D. (1977), "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review*, Vol. 84, pp. 231–259.

Nørgaard, M., and Hornbæk, K. (2009), "Exploring the Value of Usability Feedback Formats," *International Journal of Human-Computer Interaction*, Vol. 25, No. 1, pp. 49–74.

Norman, D. (1986), "Cognitive Engineering," in *User Centered System Design: New Perspectives on Human–Computer Interaction*, D. A. Norman and S. W. Draper, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 31–61.

Norman, D. A. (1983), "Design Rules Based on Analyses of Human Error," *Communications of the ACM*, Vol. 26, pp. 254–258.

Nunnally, J. C. (1978), *Psychometric Theory*, McGraw-Hill, New York.

Ohnemus, K. R., and Biers, D. W. (1993), "Retrospective versus Thinking Aloud in Usability Testing," in *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting*, Human Factors and Ergonomics Society, Seattle, WA, pp. 1127–1131.

Olmsted-Hawala, E. L., Murphy, E., Hawala, S., and Ashenfelter, K. T. (2010), "Think-Aloud Protocols: A Comparison of Three Think-Aloud Protocols for Use in Testing Data-Dissemination Web Sites for Usability," in *Proceedings of CHI 2010*, Association for Computing Machinery, Atlanta, GA, pp. 2381–2390.

Palmquist, R. A., and Kim, K. S. (2000), "Cognitive Style and On-Line Database Search Experience as Predictors of Web Search Performance," *Journal of the American Society for Information Science*, Vol. 51, pp. 558–566.

Parasuraman, A. (1986), "Nonprobability Sampling Methods," in *Marketing Research*, Addison-Wesley, Reading, MA, pp. 498–516.

Prümper, J., Zapf, D., Brodbeck, F. C., and Frese, M. (1992), "Some Surprising Differences between Novice and Expert Errors in Computerized Office Work," *Behaviour & Information Technology*, Vol. 11, pp. 319–328.

Ramli, R., and Jaafar, A. (2009), "Remote Usability Evaluation System (e-RUE)," in *Second International Conference on Computer and Electrical Engineering*, Institute of Electrical and Electronics Engineers, Dubai, UAE, pp. 639–643.

Rasmussen, J. (1986), *Information Processing and Human–Machine Interaction: An Approach to Cognitive Engineering*, Elsevier, New York.

Redish, J. (2007), "Expanding Usability Testing to Evaluate Complex Systems," *Journal of Usability Studies*, Vol. 2, No. 3, pp. 102–111.

Rengger, R. (1991), "Indicators of Usability Based on Performance," in *Human Aspects in Computing, Design and Use of Interactive Systems and Work with Terminals: Proceedings of the 4th International Conference on Human–Computer Interaction*, H. J. Bullinger, Ed., Elsevier Science, Stuttgart, Germany, pp. 656–660.

Rosenbaum, S. (1989), "Usability Evaluations versus Usability Testing: When and Why?" *IEEE Transactions on Professional Communication*, Vol. 32, pp. 210–216.

Rubin, J. (1994), *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*, Wiley, New York.

Rubin, J., and Chisnell, D. (2008), *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*, 2nd ed., Wiley, New York.

Ruthford, M. A., and Ramey, J. A. (2000), "Design Response to Usability Test Findings: A Case Study Based on Artifacts and Interviews," *IEEE International Professional Communication Conference*, IEEE Press, Piscataway, NJ, pp. 315–323.

Sadowski, W. J. (2001), "Capabilities and Limitations of Wizard of Oz Evaluations of Speech User Interfaces," in *Proceedings of HCI International 2001: Usability Evaluation and Interface Design*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 139–143.

Safar, J. A., and Turner, C. W. (2005), "Validation of a Two Factor Structure for System Trust," in *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 497–501.

Sauro, J. (2004), "Restoring Confidence in Usability Results," available: http://www.measuringusability.com/conf_intervals.htm.

Sauro, J. (2006), "Why a Completion Rate Is Better with a Confidence Interval," *The UPA Voice*, Vol. 18, No. 3, available: http://www.upassoc.org/upa_publications/upa_voice/volumes/2006/june/completion_rate_with_confidence_interval.html.

Sauro, J., and Dumas, J. S. (2009), "Comparison of Three One-Question, Post-Task Usability Questionnaires," in *Proceedings of CHI 2009*, Association for Computing Machinery, Boston, MA, pp. 1599–1608.

Sauro, J., and Kindlund, E. (2005), "A Method to Standardize Usability Metrics into a Single Score," in *Proceedings of CHI 2005*, Association for Computing Machinery, Portland, OR, pp. 401–409.

Sauro, J., and Lewis, J. R. (2005), "Estimating Completion Rates from Small Samples Using Binomial Confidence Intervals: Comparisons and Recommendations," In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 2100–2104.

Sauro, J., and Lewis, J. R. (2009), "Correlations among Prototypical Usability Metrics: Evidence for the Construct of Usability," in *Proceedings of CHI 2009*, Association for Computing Machinery, Boston, MA, pp. 1609–1618.

Sauro, J., and Lewis, J. R. (2010), "Average Task Times in Usability Tests: What to Report?" in *Proceedings of CHI 2010*, Association for Computing Machinery, Atlanta, GA, pp. 2347–2350.

Sauro, J., and Lewis, J. R. (2011), "When Designing Usability Questionnaires, Does It Hurt to Be Positive?" in *Proceedings of CHI 2011*, Association for Computing Machinery, Vancouver, Canada, pp. 2215–2223.

Sauro, J., and Lewis, J. R. (2012), *Quantifying the User Experience: Practical Statistics for User Research*, Morgan-Kaufmann, Burlington, MA.

Schmettow, M. (2008), "Heterogeneity in the Usability Evaluation Process," in *Proceedings of the 22nd British HCI Group Annual Conference on HCI 2008: People and Computers XXII: Culture, Creativity, Interaction*, Vol. 1, Association for Computing Machinery, Liverpool, England, pp. 89–98.

Schmettow, M. (2009), "Controlling the Usability Evaluation Process under Varying Defect Visibility," in *Proceedings of the 2009 British Computer Society Conference on Human-Computer Interaction*, Association for Computing Machinery, Cambridge, pp. 188–197.

Seffah, A., and Habieb-Mammar, H. (2009), "Usability Engineering Laboratories: Limitations and Challenges toward a Unifying Tools/Practices Environment," *Behaviour & Information Technology*, Vol. 28, No. 3, pp. 281–291.

Seffah, A., Donyaee, M., Kline, R. B., and Padda, H. K. (2006), "Usability Measurement and Metrics: A Consolidated Model," *Software Quality Journal*, Vol. 14, pp. 159–178.

Shackel, B. (1990), "Human Factors and Usability," in *Human–Computer Interaction: Selected Readings*, J. Preece and L. Keller, Eds., Prentice Hall International, Hemel Hempstead, Hertfordshire, England, pp. 27–41.

Shneiderman, B. (1987), *Designing the User Interface: Strategies for Effective Human–Computer Interaction*, Addison-Wesley, Reading, MA.

Smith, B., Caputi, P., Crittenden, N., Jayasuriya, R., and Rawstorne, P. (1999), "A Review of the Construct of Computer Experience," *Computers in Human Behavior*, Vol. 15, pp. 227–242.

Smith, D. C., Irby, C., Kimball, R., Verplank, B., and Harlem, E. (1982), "Designing the Star User Interface," *Byte*, Vol. 7, No. 4, pp. 242–282.

Sonderegger, A., and Sauer, J. (2010), "The Influence of Design Aesthetics in Usability Testing: Effects on User Performance and Perceived Usability," *Applied Ergonomics*, Vol. 41, pp. 403–410.

Southwest Oncology Group Statistical Center (SOGSC) (2004), "Binomial Confidence Interval," available: http://www.swogstat.org/stat/public/binomial_conf.htm.

Steele, R. G. D., and Torrie, J. H. (1960), *Principles and Procedures of Statistics*, McGraw-Hill, New York.

Stevens, S. S. (1951), "Mathematics, Measurement, and Psychophysics," in *Handbook of Experimental Psychology*, S. S. Stevens, Ed., Wiley, New York, pp. 1–49.

Sudman, S. (1976), *Applied Sampling*, Academic, New York.

Swamy, S. (2007), "How Should You Frame Questions to Measure User Attitudes Accurately?" in *Usability and Internationalization, Part II, HCII 2007*, N. Aykin, Ed., Springer-Verlag, Heidelberg, Germany, pp. 496–505.

Sy, D. (2007), "Adapting Usability Investigations for Agile User-Centered Design," *Journal of Usability Studies*, Vol. 2, No. 3, pp. 112–132.

Theofanos, M., and Quesenbery, W. (2005), "Towards the Design of Effective Formative Test Reports," *Journal of Usability Studies*, Vol. 1, No. 1, pp. 27–45.

Thimbleby, H. (2007), "User-Centered Methods Are Insufficient for Safety Critical Systems," in *Proceedings of USAB 2007*, A. Holzinger, Ed., Springer-Verlag, Heidelberg, Germany, pp. 1–20.

Tohidi, M., Buxton, W., Baecker, R., and Sellen, A. (2006), "Getting the Right Design and the Design Right: Testing Many Is Better Than One," in *Proceedings of CHI 2006*, Association for Computing Machinery, Montreal, Canada, pp. 1243–1252.

Tullis, T., and Albert, W. (2008), *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*, Morgan-Kaufmann, Burlington, MA.

Tullis, T. S., and Stetson, J. N. (2004), "A Comparison of Questionnaires for Assessing Website Usability," paper presented at the UPA Annual Conference, available: home.comcast.net/~tomtullis/publications/UPA2004TullisStetson.pdf.

Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., and Bergel, M. (2002), "An Empirical Comparison of Lab and Remote Usability Testing of Web Sites," in *Proceedings of the Usability Professionals Association*, Usability Professionals Association, Chicago, IL, pp. 1–8.

Turner, C. W., Lewis, J. R., and Nielsen, J. (2006), "Determining Usability Test Sample Size," in *The International Encyclopedia of Ergonomics and Human Factors*, W. Karowski, Ed., CRC Press, Boca Raton, FL, pp. 3084–3088.

Uldall-Espersen, T., Frøkjær, E., and Hornbæk, K. (2008), "Tracing Impact in a Usability Improvement Process," *Interacting with Computers*, Vol. 20, pp. 48–63.

van de Vijver, F. J. R., and Leung, K. (2001), "Personality in Cultural Context: Methodological Issues," *Journal of Personality*, Vol. 69, pp. 1007–1031.

van den Haak, M. J., and de Jong, D. T. M. (2003), "Exploring Two Methods of Usability Testing: Concurrent versus Retrospective Think-Aloud Protocols," in *Proceedings of the International Professional Communication Conference, IPCC 2003*, Institute of Electrical and Electronics Engineers, Orlando, FL, pp. 285–287.

van den Haak, M. J., and de Jong, M. D. T. (2005), "Analyzing the Interaction between Facilitator and Participants in Two Variants of the Think-Aloud Method," in *2005 IEEE International Professional Communication Conference Proceedings*, Institute of Electrical and Electronics Engineers, Limerick, Ireland, pp. 323–327.

van den Haak, M. J., de Jong, M. D. T., and Schellens, P. J. (2006), "Constructive Interaction: An Analysis of Verbal Interaction in a Usability Setting," *IEEE Transactions on Professional Communication*, Vol. 49, No. 4, pp. 311–324.

Vilbergsdóttir, S. G., Hvannberg, E. T., and Law, E. L. (2006), "Classification of Usability Problems (CUP) Scheme: Augmentation and Exploitation," in *Proceedings of NordiCHI 2006*, Association for Computing Machinery, Oslo, Norway, pp. 281–290.

Virzi, R. A. (1990), "Streamlining the Design Process: Running Fewer Subjects," in *Proceedings of the Human Factors Society 34th Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 291–294.

Virzi, R. A. (1992), "Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough?" *Human Factors*, Vol. 34, pp. 457–468.

Virzi, R. A. (1997). "Usability Inspection Methods," in *Handbook of Human–Computer Interaction*, 2nd ed., M. G. Helander, T. K. Landauer, and P. V. Prabhu, Eds., Elsevier, Amsterdam, pp. 705–715.

Virzi, R. A., Sorce, J. F., and Herbert, L. B. (1993), "A Comparison of Three Usability Evaluation Methods: Heuristic, Think-Aloud, and Performance Testing," in *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 309–313.

Vredenburg, K., Mao, J. Y., Smith, P. W., and Carey, T. (2002), "A Survey of User Centered Design Practice," in *Proceedings of CHI 2002*, Association for Computing Machinery, Minneapolis, MN, pp. 471–478.

Walpole, R. E. (1976), *Elementary Statistical Concepts*, Macmillan, New York.

Wenger, M. J., and Spyridakis, J. H. (1989), "The Relevance of Reliability and Validity to Usability Testing," *IEEE Transactions on Professional Communication*, Vol. 32, pp. 265–271.

West, R., and Lehman, K. R. (2006), "Automated Summative Usability Studies: An Empirical Evaluation," in *Proceedings of CHI 2006*, Association for Computing Machinery, Montreal, Canada, pp. 631–639.

Whiteside, J., Bennett, J., and Holtzblatt, K. (1988), "Usability Engineering: Our Experience and Evolution," in *Handbook of Human–Computer Interaction*, M. Helander, Ed., North-Holland, Amsterdam, pp. 791–817.

Wickens, C. D. (1998), "Commonsense Statistics," *Ergonomics in Design*, Vol. 6, No. 4, pp. 18–22.

Wiethoff, M., Arnold, A., and Houwing, E. (1992), *Measures of Cognitive Workload*, MUSiC ESPRIT Project 5429 document code TUD/M3/TD/2.

Wildman, D. (1995), "Getting the Most from Paired-User Testing," *Interactions*, Vol. 2, No. 3, pp. 21–27.

Williams, G. (1983), "The Lisa Computer System," *Byte*, Vol. 8, No. 2, pp. 33–50.

Winter, S., Wagner, S., and Deissenboeck, F. (2008), "A Comprehensive Model of Usability," in *Engineering Interactive Systems*, International Federation for Information Processing, Heidelberg, Germany, pp. 106–122.

Wixon, D. (2003), "Evaluating Usability Methods: Why the Current Literature Fails the Practitioner," *Interactions*, Vol. 10, No. 4, pp. 28–34.

Woolrych, A., and Cockton, G. (2001), "Why and When Five Test Users Aren't Enough," in *Proceedings of IHM–HCI 2001 Conference*, Vol. 2, J. Vanderdonckt, A. Blandford, and A. Derycke, Eds., Cépadèus Éditions, Toulouse, France, pp. 105–108.

Wright, P. C., and Monk, A. F. (1991), "A Cost-Effective Evaluation Method for Use by Designers," *International Journal of Man–Machine Studies*, Vol. 35, pp. 891–912.

Wright, R. B., and Converse, S. A. (1992), "Method Bias and Concurrent Verbal Protocol in Software Usability Testing," in *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1220–1224.

# CHAPTER 47

# USER REQUIREMENTS COLLECTION AND ANALYSIS

**Dieter Spath, Fabian Hermann, Matthias Peissner and Sandra Sproll**
**Fraunhofer Institute for Industrial Engineering IAO**
**Stuttgart, Germany**

## 1  INTRODUCTION

Software and interactive system development is for markets and users. To get good products which people want to buy and use, much information is needed: information about user goals and tasks, needs, preferences, and wishes as well as specific information about functions or user interface solutions that are desirable, attractive, and efficient, all part of expressed *requirements*. The use of requirements in design and development projects ensures sufficient information flow between users and markets on the one hand and system developers and designers on the other. Developers use these guidelines to plan, focus, and evaluate their work. System designers use requirements as a valuable source of inspiration to come up with new solutions. Additionally, requirements play a key role in the level of project management. They are used to explain the project goals, plan the project, invite tenders, specify contracts, and so on. Thus, requirement specifications help to coordinate different project stakeholders and partners such as clients and contractors, different units, and end users. One common approach to getting a project started is to deliver an initial system outline by compiling a set of requirements, which is then elaborated on to a more precise technical requirements document (Davis, 1993).

Many of the project stakeholders who document, use, or work with requirements are often also *sources* of requirements. In particular, technical or marketing departments define requirements of all kinds, often again relying on diverse input channels. Informal channels include user and customer contacts in sales, support, and help desks as well as user organizations and Web-based product feedback. Results from market research

or product management are often more formalized and thus can deliver general input for the definition of product features or detailed requirements and ideas for functionalities and user interface solutions.

Different levels of detail and scope of requirements and their different usages in the engineering processes have been discussed extensively, yielding a set of typical distinctions. An ongoing International Organization for Standardization (ISO) activity on system and software engineering uses the terms *user needs* and *user requirements*. In working draft ISO/IEC WD 25064 user needs are defined as "factors or conditions necessary for a user to achieve desired results" (ISO, 2010a, p. IV). User requirements, however, are "statements that provide the basis for design and evaluation of interactive systems to meet some or all of the identified user needs" (cf. ISO/IEC WD 25065; ISO, 2010b). This concept reflects the fact that user requirements are often formulated in terms of system features whereas user needs take a more general, user-centered perspective. In ISO/IEC WD 25065, diverse subtypes of requirements are distinguished: typical ergonomic aspects such as task-related and usability requirements, but also user characteristics, information about the context, or compatibility with standards is subsumed here. In software engineering (e.g., Sommerville, 2004), user requirements are seen as high-level descriptions typically formulated in natural language. They are separated from s*ystem requirements*, which are seen as more specific information resulting from an analysis phase: They are detailed specifications of system features, functions, or design solutions. Finally, the demarcation is not strict. In general, the term "user requirements" implies that the user's needs

and interests are in focus, even if technical or design solutions are specified.

A well-established distinction (e.g., ISO/IEC 24765; ISO,2009; Sommerville, 2004) is made between *functional* requirements describing system functionalities and services and *non-functional* requirements that specify quality aspects of the product (e.g., reliability, performance, usability). Nonfunctional requirements that relate to the development process, for example, delivery or procedure, serve the purpose of organizing the cooperation of an ordering and a contracting party.

## 2 METHODS AND APPROACHES FOR USER REQUIREMENTS COLLECTION

For the collection and analysis of user requirements, a number of methods have been established. These methods aim at systematically identifying user needs and specifying contextual conditions which need to be considered when designing user-friendly and efficient interactive systems.

The methods described in the following sections cover only a selection of the many requirements methods which are used in industry today. Other commonly used methods include stakeholder interviews (see, e.g., Fowler and Mangione, 1990) and extant systems analysis (Kirwan and Ainsworth, 1992), which applies well-established usability evaluation methods to existing, often competitive systems in order to identify user habits and expectations, typical problems of use, and best practice design approaches.

### 2.1 Task Analysis

Analyzing typical users' tasks and activities is often regarded as a first step toward understanding the user requirements of an interactive system. For instance, to understand the usage concerns of a specific workplace, the activities of this workplace are explored: What are the personal and organizational goals of individuals and groups in a workplace? What actions do they carry out to achieve these goals?

A popular approach for analyzing activities is hierarchical task analysis (HTA). In a HTA (e.g., Annett et al., 1971), the most relevant user tasks are identified and organized into a hierarchy. Tasks are broken down into subtasks, then operations and actions. HTA can be applied to physical tasks and break them down to a fine level of detail with atomic actions such as pressing a key. But HTA can be applied also to complex cognitive tasks such as planning and decision making. As a result, the task components are graphically represented in a structure chart. HTA entails identifying tasks, categorizing them, identifying the subtasks, and checking the overall accuracy of the model (Crystal and Ellington, 2004). Traditional task analysis assumes that there is one correct and complete description of a user task, and it aims at comprising the description consistently (Carroll, 2002).

The step-by-step transformation of a complex activity into an organized set of successive choices and actions is seen as the central strength of HTA (Rosson and Carroll, 2002). Thus, the resulting hierarchy can be analyzed in terms of completeness, complexity, and consistency.

Besides HTA, there are many other techniques for task analysis [see Kirwan and Ainsworth (1992) for an overview]. Although the techniques differ considerably in terms of focus and specific procedure, their contribution to user requirements analysis is quite the same. A detailed description of the interaction between a human and an interactive system serves as a framework for further analyses. For each step in the task completion process, human capabilities and limitations can be identified, and information needs or other enabling conditions can be specified. Thus, task analysis does not directly end in a set of user requirements, but it provides a detailed understanding and knowledge of user tasks and interactions from which requirements can be discovered.

### 2.2 Ethnography

Task analysis techniques focus on very specific aspects in human work. Some techniques focus more on the observable activities, whereas others highlight the cognitive processes in task completion. In contrast, ethnographic methods take a more holistic perspective. Ethnography developed from work in anthropology which focuses primarily on cultural aspects of humans and aims at understanding humans living within a social group, including rules, practices, and conventions (Spinuzzi, 2000). Ethnography is based on the idea that humans are best understood in the fullest possible context, including their environment and the improvements they make to it. Therefore, ethnographic methods are strong in developing a thick, rich multilayered representation of a user's work or communication habits. Originally, the focus of ethnography was on describing and interpreting cultural aspects, rather than using it as a design method (Blomberg, 1995). Nevertheless, many researchers have used ethnography to gather information for design work (e.g., Blomberg, 1995; Nardi, 1993; Nardi and O'Day, 1999). Many examples of the use of ethnography were concerned with the development of complex computer systems (e.g., Viller and Sommerville, 1999; Wales et al., 2002; D'Souza and Greenstein, 2003). These authors assess ethnography to be an effective approach for obtaining insights into individual work patterns and into the ways technology works within organizations.

In ethnographic studies, the researcher spends an extended period—months or even years—studying the users within their environment. The observation includes the users' behaviors and interactions between users and between users and devices (Martin et al., 2006). The research focuses on learning without preconceived ideas or questions. The main goal is to get immersed into an environment over time, record observations, and obtain insights which emerge from patterns they find in their observations (Spinuzzi, 2000). One prominent example using ethnography is the study of Sommerville et al. (1994) which describes the design process for a new air traffic control computer system. In this study, researchers spent several months

observing the controllers' behaviors and interactions and reading their manuals.

In contrast to the systematic approaches of task analysis, ethnography is exploratory and open ended with emphasis upon discovery. Therefore, it is often especially suitable for identifying unmet or ill-met user needs (cf. Martin et al., 2006). By recording verbal and nonverbal behavior, certain aspects of the environment and usage situation can be identified which are not even obvious to the users themselves (Martin et al., 2006). According to D'Souza and Greenstein (2003, p. 263), "this results in discovering latent needs—those needs of which a user is not aware, that when met, bring a high level of satisfaction."

Ethnographic methods can provide extremely detailed insights into an environment. However, it is also intrusive, time consuming, and therefore extremely expensive. Spinuzzi (2000) indicates it takes 6–12 months for an ethnography study in which researchers observe users working, attending meetings, and so on. Moreover, data analysis is quite costly given the huge amount of gathered data which is mainly qualitative. Therefore, in human–computer interaction (HCI), ethnography is often considered to be impractical as a requirements collection method in design projects, at least in its purest form (Martin et al., 2006).

## 2.3 Contextual Inquiry

Contextual inquiry (CI) is an ethnographic study refined to suit the needs of industrial system design projects in HCI (Holtzblatt and Beyer, 1993). It starts from the assumption that task analysis and self-reports in interviews and discussions often lack important context information. People tend to rationalize their behavior (Ericsson and Simon, 1993) and describe their activities in a prescribed or most typical version. According to Rosson and Carroll (2002), it is important to focus also on tacit or "unofficial" knowledge when collecting and analyzing user requirements. They state that a lot of the users' knowledge is unconscious until users carry out activities or are confronted with their behavior. However, tacit knowledge is considered to be valuable as it often contains the "fixes" and "enhancements" developed informally to address the problems. Thus, CI is a method for probing the end users' conscious and unconscious knowledge in the field.

CI was developed as a field method from the realms of psychology, anthropology, and sociology (Darroch and Silvers, 1982) in connection with the larger design method contextual design (Beyer and Holtzblatt, 1998), where it can be used throughout the development cycle. Thereby, it is often used to gather requirements for a variety of products and systems. Like ethnographic research methods, CI supports an understanding of current practices as a basis for developing a system model that fits users' requirements (Holtzblatt and Jones, 1993). But CI adapts "ethnographic methods to fit the time and resource constraints of engineering" (Holtzblatt and Beyer, 1993, p. 93). Thereby, CI involves collecting detailed information about user activities by observing and interviewing the users while they actually carry out their activities in their normal environments. The goal is to understand how and why something is done or why something is not done (Beyer and Holtzblatt, 1998). The main questions on which to focus include: What is the user's work? What tools are currently used? What works well and why? What are the problems that should be addressed with the new technology? (Holtzblatt and Jones, 1993).

CI is based on four principles guiding the adoption and adaption of the technique to get valuable data:

1. The first and most basic requirement of CI is the principle of data gathering in the *context* of users' work. According to Beyer and Holtzblatt (1998), the key to getting good data is to go where the work happens and observe it while it happens. The context enables us to gather ongoing experiences rather than summary experiences. This is different from the data you get from, for example, questionnaires, because people usually tend to summarize and give overall impressions with one or two highlights. However, we are interested in the detailed structure of work. Furthermore, the context enables us to gather concrete data rather than abstract data, because it is based on in-the-moment experiences (Raven and Flanders, 1996). Concrete data help to identify requirements.

2. *Partnership* describes the relationship between the person performing the CI and the user working together as equals. The goal is to make them collaborators in understanding and exploring work issues. The conversation about work helps the user to become aware of aspects that were formerly invisible (Beyer and Holtzblatt, 1998). By alternating between watching and probing, a true partnership develops in which both partners identify requirements and think about design solutions (Chin et al., 1997; Beyer and Holtzblatt, 1998).

3. For requirement analysis it is necessary to use the language of the respective domain when describing the gathered data. Therefore, an *interpretation* of the observations in terms of their implication about work structure is essential (Beyer and Holtzblatt, 1998).

4. In CI, a clear *focus* steers the conversation. The focus—a perspective or set of concerns—supports the interviewer in keeping the conversation on the central topics. Unlike a structured interview, CI does not constrain the flexibility to follow a promising pathway in a conversation that might not have been in the list of questions (Raven and Flanders, 1996).

According to Beyer and Holtzblatt (1998), the most common structure for CI is the contextual interview. The interview takes place in a one-on-one interaction lasting for 2–3 h. The user performs typical tasks and discusses it with the interviewer. The typical structure includes the following steps:

1. The *conventional interview*, gathering summary data, aims to get to know the users and their issues. After explaining the procedure, the interviewer gets an overview of the work and asks for opinions about relevant tools.

2. The *transition* is the phase where the interviewer explains the rules of a contextual interview, that is, the user is observed while doing his or her work and interrupted by the interviewer whenever something is interesting.

3. The *contextual interview* is the phase where contextual data are gathered. Following the principles of context, partnership, interpretation, and focus, the interviewer is the apprentice, observing and taking notes, asking questions, and suggesting interpretations of behavior.

4. The *wrap-up* occurs at the end of the interview, when the interviewer summarizes his or her understanding of the work structures and tools. The user should finally correct and elaborate on the understanding.

To sum up, CI can be described as a systematic adaption of ethnography for the design of interactive systems. CI is considered to be a discovery method directed at eliciting user requirements (Wixon et al., 2002). It aims at understanding users' needs and requirements by observing and interviewing them in their real environments. The strength of the method is its high external validity. Formal and informal knowledge of the user can serve as the basis for the identification of user requirements.

## 2.4 Focus Groups

Conducting focus groups is a highly accepted method for involving users in the collection and analysis of user requirements (Garmer et al., 2004). The technique is used to identify user needs at early concept stages, to explore product attributes and features required by users and their relative importance, or to identify contextual problems (Martin et al., 2006). Kuniavsky (2003, p. 201) describes focus groups as "an excellent technique for uncovering what people think about a given topic. . . . They reveal what people perceive to be their needs, which is crucial when determining what should be part of an experience and how it should be presented."

Krueger and Casey (2000) describe focus groups in terms of five typical characteristics: (1) People with (2) certain characteristics (3) provide qualitative data (4) in a focused discussion (5) to develop an understanding of a topic of interest, for example, typical usage structures and evolving user needs and requirements:

1. Focus groups typically involve 5–10 people. A focus group should not be run with fewer participants in order to support lively discussions and have a variety of perspectives represented (Nielson, 1993). If there are more than 12 participants, the focus group could fragment into subgroups, for example, by sharing information only with neighbors (Krueger and Casey, 2000).

2. Participants in focus groups have certain characteristics that are important to the research questions. The researcher has to think of "Who can provide the type of information needed?" Participants should be similar regarding the aspects of interest, for example, homogeneity of partic-

ipants is determined by the research questions (Krueger and Casey, 2000). In general, homogeneity supports communication between group members, while heterogeneity provides wider perspective and innovation (Levine and Moreland, 1998; Stewart et al., 2006).

3. Focus groups provide qualitative data. The researcher aims to compare and contrast opinions across several groups. In an inductive process the researcher intends to develop an understanding of user needs and requirements based on discussions rather than coming to one conclusion (Krueger and Casey, 2000).

4. The discussion in focus groups has a clear focus on the topic of interest. Therefore, the set of questions is predetermined, phrased, and arranged in a natural, logical sequence. Usually, a session begins with more general questions getting people to think and talk. As the session continues, questions become more focused, gathering the most useful information near the end (Krueger and Casey, 2000). The moderator has to keep the discussion on track without inhibiting the free flow of ideas (Nielson, 1993). Questions should be easy, clear, and short, usually open ended, and include only dimensions of interest (Krueger and Casey, 2000).

5. Krueger and Casey (2000) point out that a focus group is aimed not at developing an understanding of a topic in terms of overall consensus but rather at understanding user needs and requirements in terms of the feelings and thoughts of the participants. Different opinions among group members help the researcher to identify how and why particular ideas are embraced or rejected (Stewart et al., 2006).

A focus group is a technique (Nielson, 1993) to explore user needs (Caplan, 1990; Greenbaum, 1998). While the moderator follows a preplanned questioning route, from a users' perspective, a session should feel free floating (Nielson, 1993). Through the interaction between the participants, focus groups generate spontaneous reactions and valuable ideas (Caplan, 1990). The interactive environment of a focus group help the members to ponder, reflect, and listen to experiences of others and compare their own personal reality to that of others (Krueger and Casey, 2000).

Focus groups are restricted to what the participants are aware of and what they can recall and articulate (Martin et al. 2006). Martin et al. (2006) suggest focus groups to be complemented by observations of relevant situations in order to gather information about user needs and requirements that cannot be articulated. To sum up, focus groups can collect user needs and requirements quickly and with relatively low cost. They provide an opportunity to interact directly with the users and gather a lot of data in the users' own words. Focus groups are an appropriate technique for different user groups (Stewart et al., 2006). A major challenge is the analysis and interpretation of the recorded data in order to specify user requirements appropriately.

## 2.5  Scenario Analysis

Scenarios have been used as a powerful design tool throughout the entire design process of an interactive system. Scenarios can facilitate all design activities by providing a lightweight way of creating and reusing usage situations (Carroll, 2000).

In this context a scenario is a description that contains actors, background information about the actors and their environments and goals, and sequences of actions or events (Go and Carroll, 2003). Scenarios are stories which are shared among various stakeholders. They can be expressed in various media and forms, for example, textual narratives, storyboards, video mock-ups, or scripted prototypes. In HCI, scenarios typically illustrate user tasks and interactions in a story format (Go and Carroll, 2006).

It is easier for end users and other stakeholders to relate to real-life examples rather than to abstract descriptions of the functions provided by the system. Moreover, scenarios facilitate the communication between designers and users as they can act as a vehicle of knowledge (Go and Carroll, 2003). For these reasons, it can be quite useful to develop a set of scenarios as a starting point for collecting and eliciting user requirements (cf. Sommerville, 2004).

Scenario-based design (SBD) (Caroll, 2002; Rosson and Carroll, 2002) is a systematic approach to ensuring that interaction design will remain focused on the needs and concerns of users throughout the entire design process. In SBD, new designs are developed on the basis of rich and participatory descriptions of all stakeholders and on the basis of a systematic analysis of current usage environments. SBD starts with scenario-based requirements analysis (SBRA), as described by Rosson and Carroll (2002) and illustrated in Figure 1.

At the beginning of the SBRA, analysts compose a *root concept* which describes the project vision and rationale, the assumptions that guide the development process and the initial analysis of project stakeholders in order to develop a shared understanding of the project's high-level goals (Rosson and Carroll, 2002).

In the subsequent *field studies,* current practise and activities that will be transformed by the future system are analyzed based on the root concept. Special attention is given to the needs and concerns of all stakeholders (Rosson and Carroll, 2002). Techniques of qualitative research, such as task observation and recording (Diaper, 1989) and stakeholder interviews or artefacts analysis, are employed.

Then, the collected data is discussed in several *summary* representations: Stakeholder profiles summarize general characteristics of each group and stakeholder diagrams show the relation among the different groups. Another summary documents the tasks of each stakeholder group. Also, tools and artefacts as well as general project themes are summarized (Rosson and Carroll, 2002). During this process, task analysis methods such as hierarchical task analysis (Shepard, 1989) and methods similar to the affinity diagram in contextual design (Beyer and Holzblatt, 1998) are valuable.

*Problem scenarios* tell stories of current practice by describing activities in the problem domain. They contain summary information on identified stakeholders, their key tasks and tools, and the artefacts they use. In a creative process, scenarios are developed. These scenarios are often entirely fictional but are based on real-world characters or observed episodes.

According to Rosson and Carroll (2002), the themes and relationships implicit in a scenario can be made more explicit by analyzing them in *claims*. A claim is seen as a description of trade-offs, that is, pros and
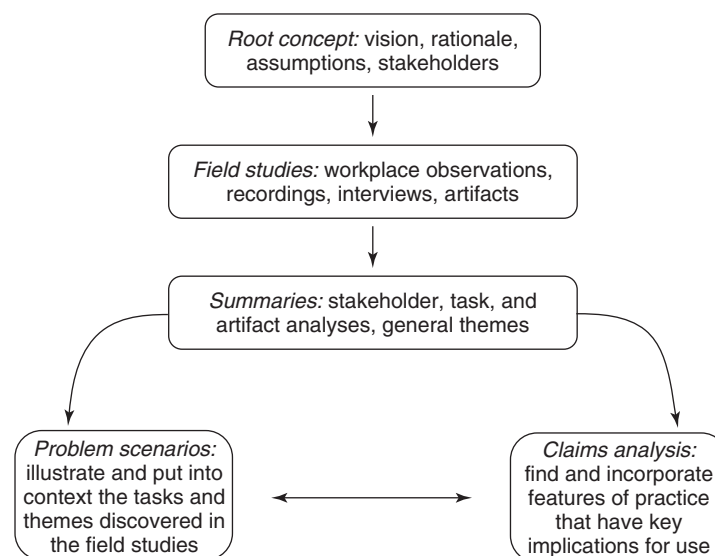


**Figure 1**  Overview of SBRA. (Rosson and Carroll, 2002.)

cons related to specific usability concerns (Go and Carroll, 2006). During requirement analysis, claims are analyzed by identifying features of a scenario that have significant positive or negative consequences for system use. A consequence is some impact on a person's ability to perform a task or enjoy the interaction (Rosson and Carroll, 2002). By explicitly stating the advantages and disadvantages of a specific feature, claims analysis provides a balanced view of both problems and opportunities. During requirements analysis, usually there is a tendency to focus on difficulties of current practice (Bødker, 1991; Nardi, 1996). The methodology of claim analysis ensures that well-working aspects of the current situation are also considered and serve as a basis for the new design (Rosson and Carroll, 2002).

Scenarios can serve as a vehicle for analyzing and specifying requirements and for supporting their transition into the next development phases. In contrast to other methods and formats with the same purposes, scenarios are written from a system's point of view and in a concrete, process-oriented way (Go and Carroll, 2004). Hooper and Hsia (1982) used scenarios in one of the earliest works as prototypes for representing the system for selected sequences of events. The users can simulate the real operations of a system and thus identify their actual needs and requirements. Whereas scenarios seem to provide significant advantages in communication concerns, they are criticized for their shortcomings in completeness and precision (cf. Diaper, 2002). Even an interactive system of average complexity can hardly be fully covered by a manageable number of interaction scenarios. For requirements analysis this means that the actual selection of scenarios will heavily influence and sometimes bias the set of requirements which can be identified. Diaper (2002) also criticizes the common strategy of defining scenarios in a very general (i.e., high level of abstraction) and vague (i.e., lack of detail) manner which allows for different interpretations and therefore misunderstandings in the requirements and design process.

### 2.6 Cultural Probes

"Cultural Probes" (Gaver et al., 1999) have become a prominent approach in interactive system design for exploring design spaces and learning more about user needs by engaging with future users in their situational contexts (e.g., Boehner et al., 2007; Boettcher, 2006; Crabtree et al., 2003; Gaver et al., 2004). Cultural probes originated in artist design and have been used in a number of innovate design projects (e.g., Presence project, and Equator IRC) to primarily inspire design activity (Blythe et al., 2003).

Gaver et al. (1999) introduced cultural probes to address a common dilemma in projects developing interactive products and services for unfamiliar user groups. This design-led approach (Gaver et al., 2004) meets the challenges of both understanding local cultures so that the design fits relevant needs and context aspects and ensuring that the design is not constrained by focusing on needs that are already understood. The aim is to lead discussions toward unexpected user needs and ideas. Unlike traditional approaches, Gaver et al. (1999)

intended to develop an approach for designing for pleasure, rather than for utility, that is, attention is directed to "playful" aspects and activities that are "meaningful and valuable" for humans, in contrast to production and efficiency (Gaver, 2001). Peoples' lives should be enriched in new and pleasurable ways by new technologies. The focus is on new understandings of technology by exploring experiences and functions beyond the norm. New pleasures, new forms of sociability, and unexpected user requirements can be discovered. Cultural probes elicit information of members of a user group and shed light on users' social, emotional, and aesthetic values and habits (Blythe et al., 2003).

Gaver et al. (2004) consider the probes as collections of evocative tasks provoking inspirational responses from the user groups. Thereby, cultural probes are packages of materials given to the participants. The participants use these materials to tell about their lives over a period of time and return the probes to the researchers. Within a probe package, participants can find postcards with images on the front or questions at the back about attitudes towards their lives, environments, or technology (e.g., "Tell us about your favorite device"), maps where they can mark zones (e.g., the zone where they would go to meet people), a camera to take pictures (e.g., of their home, their clothes, something desirable), or a photo album and media diary (e.g., for illustrating their past, current life, or anything meaningful). In this way, fragmentary data are gathered over a period of time providing subjective but inspirational glimpses into the lives and situational contexts of the participants (Boettcher, 2006).

Rather than defining a final set of requirements, the "inspirational data" gathered with the cultural probes stimulate the designers' imagination (Gaver et al., 1999). What should be reached is not an objective view of users' needs but a more impressionistic account of their desires and needs. Against the background that knowledge has limits, Gaver et al. (2004, p. 53) even stress that their "approach values uncertainty, play, exploration, and subjective interpretation as ways of dealing with those limits." Summarizing cultural probe data and using these data for classical requirements analysis seems to contradict Gaver's intentions. He argues against "asking specific questions" and "rationalizing" the probes and "summarizing returns" as this will lead to a loss of valuable information (Gaver et al., 2004). Nevertheless, methods were developed from moving cultural probe data toward classical requirements engineering considering subjectivity and interpretation [e.g., inclusion of agent-oriented software engineering (AOSE) in the cycle of cultural probe observation to production (Boettcher, 2006)].

## 3  ANALYSIS AND MANAGEMENT OF USER REQUIREMENTS

Most of the above-mentioned approaches deliver results that describe the user's needs, situations, and contexts in depth. Their goal is to give a vivid impression of the users and opportunities to assist them with the technology. In this way perhaps partial but valid information is

delivered which can inspire and guide the design work. Many user-oriented approaches do not claim to provide systematic requirements engineering: For example, Gaver et al. (2004) explicitly state that a systematic set of requirements is not delivered in the sense that working through the list will guarantee a satisfying product. In complex product development, user requirements will be integrated in systematic engineering approaches (e.g., Boettcher, 2006) to solve typical problems (e.g., Aurum and Wohlin, 2005; Kotonya and Sommerville, 1998; ISO, 2010b): User requirements must be balanced and negotiated with stakeholders; trade-offs, inconsistencies, and ambiguities must be solved; abstract goals must be elaborated to feasible design and technical solutions; priorities must be determined in case of restricted time and budget; and requirements must be traceable and validated.

### 3.1 Documentation Formats

A first step in working with user requirements is to record them in a standardized way. If empirical results serve as an input, for example, from interviews or observations, standardized documentation is a great first step in the interpretation and reduction of data. In most cases, raw empirical data will not have the quality needed to support decisions of developers, designers, managers, and whoever else uses the requirements documentation. Diverse standards formulate quality aspects for well-formed requirements and documents [e.g. Institute of Electrical and Electronics Engineers/American National Standards Institute (IEEE/ANSI) 830-1998]. ISO/IEC WD 25065 (ISO, 2010b) recommends a clear syntax to ensure the suitability for design and the verifiability of a user requirement. Different levels of abstraction are suggested, from the presence or absence of a particular system feature to quantifiable measurements, that is, in terms of user behavior, system response times, and so on.

A requirements specification document in practice may comprise diverse formats and types of information. Textual and narrative formats are often used in the beginning of projects while in later stages of the requirements elaboration more formalized formats may be used (Kotonya and Sommerville, 1998). In particular, functional requirements can be specified in task-based formats such as use cases. In addition to the differences, many approaches have some writing guidelines in common: for example, avoid giving detailed information on technical realization or user interface design but focus on user goals and system responsibility to achieve them (see, e.g., Constantine and Lockwood, 1999; Beyer & Holtzblatt, 1998).

Narrative formats such as *problem scenarios* and *claims* in scenario-based-design (see above) do not follow these recommendations. They do not specify features of a particular system, as their primary goal is to express user requirements by describing the needs and opportunities in the current situation. Similarly, *user stories* (e.g., Ramsin and Paige, 2008) are narrative descriptions of use situations. They are characterized by low levels of detail and should be as short as possible (Cohn, 2004). User stories should give an overall picture of the users' needs and values of the functionality as well as estimate development efforts but avoid giving too much information (e.g., on interaction flows). Again, the format reflects the status and role of requirements in the particular approach, in this case agile development (see below).

In contrast to narrative formats, *use cases* typically describe functional requirements in a more detailed manner. They focus on tasks and have a formal syntax, often as diagrams and tables. Typically, use cases comprise a task or function title and a sequence of system–user interactions. Use case models are mostly procedurally structured so that within a use case other sub–use cases can be called conditionally or unconditionally [e.g., the prominent format of Unified Modeling Language (UML) use cases in the unified process; Booch et al., 1998]. Some approaches stress the relation to other models in the software development process, such as the link of essential use cases (Constantine and Lockwood, 1999) to view models in graphical user interfaces. Many other task-related formats also follow a formalized syntax (e.g., using task diagrams).

### 3.2 User Requirements in Engineering Processes

In software engineering, the management of requirements is a central aspect described in all established development process models. Requirements are always the starting point in a project, but often requirements are changing or need to be detailed during the process. Since the beginning of formalized process descriptions, the question of how to deal with that problem in complex system development was discussed and "waterfall"-like development with successive steps was criticized as inappropriate (Royce, 1970).

Process models mostly define sequential phases of activities but foresee controlled steps back in order to correct or substantiate requirements specifications or other previous project results. Typically, an initial phase is dedicated to the collection, analysis, and documentation of requirements. In earlier models, iterations to change them are treated as rather exceptional cases (e.g., in object modeling technique; Rumbaugh et al., 1991). So-called evolutionary or incremental approaches (Bunse and von Knethen, 2008) implement preplanned iterations of analysis, development, prototype, and risk reviews (e.g., the spiral model; Boehm, 1988). In the iterations, different increments of the software system are focused, covering different parts of requirements. Incremental development often incorporates a step-by-step refinement of requirements. In the first phase ("inception") of the unified process (Booch et al., 1998), for example, only overall requirements on the business level are identified. The majority of requirements should be specified only in the second phase ("elaboration").

Together with iterative refinement, many process models build on prototyping in order to solve the problem of initially imprecise or unclear requirements. While requirements leave freedom for design, prototypes are concrete and can be used for evaluation and validation also for customers and users. This also is one of the

core ideas in human-centered design (ISO, 2010c), the central approach to implement usability engineering in software development. Human-centered design involves end users throughout the entire software development process by means of evaluations of system prototypes. In repeated iterations, requirements (and system design) are then refined on the basis of evaluation results.

Iterative and other "traditional" software development processes have been criticized as being inefficient for volatile projects dealing with fast-changing organizations, markets, and technologies. Predefined iterations are regarded as inflexible. Extensive specifications for requirements are considered to cause unjustifiable efforts of adapting, updating, and communication (see, e.g., Batra, 2009; Sillitti and Succi, 2005. Since the late 1990s, *agile software engineering approaches* such as extreme programming (Beck, 2003), Scrum (Schwaber, 2004), and Crystal (Cockburn, 2004) have tried to overcome these problems by focusing on communication and cooperation (Jiang and Eberlein, 2008). The "lightweight" management of requirements plays a key role in achieving the goal of effective processes: Lean description formats, in particular user stories, are used in the early phases to get an overall picture of the users' needs and values of a functionality and to estimate development efforts. Later, they are refined and elaborated as necessary by customers and users. By gathering information not before needed, no dispensable efforts are spent, and the current situation of the project is always considered. In Scrum, the project roles dealing with requirements are defined: A *product owner* representative of the customer is responsible for prioritizing requirements and passing them on to the executing team who decides which requirements to take and when. This approach of lean requirement documentation to reduce efforts and ensure direct, well-fitting information flow is still under discussion. While the potential strengths are obvious, it appears that the great responsibility of project stakeholders to ensure continuous and sound input also bears risk, for example, for requirements retention (Kelly, 2010).

## 4   CONCLUSIONS

Collecting and analyzing user requirements comprise the first and important step in the process of developing an interactive system. A deep understanding of contextual conditions and forces, user tasks and activities, and user needs is a major precondition for designing successful and user-friendly systems.

This chapter has presented a selection of methods for user requirements collection and analysis. The differences between the methods described here reflect different approaches and perspectives which can be taken in the early phases of system development. Task analysis is a very well-structured and analytical approach which fits directly into traditional engineering processes. It focuses on user activities, whereas other methods emphasize the importance of physical, organizational, and social contexts and environments (e.g., ethnography, contextual inquiry). These methods have a more

sociocultural origin. Methods such as scenario analysis have their strengths in supporting an easy communication between project stakeholders. Compared to more analytical approaches, these narrative methods can establish a vivid impression of use scenarios while they make it more difficult to cover requirements in a concise and complete manner. Cultural probes go one step further. Their purpose is not a complete or structured set of requirements. Cultural probes are appreciated for their exploratory and playful character, which supports inspiration and stimulates the designers' imagination. Requirements collection methods also differ in the extent to which they involve end users and rely on empirical data. In most cases, a combination of more analytical/theoretical (e.g., task analysis, scenario analysis) and empirical methods such as focus groups or contextual inquiries will yield best results.

To achieve the best value of collected and analyzed requirements, requirements must be effectively incorporated into the design and development process. The formats, in which requirements are delivered into the design process, highly influence the potential uptake by other project members. This chapter has discussed different formats from narrative to formal descriptions. From a modern perspective which appreciates lean and agile processes, it is not a question of preference between detailed formal descriptions and narratives but rather a question of when (in the development process) to use which format. In the early phases of the project, narrative user stories can support a common understanding of high-level objectives and general user needs. Later in the process, more elaborate and formal requirements descriptions will be needed in order to inform detailed design and development activities. The management of user requirements in today's development processes faces the challenge of permanent changes and refinements. Iterative design optimizations have to be supported and reflected on by iterative refinements on the level of user requirements.

## REFERENCES

Annett, J., Duncan, K., Stammers, R., and Gray, M. (1971), *Task Analysis*, Her Majesty's Stationery Office, London.

Aurum, A., and Wohlin, C., Eds. (2005), *Engineering and Managing Software Requirements*, Springer, Berlin.

Batra, D. (2009), "Modified Agile Practices for Outsourced Software Projects," *Communications of the ACM*, Vol. 52, pp. 143–148.

Beck, K. (2003), *Extreme Programming*, Addison-Wesley, New York.

Beyer, H., and Holtzblatt, K. (1998), *Contextual Design: Defining Customer-Centered Systems*, Morgan Kaufmann, San Francisco.

Blomberg, J. (1995), "Ethnography: Aligning Field Studies of Work and System Design," in A. F. Monk and G. N. Gilbert, Eds., *Perspectives on HCI: Divers Approaches*, Harcourt Brace & Company, New York, pp. 175–198.

Blythe, M., Baxter, G., Wright, P., Cheverst, K., Clarke, K., Dewsbury, G., Rouncefield, M., and Sommerville, I. (2003), "Gathering Requirements for Inclusive Design,"

paper presented at the Second BSC HCI Group Workshop on Culture and HCI, Bridging Cultural and Digital Divides, pp. 65–71.

Bødker, S. (1991), *Through the Interface: A Human Activity Approach to User Interface Design*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Boehm, B.W. (1988), "A Spiral Model of Software Development and Enhancement," *IEEE Computer*, Vol. 21, pp. 61–72.

Boehner, K., Vertesi, J., Sengers, P., and Dourish, P. (2007), "How HCI Interprets the Probes," in *Proceedings CHI 2007*, ACM Press, New York, USA, pp. 1077–1086.

Boettcher, A. (2006), "Moving from Cultural Probes to Agent-Oriented Requirements Engineering," in *Proceedings of OZCHI*, ACM, Melbourne, Australia.

Booch, G., Jacobson, I., and Rumbaugh, J. (1998), *The Unified Software Development Process*, Addison-Wesley, Amsterdam, The Netherlands.

Bunse, C., and von Knethen, A. (2008), *Vorgehensmodelle kompakt*, Spektrum Akademischer Verlag, Heidelberg.

Caplan, S., (1990), Using focus group methodology for ergonomic design. *Ergonomics* 33 (5), pp. 527–533.

Carroll, J. M. (2000), Making use: Scenario-based design of human-computer interactions, MIT Press, Cambridge, MA.

Carroll, J. M. (2002), Making Use is more than a matter of task analysis. *Interacting with Computers 14 (5)*, pp. 629–637.

Carroll, J. M., and Rosson, M. B. (1990), "Human-Computer Interaction Scenarios as a Design Representation," in *Proceedings of the 23rd Annual Hawaii International Conference on Systems Sciences*, Kailua- Kona, HI, January 2–5, Lee W. Hoevel, Ed., IEEE Computer Society Press, Los Alamitos, CA, pp. 555–561.

Chin, G., Rosson, M. B., and Carroll, J. M. (1997), "Participatory Analysis: Shared Development of Requirements from Scenarios," in *Proceedings of Human Factors in Computing Systems, CHI'97Conference*, Association for Computing Machinery, New York, pp. 162–169.

Cockburn, A. (2004), *Crystal Clear a Human-Powered Methodology for Small Teams*, Addison-Wesley, New York.

Cohn, M. (2004), "Effective User Stories," in *Extreme Programming and Agile Methods: XP/Agile Universe*, Carmen Zannier, Hakan Erdogmus, Lowell Lindstrom, Eds., Springer, Berlin, pp. 525–540.

Constantine, L., and Lockwood, L. (1999), *Software for Use*, Addison-Wesley, New York.

Crabtree, A., Hemmings, T., Rodden, T., Cheverst, K., Clarke, K., Dewsbury, G., and Rouncefield, M. (2003), "Designing with Care: Adapting Cultural Probes to Inform Design in Sensitive Settings," in *Proceedings of OZCHI 2003, Brisbane, Australia: Ergonomics Society of Australia*, pp. 4–13.

Crystal, A., and Ellington, B. (2004), "Task Analysis and Human-Computer Interaction: Approaches, Techniques, and Levels of Analysis," in *Proceedings of the Tenth Americas Conference on Information Systems,* Association for Information Systems, New York, August 2004.

Darroch, V. and Silvers, R. (Eds.), (1982). *Interpretive Human Studies: An Introduction to Phenomenological Research.* Washington, D.C.: University Press of America.

Davis, A. M. (1993), *Software Requirements: Objects, Functions and States*, Prentice Hall, Englewood Cliffs, NJ.

Diaper, D. (1989), "Task Observation for Human-Computer Interaction," in *Task Analysis for Human-Computer Interaction*, D. Diaper, Ed., Ellis Horwood Ltd, Chichester, England, pp. 210–237.

Diaper, D. (2002), "Scenarios and Task Analysis," *Interacting with Computers*, Vol. 14, No. 4, pp. 379–395.

D'Souza, M. E., and Greenstein, J. S. (2003), Listening to Users in a Manufacturing Organization: A Contextbased Approach to the Development of a Computer-Supported Collaborative Work System," *International Journal of Industrial Ergonomics*, Vol. 32, pp. 251–264.

Ericsson, K. A., and Simon, H. A. (1993), *Protocol analysis: Verbal reports as data,* (Rev. ed.), MIT Press, Cambridge, Massachusetts.

Fowler, F. J. Jr., and Mangione, T. W. (1990), *Standardised Survey Interviewing*, Sage Publications, Newbury Park.

Garmer, K., Ylven, J., and Karlsson, I. C. M (2004), "User Participation in Requirements Elicitation Comparing Focus Group Interviews and Usability Tests for Eliciting Usability Requirements for Medical Equipment: A Case Study," *International Journal of Industrial Ergonomics*, Vol. 33, pp. 85–98.

Gaver, W. (2001), "Designing for Ludic Aspects of Everyday Life,"*ERCIM News*, No. 47, available: http://www.ercim .org/publication/Ercim_News/enw47/gaver.html.

Gaver W., Dunne, A., and Pacenti, E. (1999), "Cultural Probes", *Interactions*, Vol. 6, No. 1, pp. 21–29.

Gaver, W., Boucher, A., Pennington, S., and Walker, B. (2004), "Cultural Probes and the Value of Uncertainty," *Interactions*, Vol. 11, pp. 53–56.

Go, K., and Carroll, J. M. (2003), "Scenario-Based Task Analysis," in *The Handbook of Task Analysis for Human-Computer Interaction*, D. Diaper and N. Stanton, Lawrence Erlbaum Assoc Inc Eds., New Jersey pp. 117–134.

Go, K., and Carroll, J. M. (2004), "The Blind Men and The Elephant: Views of Scenario-Based System Design," *ACM Interactions*, Vol. 11, No. 6, pp. 44–53.

Go, K., and Carroll, J. M. (2006), "TEAUS: Task Exploration and Analysis Using Scenarios," *Bulletin of the YINS*, Vol. 10, available: http://ci.nii.ac.jp/naid/110006283075/en/.

Greenbaum, T. L. (1998), *The handbook for focus group research* (2nd ed.), Sage Publications, Thousand Oaks, California.

Holtzblatt, K., and Beyer, H. (1993), "Making Customer-Centered Design in Work Teams," *Communications of the ACM*, Vol. 36, pp. 93–103.

Holtzblatt, K., and Jones, S. (1993), "Contextual Inquiry: A Participatory Technique for System Design", in *Participatory Design: Principles and Practice*, A. Namioka and D. Schuler, Eds., Lawrence Erlbaum, Hillsdale, NJ, pp. 177–210.

Hooper, J. W., and Hsia, P. (1982), "Scenario-Based Prototyping for Requirements Identification," *Software Engineering Notes*, Vol. 7, No. 5, pp. 88–93.

Institute of Electrical and Electronics Engineers (IEEE) (1998), *IEEE Guide to Software Requirements Specification*, ANSI/IEEE Std. 830-1998, IEEE Press, Piscataway, NJ.

International Organization for Standardization (ISO) (2009), *Systems and Software Engineering—Vocabulary*, ISO/IEC 24765, ISO, Geneva.

International Organization for Standardization (ISO) (2010a), *Systems and Software Engineering—Software Product Quality Requirements and Evaluation (SQuaRE)—Common Industry Format (CIF) for*

*Usability: User Needs Report, Working Draft*, ISO/IEC WD 25064, ISO, Geneva.

International Organization for Standardization (ISO) (2010b), *Systems and Software Engineering—Software Product Quality Requirements and Evaluation (SQuaRE)—Common Industry Format (CIF) for Usability: User Requirements Specification, Working Draft*, ISO/IEC WD 25065, ISO, Geneva.

International Organization for Standardization (ISO) (2010c), *Ergonomics of Human System Interaction—Part 210: Human-Centred Design for Interactive Systems*, ISO/IEC 9241-210: ISO, Geneva.

Jiang, L., and Eberlein, A. (2008), "Towards a Framework for Understanding the Relationships between Classical Software Engineering and Agile Methodologies," in *Proceedings of the 2008 international Workshop on Scrutinizing Agile Practices Or Shoot-Out At the Agile Corral APOS '08*, Kruchten, P., and Adolph, S. (Eds.) Association for Computing Machinery, New York, pp. 9–14.

Kelly, S. (2010), "Towards an Evolutionary Framework for Agile Requirements Elicitation," in *Proceedings of the 2nd ACM SIGCHI Symposium on Engineering Interactive Computing Systems* Sukaviriya, N. and Vanderdonckt, J., Eds, Berlin, Germany, Association for Computing Machinery, New York, NY, USA, pp. 349–352.

Kirwan, B., and Ainsworth, L. K (1992), *A Guide to Task Analysis*, Taylor and Francis, London and New York.

Kotonya, G., and Sommerville, I. (1998), *Requirements Engineering*, Wiley, New York.

Krueger, R. A., & Casey, M. A. (2000), *Focus groups: A practical guide for applied research* (3rd ed.), Sage Publications, Thousand Oaks, California.

Kuniavsky, M. (2003), *Observing the user experience: A practitioner's guide for user research*, Morgan Kaufmann, Amsterdam.

Levine, J. M., and Moreland, R. L. (1998), "Small Groups," in *Handbook of Social Psychology*, 4th ed., Vol. 2, D. Gilbert, S. Fiske, and G. Lindzey, Eds., McGraw-Hill, Boston, pp. 415–469.

Martin, J. L., Murphy, E., Crowe, J. A., and Norris, B. J. (2006), "Capturing User Requirements in Medical Device Development: The Role of Ergonomics," *Physiological Measurement*, Vol. 27, pp. 49–62.

Nardi, B. A. (1993), A *Small Matter of Programming: Perspectives on End User Computing*, MIT Press, Cambridge, MA.

Nardi, B. A. (1996), *Context and Consciousness: Activity Theory and Human-Computer Interaction*, MIT Press, Cambridge, MA.

Nardi, B. A., Vicki L. and O'Day, V. L. (1999), *Information Ecologies: Using Technology with Heart*, MIT Press, Cambridge. MA.

Nielsen, J. (1993), *Usability engineering*, Morgan Kaufmann Publishers, San Francisco, California.

Ramsin, R., and Paige, R. F. (2008), "Process-Centered Review of Object Oriented Software Development Methodologies," *ACM Computing Survey*, Vol. 40, pp. 1–89.

Raven, M. E., and Flanders, A. (1996), "Using Contextual Inquiry to Learn About Your Audiences," *Asterisk Journal of Computer Documentation*, Vol. 20, No. 1, pp. 1–13.

Rolland, C., Achour, C. B., Cauvet, C., Ralyte, J., Sutcliffe, A., Maiden, N. A. M., Jarke, M., Haumer, P., Pohl, K., Dubois, E., and Heymans, P. (1996), "A Proposal for a Scenario Classification Framework," *Requirements Engineering*, Vol. 3, No. 1, pp. 23–47.

Rosson M. B., and Carroll, J. M. (2002), *Usability Engineering: Scenario-Based Development of Human-Computer Interaction*, Morgan Kaufmann, San Diego.

Rosson, M. B., and Carroll, J. M. (2002), "Evolving Technologies and Emerging Applications," in *The Human-Computer Interaction Handbook: Fundamentals*, J. Jacko and A. Sears, Eds., Lawrence Erlbaum Associates, Mahwah, New Jersey Chapter 53.

Royce, W. (1970), "Managing the Development of Large Software Systems," in *Proceedings of IEEE WESCON 26*, pp. 1–9, available: http://www.cs.umd.edu/class/spring 2003/cmsc838p/Process/waterfall.pdf

Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., and Lorensen, W. (1991), *Object Oriented Modeling and Design*, Prentice Hall, New Jersey

Schwaber, K. (2004), *Agile Project Management with Scrum*, Microsoft Press, Washington.

Shepherd, A. (1989), "Analysis and Training in Information Technology Tasks," in *Task Analysis for Human-Computer Interaction*, D. Diaper, Ed., Ellis Horwood Ltd, Chichester, England, pp. 15–55.

Sillitti, A., and Succi, G. (2005), "Requirements Engineering for Agile Methods," in *Engineering and Managing Software Requirements*, A. Aurum and C. Wohlin, Eds., Springer, Berlin.

Sommerville, I. (2004), *Software Engineering*, 7th ed., Pearson Education, Harlow, England.

Sommerville, I., Bentley, R.,. Rodden, T., and Sawyer, P. (1994), "Cooperative Systems Design," *Computer. Journal*, Vol. 37, pp. 357–366.

Spinuzzi, C. (2000), "Investigating the Technology-Work Relationship: A Critical Comparison of Three Qualitative Field Methods," in *Proceedings of IEEE Professional Communication Society International Professional Communication Conference* and *Proceedings of the 18th Annual ACM International Conference on Computer Documentation: Technology & Teamwork* (Cambridge, MA), IEEE Educational Activities Department Piscataway, New Jersey, pp. 419–432.

Stewart, D. W., Shamdasani, P. N., and Rook, D. W. (2006), *Focus Groups: Theory and Practice*, Sage, Thousand Oaks, CA.

Viller, S., and Sommerville, I. (1999), "Coherence: An Approach to Representing Ethnographic Analyses in Systems Design," *Human-Computer Interaction*, Vol. 14, pp. 9–41.

Wales, R., O'Neill, J., and Mirmalek, Z. (2002), "Ethnography, Customers, and Negotiated Interactions at the Airport," *IEEE Intelligent Systems*, Vol. 17, pp. 15–23.

Wixon, D. R., Ramey, J., Holzblatt, K., Beyer, H., Hackos, J., Rosenbaum, S., Page, C, Laakso, S. A., and Laakso, K.-P. (2002), "Usability in Practice: Field Methods Evolution and Revolution," in *Proceedings of CHI 2002 Extended Abstracts* Conference on Human Factors in Computing Systems, Minneapolis, Minnesota, April 20–25, 2002, Association for Computing Machinery, New York.

# CHAPTER 48

# WEBSITE DESIGN AND EVALUATION

**Kim-Phuong L. Vu**
California State University–Long Beach
Long Beach, California


**Robert W. Proctor**
Purdue University
West Lafayette, Indiana


**Fredrick P. Garcia**
Center for Usability in Design and Accessibility
Long Beach, California

## 1 INTRODUCTION

Virtually all users have encountered examples of good and bad sites in their interactions with the World Wide Web. Poor Web designs often lead to user frustration because users cannot easily access the information they are seeking, and this frustration may cause them to abandon the site. Having users abandon a site is undesirable in almost any case, as it defeats the purpose of disseminating information or providing services on the Web. Moreover, abandonment of a site is particularly harmful to e-businesses because it results in the loss of potential customers. Organizations that are successful in developing a good website and maintaining it will thus have a competitive edge over their rivals.

The first question to ask when designing a website should be: What is the goal of the website? Is the site's goal to sell the most widgets that company XYZ produces? Is it to provide a resource where scientific information relating to a certain topic can be found? Or, is it to be an entertainment site that users can visit in their recreation time? Although the task of posing this question may be simple, the answer is critical for determining the nature of the website as well as for driving the design and evaluation process for the site (see, e.g., Vu et al.,

2011). Clearly defining the goals for the website helps the designer focus on those aspects that are important to attaining the goal and determining the emphasis that is to be placed on each component.

Several basic types of websites are commonly encountered (Irie, 2004): information dissemination, portal, social networking/community, search, e-commerce, company information, and entertainment. Each of these types of websites is designed for specific purposes or goals and has characteristics unique to achieving those goals:

- News/information dissemination
  - *Goal*: to provide users with news or information
  - *Characteristic*: mostly text based, with minor graphics; simple and consistent navigation
  - *Example*: www.npr.com—NPR is a website that provides information about noncommercial news, talk, and entertainment programs.
- Portal
  - *Goal*: to provide links to other websites or resources
  - *Characteristic*: mostly uniform resource locator (URL) links, with minor descriptions of

the linked resources; usually organized alphabetically or by a keyword or theme

- *Example*: www.merlot.org—The Multimedia Educational Resource for Learning and Online Training (MERLOT) website provides free access to online learning materials for faculty and students of higher education.

- Social networking/community/communication
  - *Goal*: to provide a medium that promotes the community and provides opportunity for communication and interaction among users
  - *Characteristic*: usually in the form of message boards or chat rooms
  - *Example*: www.facebook.com—Facebook is a social networking site that allows individuals to set up profile pages to connect and share information with other people in their life.

- Search
  - *Goal*: to facilitate retrieval of specific information or resources
  - *Characteristic*: usually in the form of a search engine; the returned results page resembles a portal
  - *Example*: www.google.com—The Google website consists of a search engine that allows users to find information relating to a variety of topics.

- E-commerce
  - *Goal*: to allow companies to sell products or services; to allow users to purchase products or services
  - *Characteristic*: usually consists of a searchable catalog of products; must include a mechanism for secure online monetary transactions
  - *Example*: www.amazon.com—Amazon is a website that sells books and other products to users.

- Company/organization/product information
  - *Goal*: to provide specific information about a company, organization, or product
  - *Characteristic*: emphasis on company's or product's image/logo
  - *Example*: www.apple.com—The Apple website features products and services provided by the company, such as the iPhone4 and iPad.

- Entertainment
  - *Goal*: to provide pleasant interactions or entertainment resources for users
  - *Characteristic*: usually places emphasis on aesthetics; games and videos
  - *Example*: www.d-9.com—The District 9 website features video trailers and advertisement for the movie.

Once the goal for the website has been established, the next step is to determine what content should be placed in the site (Liao et al., 2010). It is important to distinguish content from aesthetics. Content design focuses primarily on the substance or information that is contained in the website, whereas aesthetics design focuses primarily on making the site visually pleasant or enjoyable. A good website can be informative as well as pretty and creative, which means that designing for content and aesthetics is not mutually exclusive. However, because the goal of many websites is to provide some type of service, it is usually more important to design for content than for aesthetics.

## 2 WEBSITE CONTENT

Websites should be designed with the end users in mind at all times because the sites are intended to support user activities (Proctor and Vu, 2010). The design process for Web interfaces proceeds in a manner similar to traditional usability engineering life cycles, which include a requirements analysis phase, a design, testing, and development phase, and an installation phase (see, e.g., Mayhew, 2011). It is important to note that design, evaluation, and development are iterative processes. That is, the design of the website should be evaluated as it is being developed so that usability problems can be identified and fixed. However, for ease of presentation, Web design issues are covered in this section and Section 3 of the chapter, and Web evaluation techniques are covered separately in Section 4.

### 2.1 Components of a Website

There are several major components of a website: its content, architecture and organization, the presentation of the content, and the programming logic that is used to integrate the content and its presentation within the site's structure. Content design and presentation can be broken down even further into specific components, such as page design, navigation, use of multimedia, search design, and URL design (see, e.g., Nielsen, 2000). On the recommendation of the World Wide Web Consortium (W3C), the international organization that creates Web standards [HyperText Markup Language (HTML), Extensible Markup Language (XML), etc.], Web designers have begun to programmatically separate content from presentation as much as possible. The main reason for this change is to build more flexibility into websites. When content is separated from presentation, it allows the content to be rendered on many different devices (e.g., mobile devices), displayed in many different formats (e.g., different color schemes), and changed quickly and easily. Separating content from presentation requires that the designer provide proper structure for content by marking it up with appropriate HTML tags (e.g., <p>) and employ cascading style sheets (CSSs) for presentation. The separation of content from presentation is also necessary for creating accessible websites because assistive technologies use the structural HTML tags to convey information to the user and present the content in accessible formats. Examples of how the same content marked up with identical HTML tags can be presented in many different ways with CSSs are available at the CSS Zen Garden (http://www.csszengarden.com/).

One way to determine the process involved in website design is to look at common design practices.

Newman et al. (2003; see also Newman and Landay, 2000) conducted a study in which 11 expert website designers were interviewed about various Web design projects to determine the practices in which Web designers customarily engage during the design process. They identified several specific areas to which the expert designers referred when describing the design process of a website:

- *Information design* includes identifying and grouping content so that individual components can be integrated and organized into a coherent whole.
- *Navigation design* includes methods for users to move through or access different parts of the website.
- *Graphics design* includes how to present individual pieces of information or content visually to the users (e.g., through images).
- *Information architecture* includes how to combine the information and navigation components so that the entire website functions as a unified entity.
- *User interface design* refers to designing and evaluating the usability of the website, including its informational and navigational components.

Although the areas listed above emphasize the major ones of concern for Web design, as identified by several Web design experts, there is some degree of overlap between the different areas. Newman et al. (2003) noted that most of the time the designers indicated that work in the areas of information design and navigation design preceded work on the graphics design. Because the Web consists of many individual components, designers should use the goal of the website to determine which components are more critical than others.

## 2.2 Content Preparation

The Web provides a medium for exchange of information and services to a global audience. Because of the potential impact that the Web can have on the success of an organization or company, it is important to have an effective content design that promotes the goals of the website. Unfortunately, it is often the case that websites are not designed in an effective manner. For example, although the goal of an e-commerce website is to sell the products, one study showed that users were not able to find specific items on e-commerce websites 36% of the time (Nielsen et al., 2000). After ten years, though, the situation does not seem to have improved much. The Web analytics firm iperceptions (2010) reported, in a survey of immediate postexperience feedback from over 400,000 e-commerce users, that 47% of users could not find what they were looking for at a site. If users cannot find an item, they cannot buy it. Thus, e-commerce websites that are designed to structure and organize their content in a manner that promotes the ease with which users can locate specific items will have a competitive edge over rival websites designed to achieve the same goal.

Content preparation refers to the processes involved in determining the information for the website to convey and how to organize, structure, and present that information so that it can be retrieved easily and efficiently when needed. Proctor et al. (2002b) summarized four major areas that Web designers should emphasize for content preparation [see also Proctor et al. (2003) and Liao et al. (2010)]. These areas include:

1. *Knowledge elicitation*: determining what type of information should be conveyed
2. *Structure and organization of information*: determining the best method to structure and organize information
3. *Retrieval of information*: determining the best methods for helping users search and retrieve information
4. *Presentation of information*: determining the medium in which information should be presented to the user

## 2.3 Knowledge Elicitation

When designing for the Web, knowledge should be elicited from two classes of users: experts and end users (Proctor et al., 2002b). Experts in Web design can provide valuable information, including (1) how to organize and present information in a manner that is consistent with human information-processing capabilities, (2) the functions and features that good websites should possess, (3) methods for enhancing the website's efficiency or effectiveness; and so on. The information that end users provide is usually not expert advice regarding how to design a website, but rather, information that is intended to help designers understand (1) the computing skills or level of knowledge of the end users; (2) the users' mental models, or representations, of the content that is contained in the website; (3) the specific pieces of information that users need when performing a task; and so on. Another way to characterize the different information provided by experts and end users is that end users reveal the type of information needed by them to achieve their goals, whereas experts determine how to organize and present that information in the most effective manner.

### 2.3.1 Eliciting Knowledge from Experts

A lot of information can be gained about how to design for the Web from interviewing experts and observing their work. Interviews are a well-known knowledge elicitation technique (e.g., Shadbolt and Burton, 1995) and can be one of the best methods for obtaining in-depth data from experts. In the study by Newman et al. (2003) mentioned earlier, expert designers were interviewed extensively to obtain information about their thought processes and involvement during an entire design process. The experts were asked to walk the interviewer through each phase of the project, showing examples of their "work in progress" if possible. Newman et al. report in detail the steps involved in creating a tutorial for a suite of computer-aided design (CAD) tools that can be accessed through the intranet

within a company or through the Internet remotely. The key steps identified by the expert are as follows:

1. *Discovery phase*: gathering background information about the project to determine the project scope, goals, and timeline for completion
2. *Design exploration phase*: generating sketches of initial variations of the design, including how the content will be structured, the individual pages, navigation, and the interaction sequence
3. *Design refinement phase*: creating high-fidelity mock-ups of the site, including the home page and second-level pages that can be accessed from the home page; choosing a limited set of the mock-up designs for further refinement and development; selecting a single refined mock-up to prototype in HTML
4. *Production phase*: writing up guidelines for the prototype so that a design team can turn the prototype into a working product

As illustrated by Newman et al.'s (2003) study, interviews from experts and observation of artifacts of their work can provide valuable information about the design process. However, there are numerous other methods that can be used to elicit knowledge from experts. Many of these techniques are described by Proctor et al. (2002b) and include the following:

- *Verbal Protocol Analysis*. Analysis of an expert's problem-solving strategies through examining the verbal protocols, obtained by having the experts "think aloud" as they solve problems or reflect on their thought processes when reviewing recordings of their performance during a problem-solving task.

- *Group Task Analysis*. Analysis of a group of experts' joint depiction about how a specific task is represented and processed. Usually, a flowchart is used to depict the individual steps required for performing a specific task.

- *Narratives and Scenarios*. Analysis of information contained in stories or narratives that experts tell about their activities. They often reveal valuable information about the goals of a particular task and the sequence of actions and events that lead to particular decisions and outcomes.

- *Critical Incident Reports*. Analysis of critical incidents that tested a person's expertise for the insight they provide into the processes involved an expert's decision making and reasoning when an unexpected or unusual event occurs.

When interviewing design experts is not feasible, a designer should refer to published papers and chapters on specific Web design issues of interest written by experts in the field [see Ratner (2003), Bidgoli (2004), and Vu and Proctor (2011) for edited volumes on the Internet and Web design]. Dorn and Guzdial (2010) found that, in addition to traditional books, technical reports, and information exchange between colleagues, Web designers often consult the Web to obtain sample

codes/demos and to gain access to learning materials such as online walkthroughs/tutorials, forums/blogs, podcasts, and so on.

When it comes to evaluating expertise, one question that can be asked is: Who is the expert (Flach, 2000)? Should websites be designed according to the wishes of the expert designers or are the end users the expert in this case? After all, the end users are the people for whom the website is intended. Most Web professionals would agree that the input of the end users is important when designing a website and evaluating its effectiveness. However, the end users should not be allowed to direct the design process because the features they desire for the website may be hard to implement or may lead to a less than optimal design when all factors are considered.

Below, different techniques for eliciting knowledge from end users are discussed. Many of these methods and techniques are used in the exploratory phases of Web design to understand the end users and their goals, representations of the task, likes and dislikes, and preferences. As a result, the term *understanding the user* is used to refer to this process rather than *knowledge elicitation* (see, e.g., Volk et al., 2011).

### 2.3.2 Understanding the User

The Web affords the opportunity for a site to be accessed anywhere, anytime, and with various devices (e.g., computer, laptop, personal data assistant, and cell phone). As a result, in initial or exploratory phases of Web design, it is important to obtain background information about the targeted user population, including the users' cognitive and physical capabilities, the tasks that they are likely to perform, the information needed to perform those tasks, and their roles and responsibilities (Stanney et al., 1997). For example, if the website is targeted for use by older adults, it is recommended that designers avoid using colors from the short-wavelength end of the visual spectrum (i.e., blue and green) and increase the resolution of elements on the screen so that the items can be seen more easily (Bitterman and Shalev, 2004). Moreover, different types of users may need to access different parts of the website. For example, an e-commerce merchant may want to access the U.S. Postal Service's website to request or print out mailing labels or to schedule a package pickup, whereas both the merchant and the consumer may access the site to track a package. Because different users have unique goals, the website must be designed to accommodate the tasks that the different user groups want and need to perform.

There are many different types of methods that can be employed to understand the users (see Kuniavsky, 2008; Proctor et al., 2002b; Volk et al., 2011, for reviews). Most of these methods are aimed at obtaining information regarding what users need to complete their task on the website, along with users' preferences for the site's options and features. Below are brief descriptions of the major methods, along with their goals and characteristics.

***Interviews***   As with experts, end users can be interviewed to obtain in-depth knowledge about their characteristics, opinions, and preferences. There are two

main types of interviews: structured and unstructured. In a structured interview, the interviewer asks the users questions from a prearranged list. In an unstructured interview, the interviewer allows users to express their thoughts on a topic freely. Structure and unstructured are two ends of a continuum. Most interviews fall somewhere in between these two ends, with the interviewer asking general questions, but the direction of the interview is dependent on the user's answers. Interviews of end users provide a large amount of qualitative data in which the evaluators often organize into topics or perform a content analysis of the data to identify themes and categories of information. The interviewers should try to avoid imposing their beliefs or opinions during the process. Caution must be taken when framing the questions to avoid biasing the users' answers, and at least two different evaluators should code the data to ensure that their interpretations of the answers match to some degree.

**Surveys and Questionnaires** Surveys and questionnaires consist of questions used to gather information about a user (e.g., How old are you? How often do you make online purchases?), obtain users' likes and dislikes [e.g., On a scale of 1 (do not like it at all) to 10 (like it a lot), indicate how much you like pop-up windows], and preferences (e.g., Do you prefer a dark font color on a light background or a light font color on a dark background?). Some advantages of using surveys and questionnaires include (1) obtaining data from a large sample in different demographic areas or specified user groups (e.g., previous customers or persons on an email list) is relatively easy (there are now many online survey service providers, e.g., SurveyMonkey.com), (2) the data can be obtained in a relatively short period of time, (3) if there are no open-ended questions, the data can be coded and summarized relatively easily, and (4) data from questionnaires and surveys can also be used to develop user profiles (see Kuniavsky, 2008). However, some disadvantages of using surveys and questionnaires include (1) not all targeted users chose to fill them out and submit their answers; (2) it may be difficult to verify the identity of the participants who submit their answers through the Web; (3) if the default code for "no response" on an item is not coded as such, the default value may be mistaken as the user's response rather than no response; (4) the framing of questions may affect the validity of the results; and (5) users' judgments and preferences may not correlate with their actual performance.

**Focus Groups** Unlike surveys and questionnaires, focus groups consist of a smaller number of users (usually, 5–10) but allow the users to interact with one another when discussing and evaluating different aspects or issues of the design. Problems and concerns about the website that may not have been identified by a single user can emerge during the group's interactions. A moderator usually directs the group to ensure that everyone participates and stays on the task so that all the topics wanting to be covered will get covered and that no single participant dominates the session. Proctor et al.

(2002b) noted that focus groups are best used to attain high-level goals, such as generating a list of functions or features for a product. Focus groups yield large amounts of qualitative data that must be organized and summarized. As with the previous methods, focus groups also suffer from the fact that the design judgments and preferences produced by the group may not correspond to designs that benefit performance. Furthermore, since the tasks and interface are evaluated in a context different from what the user is likely to experience during "real" use of the website, issues identified in the focus group may not always be of items of major concern.

**Naturalistic Observation** Observation and nonparticipatory contextual inquiries reflect more naturalistic techniques for understanding users. These methods rely on observing users' everyday interactions with the website in their natural surroundings and provide researchers with background information about the context in which a product is being used. This can be done by having researchers observe users performing Web tasks from "afar." It is important that the evaluator remain unnoticed, so that users' natural behaviors can be observed. Often, it is difficult for evaluators to remain unobtrusive. However, this can be done through the use of one-way mirrors or video cameras. Connecting rooms with one-way mirrors are often found in usability labs (described later) but do not capture the users in their natural environment unless they access the website in public locations (e.g., a library). Video cameras (or Web cameras) can be set in public locations to observe users. However, it is important that these cameras not be visible to users because they may change their interaction pattern if they know that they are being observed. Because data from observational methods are based on the users' actions, the conflict between users' verbal report and their actions is no longer an issue. However, it is often difficult for the observer to remain unnoticed, and the data obtained from the observer may reflect his or her biases and/or interpretation of the observed events.

**Ethnographic Studies** With ethnographic studies, which emerged from the field of anthropology, the researcher seeks to understand the users by immersing himself or herself in the targeted users' culture or work environment (Millen, 2000). The goal of the researcher is to become a natural member of the group so that he or she will be able to understand the views of the user groups and work with them to design products that meet their needs. Ethnographic studies can yield a customer–partner relationship that results in an effective medium for identifying the users' needs (Volk et al., 2011). However, there are many drawbacks to ethnographic methods as well. Ethnographic studies take a long time to conduct, suffer from the same disadvantages of interpretation and bias as other methods based on self-report or observational data, and may not result in general guidelines for product designs since the data are based on a very specific group of people.

**User Diary** The user diary allows the user to observe and record, in a diary, his or her actions with a product over a period of time. A typical diary study involves

asking participants to keep a diary of their daily activities with the product for a few days or weeks. However, Rieman (1993) noted that is often inconvenient for participants to keep a diary for more than two weeks. The diary method is based on a real-time tracking system. Diary logs can reveal qualitative data such as "critical incidents" that have occurred as well as descriptive data such as the time spent on each task. The disadvantage of traditional diary methods is that they rely on the conscientiousness of the participants, which may result in diaries of different qualities. Participants may forget to log the activities, may log information that they think the evaluators want, or provide vague or general summaries. However, with the availability of wireless connections, an online or video diary can be obtained by wearing a wireless video camera or Web camera that records the users' interaction with different websites. The video diary circumvents the traditional problems of users waiting until a convenient time to enter the data in their diary.

***Web Server Log Files***   A benefit of collecting data about website usage is that the site itself can be used as a data collection tool because the actions of the users are recorded as the users interact with the website. Server logs can provide designers valuable information about existing websites, such as "who" is visiting the site, the pages within the site that the users visit and the order in which those pages are accessed, how long the user spends looking at a particular page, and what items users search for (Pearrow, 2000). The benefits of evaluating data from log files are that a large amount of data can be obtained from users who access the site and the data collection process does not interfere with the users' interaction with the site. However, because information irrelevant to the design goals is logged, it may be difficult to sort through log files and important variables of interest may not have been logged. Zeng and Duan (2011) describe a technique for analyzing user activity such as clicking (analyzing mouse click sequences), selection (collecting Web postings), and propagation (estimating where users are likely to go). These activities can then be used in models to capture user Web behaviors.

### 2.3.3   Summary

This section summarized the major methods that are used to determine the processes involved in the website's design life cycle and how to extract the content to be conveyed by the website from experts and end users. Although these knowledge elicitation methods can yield substantial insight and data regarding the knowledge, expertise, and characteristics of designers and end users, the main drawback is that the methods rely on self-report data. Self-report data are vulnerable because there are well-known biases in self-reports (e.g., Isaacs, 1997), not all cognitive processes can be articulated clearly [see Anderson (1982) for a discussion of declarative versus procedural knowledge], analysis is based on the interpretation of the evaluators (see Proctor et al., 2002b; Volk et al., 2011), and users' preferences and judgments may not correlate with their performance (e.g., Bailey, 1993; Nielsen, 2001; Vu and Proctor, 2003).

### 2.4   Structuring and Organizing the Content

Once designers have identified or elicited the content that needs to be conveyed, they must structure and organize it in a manner that will allow effective presentation and retrieval of that information. It is important for designers to plan carefully how to structure and organize the content to be conveyed by the website because poor design of the information architecture leads to poor usability (Nielsen, 2000; iperceptions, 2010), making the website more difficult to use and less enjoyable to the end user. To create a good website, designers should be familiar with the following concepts (e.g., Chou, 2002):

- *Organizational schemes*: how the information is organized (e.g., by name, date, type, association).
- *Organizational structure*: how the website is structured (e.g., hierarchical structure) and the programming logic that is used to access, retrieve, and present different pieces of information.
- *Labeling*: how specific items are referred to. Labels should be used consistently throughout the website.
- *Navigation*: how to set paths that allow users to find their way through the website.

### 2.4.1   Organizational Schemes

One simple, yet successful method that can be used to determine relationships between fixed sets of categories is concept, or card, sorting. In a concept-sorting task, users are given cards that contain concept words and are asked to organize these words into discrete categories based on the relationships between the words. Users are often also asked to come up with a global label for each pile, or category, and then the categories can be used to determine how the various concepts should be organized (see, e.g., Vaughan et al., 2001). Tullis and Albert (2008) describe how a matrix of perceived distances can be created to capture the relation between pairs of cards. Pairs of cards grouped into the same pile would receive a value of 1 and those not grouped together would receive a value of 0 in the matrix. After the matrix has been completed, then statistical analyses such as hierarchical cluster analysis or multidimensional scaling can be applied to extract measures of similarity between the items on the cards. There are many other sophisticated methods for organizing and structuring content for the Web (see Proctor et al., 2002b), some of which are described briefly below.

***Objects/Actions Interface Model***   (Shneiderman, 1997) This model focuses on decomposing complex information found in websites into manageable hierarchies of objects (e.g., networks) and actions (e.g., searching). Specifically, the designers should focus on how the objects and actions are represented in the task and interface. For example, an e-commerce site may consist of individual objects such as pencils and paper. These objects can be aggregated into classes such as school supplies or stationery. Similarly, users can perform individual actions in the site, such as clicking on

a hyperlink to get from one page to another, or aggregate actions, such as scanning a list of different types of office supplies to find and link to a specific category of supplies. To link the objects and actions into the Web interface, designers can use common metaphors (e.g., a file cabinet) to organize objects and "handles" (e.g., pull-down menu) or a magnifying glass to represent an action (zooming feature) or type of action that can be performed. Hadar and Leron (2008) noted that although object-oriented program is useful, it can be for designers to learn how to use object-orienting programming and put it into practice due to difficulties in translating intuition into design parameters.

***Ecological Interface Design*** (Vicente and Rasmussen, 1992) This design process emphasizes that the way in which information is represented in a good display depends on the users' knowledge level and the type of behavior in which the users' engage. The ecological interface design is built on two concepts from cognitive engineering: the abstraction hierarchy and the skills–rules–knowledge behavior framework. The abstraction hierarchy is a multilevel knowledge representation framework that can be used to identify the information content and structure of the Web interface. The skills–rules–knowledge framework is used to distinguish the three modes of behaviors of users (Rasmussen, 1986). *Skill-based behavior* arises when users interact with a system on a regular basis and routine commands can be performed "automatically." *Rule-based behavior* occurs when users are confronted with a novel situation but can apply rules to solve them that they have learned previously. *Knowledge-based behavior* arises when a completely unfamiliar event occurs and the user must invoke his or her problem-solving skills to continue the task. By understanding the behavioral mode and constraints placed on the end user, the goal of ecological interface design is to organize and structure the information in the most meaningful manner to display to the user.

***Latent Semantic Analysis*** (Landauer et al., 2007) This analysis provides a valuable tool for organizing and structuring conceptual knowledge based on the relatedness of the items. Latent semantic analysis considers the frequencies with which words occur in various contexts. Each word then occupies a position in the semantic space based on a large corpus of text and is linked with other words through common labels or features. Words that are more highly associated are represented closer in space. Latent semantic analysis can be used as a tool to structure and organize information because it can easily determine similarity relationships between words, sentences, or larger text units. For example, Katsanos et al. (2008) used latent semantic analysis and other clustering algorithms to develop an AutoCardSorter tool for automating the data collection and analysis of card-sorting tasks.

***Extensible Markup Language*** XML is a markup language that provides designers with the option of tagging the content with standard markup indicators (e.g., </TITLE>) as well as new indicators that may

not have been used before. XML provides a mechanism to impose standards or constraints on how the content is specified so that it can be stored, retrieved, and organized easily.

***Use of the Semantic Web*** The Semantic Web (http://www.w3.org/2001/sw/) provides a framework to transform documents written in natural language into machine-readable form. The machine-readable annotations can be used to for organizing and retrieving Web content. There are several tools available for creating Semantic Web markups (see Michaelis et al., 2011).

### 2.4.2 Labeling

Labels are used extensively on the Web to represent particular pieces of information or categories of information. Labels are usually keywords or short descriptors that highlight the type of content that the user will encounter when accessing information categorized by the label. Category labels can be assigned manually by the designer or can be generated automatically or by computer programs. Qin (2004) recommends consulting existing encyclopedias and reference books in the domain in which the website is being created so that the designer can be familiar with the classification schemes already being used. Designers can also use key terms or keywords from indexes of books in the domain to familiarize themselves with the vocabulary and ontology with which the end users are also familiar.

Labels are used to identify different components of the website, including:

1. *Page Titles*. Every page of a website has a title. Coming up with a good label and description for the page is critical because "you get 40 to 60 characters to explain what people will find on your page. Unless the title makes it absolutely clear what the page is about, users will never open it" (Nielsen, 2000, p. 123).

2. *Headings*. Heading labels are used to signal to the user what content is forthcoming as he or she scans or scrolls down a Web page. Because many users tend to skim a Web page when looking for information rather than reading the content carefully, a good heading will alert the user to pay attention to the forthcoming material. To help make headings as easily to skim as possible, it is important to place keywords to the front of the heading when possible. Headings should also be easily distinguishable from the rest of the text on the screen (see http://www.usability.gov/pdfs/chapter9.pdf).

3. *Hyperlinks*. Hyperlinks take users to other places inside or outside the website. Hyperlinks have brief labels that inform users about the content that they will access by clicking on the link (e.g., if the link is to a file) or information about where the link will take them (e.g., if the link is taking users to another page). Hyperlinks should be made self-explanatory so that users do not have to guess where they will be going by clicking on them (Pearrow, 2000).

4. *Images*. There are many images that can be displayed on a website. When using images, ALT (alternative) tags should be used so that a textual label or description of the image can be displayed to the user when the image cannot be seen (e.g., text-based browser). Von Ahn and Dabbish (2004) introduced a method of labeling images with a computer game called ESP. The ESP game has a pair of users' guess labels for images until a match is obtained between the pair. They found that, on average, 3.89 labels are agreed upon each minute by participants in the game. The quality of these labels was validated by independent tests of search precision and comparison with descriptors generated by participants in an experiment. Coming up with good labels for the images can help produce more efficient image search as well as organization of the images.

5. *Icons*. Icons are symbolic representations of items, actions, or concepts. A good icon conveys an appropriate label for the content. In addition to visual icons, earcons (or auditory icons) are also being used on the Web (see Altinsoy and Hempel, 2011). All icons should include a descriptive label so that they can be indexed and referred to. The label should also be displayed to the users when a cursor is placed over the icon in case the user has trouble identifying the meaning of the icon.

It is important that the same item and/or category within a website be given the same label consistently throughout the website. It may confuse users, for example, to refer to the search feature as "search" in one part of the website and "find" in another part of the site. The label should also be representative of the type of information that it is linked to and should be a term that is recognized by the end users. By producing good labels for the individual components of the website, the designer should have an easier time organizing and structuring the content. Kantor (2003b) suggests the following guidelines for coming up with good labels:

- Never assume that users are familiar with the acronyms, programs, and jargon that the company or designer may use or that they are familiar with how the site is structured.
- Be consistent.
- Use labels that are familiar to users.

### 2.4.3  Organizational Structure

To design an information architecture for a website successfully, the designer must take into account the goals and purpose of the website, the nature of the information that needs to be conveyed, and the type of information that is needed to perform a particular task. The organizational structure refers to the physical structuring of information based on its relationship with other components of the site. The most common types of organizational structures are hierarchical, network, linear, and database oriented (Kantor, 2003a).

*Hierarchical Structures*   Most websites are organized using some type of hierarchical structure. Basically, this structure consists of a high-level, or global, category and is branched into subcategories underneath it (see Figure 1a). The subcategories can be broken down into smaller components until the elemental objects are reached. Sometimes, hierarchical structures are represented as tree diagrams. Shneiderman's (1997) objects–action interface model is based on a hierarchical organization. Hierarchical organizations are good because they represent how users naturally link concepts and categories together. Kantor (2003a) notes that although categories within the hierarchy should be both inclusive (provide all the relevant data) and exclusive (fit into only one category), this is rarely the case in practice because information can be categorized in a variety of ways. When designing hierarchies, the designer must also take into account the issue of breadth (how much should be included in each level) and depth (how deep the levels should go). Tullis et al. (2011) concluded from a review of several studies examining the breadth-versus-depth issue that breadth wins over depth in complex or ambiguous situations and in simple situations depth wins over breadth.

*Linear Structures*   Linear information structures connect pieces of information serially and do not allow for branching of information as hierarchical structures do (see Figure 1c). In essence, linear structures have depth and no breadth. A linear structure is most appropriate for organizing and structuring information that should be accessed in order (e.g., step-by-step instructions or alphabetical listing of items). Linear structures can be embedded into a larger hierarchical structure. For example, the index of an online book can be structured hierarchically according to the first letter (A–Z) of the keywords or concepts, but the concepts themselves can be organized linearly or alphabetically under the first letters.

*Networked Structures*   The Web itself is organized using networked structures. With this organization, users can move around the information architect by clicking on *hyperlinks* (see Figure 1b). The hyperlink can transport the user to an area of the website that does not have to be adjacent to their starting point. This structure is useful for organizing information that is related to but does not necessarily fall into the same category. For example, when searching for specific items such as dishes or plates on an e-commerce site, the results returned could include links to other kitchenware, such as cutlery sets.

*Database Structures*   Information within a website can also be organized and stored in a database structure (see Figure 1d). Database structures usually rely heavily on a search feature that allows users to retrieve the information. For example, the catalog website for a library may use a database structure in which users search for information within well-defined categories (e.g., call number, author, title, and date).

The type of structure used to design a website should reflect the purpose and context of the site as well as
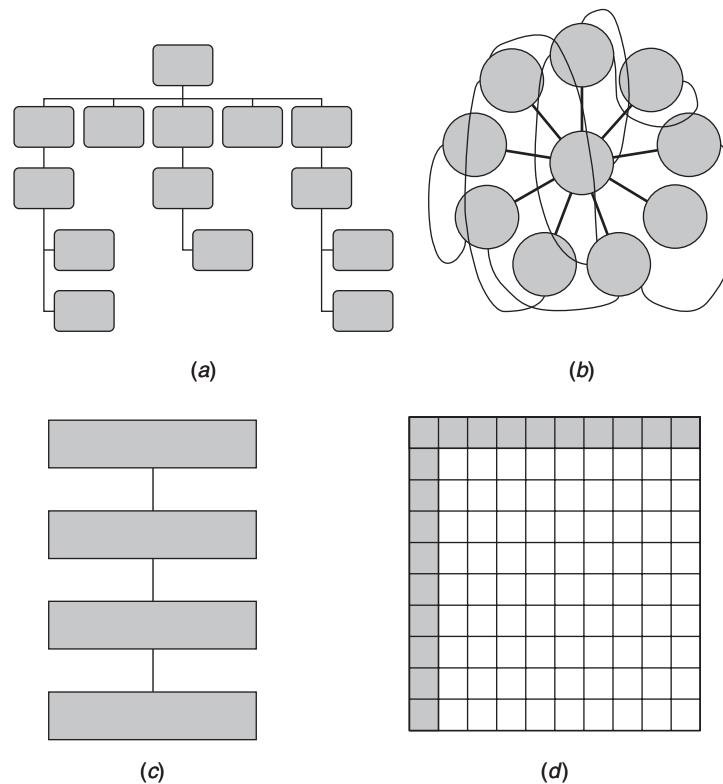
**Figure 1**   (*a*) Hierarchical, (*b*) networked, (*c*) linear, and (*d*) database structures.

the type of tasks the site is designed to support. It is also important to note that websites do not have to be organized solely in terms of one type of structure. Hybrid structures can be developed to support the specific needs of the users or for specific tasks that are performed within the website.

### 2.4.4   Navigation

Once the structure of the website is determined, designers must focus on the navigation system that allows users to access the content from different areas of the site. Navigation through a website is usually implemented through the use of hyperlinks that connect one Web page to another or a search engine that returns results that are relevant to the keyword or terms that a user enters. Because issues relating to information search are discussed in the following section, the primary emphasis of this section will be on navigation. A navigation scheme provides users with a set of paths that allow them to move through a website, such as from the homepage to a specific section of the site.

***Avoiding Getting Users Lost***   Because a website can contain a huge amount of information that is distributed throughout the site, the content needs to be structured in a simple manner that is intuitive to the users. If users become overwhelmed by the information or if the site is structured poorly, users can get lost in

a website. Once the user is lost, he or she must either start over from the beginning or abandon the task. The former adds to the time needed to complete the task, and the latter results in failure of the task.

One way to help users navigate through a website is to communicate the structure of the site to them (Proctor et al., 2002b). If users understand the nature of how the site is structured and organized, they may be able to navigate through it more easily and quickly. Some techniques that can help communicate the structure of the site to users include:

1. *Site Maps*. An outline of the website illustrates how the various sections and individual pages are organized within the site and where they are located.
2. *Interactive Navigation Displays*. Interactive navigation displays provide users with a graphical display of their current position within a website and provide information about how the users got there, how they can get back to a previously visited page, and where they can go next. For example, navigational breadcrumbs are trails of hyperlinked page titles usually located at the top of each page that shows the users how they arrived at the current page. Users can click on the hyperlinks of pages visited previously to move back.

3. *Obvious and Consistent Major Navigation Controls*. Navigation controls for the major sections should be placed in prominent locations on each page of the site. The navigational controls should also be located in the same place on each page so the user can know where to find them (Najjar, 2001). Usually, navigation controls are placed on the top or left side of the screen (Pearrow, 2000). If the page links to lower levels, navigational controls to access those levels should also be placed on the page.

4. *Avoidance of Orphan Pages*. Make sure that users are not stranded when they click on a link that transports them to a page that does not link back to the site. If the link is to an external Web page, warn the user that he or she is leaving the site and/or have the linked page open in a separate window.

5. *Mark for Links Visited*. Separate pages that users have visited from ones that they have not by using different colors for the links. Most websites tend to use the color blue to indicate active hyperlinks that have not been visited and purple or burgundy to mark hyperlinks that lead to visited pages. Halverson and Hornof (2004) showed that differentiating visited and unvisited hyperlinks by color can improve search performance by narrowing the search space.

6. *Back Button and History*. Users use the back button to leave a page approximately two-thirds of the time (see Dix and Shabir, 2011), especially as a means to correct mistakes, to avoid being stuck at a "dead end," and to explore or browse the site. Because the back button is used often, users should be allowed the option of returning to a site visited previously by clicking on the back button. The history feature provided by the Web browser is also a means for users to return to a site visited previously. Because history logs can be set to record the navigation paths for long periods of time (i.e., days and months), it is important to give informative page titles so that users can find the page again when scanning the history log.

**Presenting Navigation Options**    Tullis et al. (2011) provided a summary of several different techniques that are being used for presenting navigation options, including:

- *List of Static Links*. A list of links leads to subsections on sequential pages.
- *Index or Table of Contents Layout*. Links are organized by topics in tables or by columns and rows.
- *Expanding and Contracting Outlines*. Links can expand when clicked to reveal subsections or contract to hide them.
- *Pull-Down Menus*. Heading links result in a pull-down menu of subsections either beneath the heading or to its side.

Tullis et al. concluded that users' performance was better when the navigation controls were presented with simple static listing of links or index/table of contents organization rather than with dynamic techniques.

### 2.4.5   Summary

Many methods can be used to connect or link various parts of a site so that it becomes a unified whole. As aforementioned, the most common technique is the use of hyperlinks. When a hyperlink is activated by clicking on it, the user is moved automatically to a different Web page. Although users can move throughout the site serially by relying on hyperlinks on successive pages, they can take shortcuts or jump from one section of the website to another by using the search function of the site and accessing the links returned.

## 3   PRESENTATION OF AND ACCESS TO INFORMATION

### 3.1   Retrieval of Information

The search function is one of the most important elements of a home page and should be placed in a location that allows users to find it easily. Why is the search feature so important? The answer is simple. Users like to use the search feature to locate information that is of interest to them. According to a survey by the Pew Internet & American Life Project (2009a), 88% of American Internet users find information by using a search engine, and 50% do so daily (2009b). In fact, the Pew Internet & American Life Project survey (2009a) found that searching for information is one of the most common online activities carried out by Internet users, second only to sending/reading email (89%). A summary of statistics by SearchEngineWatch.com (2009) indicated that in August 2009 the three major search engines were Google (with 6,986,580 searches), Yahoo! (with 1,726,060 searches), and MSN/WindowsLive/Bing (with 1,156,415 searches).

Although there has been much research about search engine designs for the Web [see Kammerer and Gerjets (2011) for a review], the focus in this chapter is on local search engines designed for searching within a particular website. One benefit of designing a search function for a particular website is that the content that has to be located and retrieved is much smaller than that on the entire World Wide Web. However, many issues and techniques involved in designing global search engines for searching the Web are applicable to designing search features for an individual website.

Users can engage in two types of behavior when searching for information: browsing and keyword search (Chen et al., 1997). *Browsing* refers to the act of scanning, reviewing, or skimming the contents of a Web page to find interesting or relevant information. *Keyword searching* refers to the process of entering a keyword, term, or phrase into a search engine in an attempt to find a particular piece of information. Of course, these two searching behaviors are not mutually exclusive. For example, even if users engage in keyword searching,

they are likely to browse the returned results page to find a particular link to which they want to go.

### 3.1.1 Browsing for Information

Browsing is a strategy used by beginners when they do not know exactly what they are looking for. However, even more experienced Web users tend to engage in browsing behavior when they are exploring or investigating a topic. Because there is much information that a website can contain, browsing may not be a very efficient way to look for information in the website. Moreover, if a website is large, users can get sidetracked by the various links that they encounter even when they have a purpose for their search. Fang et al. (2005) summarized several types of browsing behaviors that reflect goal-directed or nondirected search behavior:

1. *Search-Oriented Browsing*. Directed search aimed at accomplishing a specific goal. *Example:* looking for a specific section or link in an e-commerce website that contains information about laptop computers.

2. *Reviewing or General-Purpose Browsing*. Scanning through and reviewing information or Web pages related to the users' general goals but not necessarily needed to accomplish a specific task. *Example:* browsing through the electronics section of an e-commerce website to determine what different types of computer products are available and reading information about them.

3. *Scanning Browsing*. Scanning through information without reviewing it. *Example:* scanning the headings on the home page of a website to find interesting topics.

4. *Serendipitous Browsing*. Just looking to see what is in the website, without a specific goal but with the possibility that the user may stumble into something of interest.

### 3.1.2 Keyword Search

Keyword search reflects more goal-directed behavior. That is, users are looking for a particular piece of information at the website that is relevant to attaining their goals. Keyword search is conducted through a website's search engine, in which users enter a query and results matching the query or relevant to it are returned. Keyword search gives users direct access to Web pages within the site without navigating through it serially. Kammerer and Gerjects (2011) proposed a framework for characterizing text searches on the Web that includes four stages: formulation, action, review of results, and refinement. In the first stage, users determine the information for which they want to search. At this stage the information can be a topic such as "Web design," a product such as the "ipad," or even a question such as "What is a Web mashup?" After formulating the search term, users enter the action stage where they click on the "search" button. After the search results are displayed, users enter the review phase in which they evaluate whether the displayed results are relevant and decide which result or results to further review (i.e.,

which website to visit, documents to open, etc.) If users do not find what they need or decide that too many results are displayed, then the refinement stage begins where alternative search terms are considered in order to narrow the results. Keyword search can be more efficient than browsing if the search engine is powerful (i.e., does not return many irrelevant results) because if multiple refinements have to be made, the search may become too time consuming. In the following sections, users' keyword search behaviors are reviewed.

***Finding the Search Function***   For users to use the search function of the site, they must be able to find it. Thus, the search feature should be represented in a manner that will be recognized by users and located in a place where users expect to find the search feature. The former point is captured in Nielsen et al.'s (2000) study, in which "users told us that when they looked for the search function, they looked for 'one of the little boxes.' Tabs and links to a separate Search page just didn't work for them" (p. 7). Search features are often located at the top or bottom of a Web page and should be placed on every Web page so that users can have the option of performing a search instead of browsing through the contents or returning to the home page.

***Simple versus Advanced Search***   *Simple search* refers to search engines that primarily use keywords to find information, and *advanced search* engines allow additional filters. Many advanced search features include Boolean operators (OR and AND). Eastman and Jansen (2003) noted that 90% of Web searches used extremely simple queries and only 10% used advanced query options. They conducted a study examining the effectiveness of searches with and without advanced operators on global search engines and found that the use of advanced operators did not result in significantly better results. Furthermore, Nielsen (2000) reported that users have trouble with Boolean operators because they often confuse AND with OR and vice versa. Given that a majority of users have trouble with advanced search, the website's default search feature should be a simple search, providing an option for advanced search for users who want to use the feature.

***Fast and Accurate Search***   A good search feature or engine should be fast and accurate. Speed refers to how long a user has to wait before the results returned are displayed. Users tend to rate speed as an important factor affecting their preference for a website. For example, users in Lightner's (2003) survey on preferences for e-commerce sites indicated that navigation speed and buying speed were the third and fourth characteristics that determined their overall satisfaction. Thus, the search feature for a website should provide fast, high-quality results. The speed at which a search can be performed depends on the size of the search space and the power of the search engine, and the quality of the information retrieved depends on the precision of the engine (see, e.g., Kobayashi and Takeda, 2000). *Precision* can be defined simply as the proportion of the number of relevant documents retrieved by the search over the number of relevant documents in the search space. The use of

wide search boxes can improve the quality of the search because it allows more room for users to type more descriptive words that will help result in better matches (Nielsen, 2000).

***Dictionaries and Thesauruses*** Typographical errors (typos) occur when users enter keywords or terms into a search engine. Dictionaries can be used to help correct typos. When dictionaries are used to correct a user's spelling, users should be notified of it (Proctor et al., 2002b). For example, before results are returned, a message can be presented indicating that there were "no matches for one-bedroom apartments, did you mean one-bedroom apartments instead?" By notifying the user that the search engine noted a potential typo and corrected it, users will understand why they are receiving the results that are being displayed. Different users may also use different words to describe the same item, so including a thesaurus can help users find items that are stored in the website even if it does not exactly match the label that was assigned to the item.

***Better Indexing of Terms*** Queried terms are usually matched against index terms in order to find matches. Inclusion of a thesaurus will help avoid the problem of different people using different terms to describe the same item. Metatags can also be used to index Web pages within a site. Metatags allow designers to specify additional keywords for a Web page that is indexed. It is very labor intensive to create good metatags for a Web page, but the investment in creating them may be worthwhile because global search engines can use metatags to lead users to the Web page. However, not all search engines support metatags (Sullivan and Sherman, 2001).

***Using Relevance to Rank the Presentation of Results*** Search engines for a site usually return a list of Web pages that "match" the queried topic. Usually, short descriptions of the Web page are provided underneath or next to the hyperlink that will point the user to the page itself. Depending on the nature of the website, many Web pages can match the queried topic. Because users do not typically read or skim all the results that are returned, the linked Web pages should be ordered according to their judged relevance to the queried topic to help users locate the desired information quickly. There are many tools and algorithms that have been developed for determining relevance (see Kobayashi and Takeda, 2000; Fang et al., 2005).

***Allowing Users to Return to the Results Page*** When a user follows a link to a Web page listed on the results page, it may or may not be the particular Web page for which the user is looking. Thus, users should always be allowed to link back to the results page or navigate backward (use the back button of the browser) to return to the results page. When users return to the results page, links visited should be made distinct from new links to show the user where they have been and new Web pages that they have not visited. As noted in Section 2.4.4, blue links should be used to designate unvisited sites and purple or burgundy links to mark sites that have been visited.

### 3.1.3 Summary

The search feature of a website is important because users are prone to use search features to locate information. If users cannot find what they are looking for, they may abandon a website. The search feature should be in the form of a "box" and should include algorithms that return relevant results quickly.

## 3.2 Information Presentation

Successful structuring and organization of the components in a website will lead to more efficient navigation and search of information. The next step is to determine the best methods to present to users the information included in a website. Effective presentation of information is critical because it allows the user to extract the relevant information that they need to accomplish their goals. Fogg et al. (2003) conducted a survey in which 2684 participants evaluated the credibility of websites. They found that the "look" of the website was more important than any other factor in determining the credibility of the site. In fact, 46.1% of the comments received were devoted to issues of presentation and design, including the visual design, layout, color schemes, and so on. Because designers of the website built it, they have a good understanding of where everything is located and know why the information is presented in the manner that it is. However, some designers may fail to recognize that just because the organization and presentation of information is intuitive to them does not mean that it will be intuitive to the end users. Subsequent sections are devoted to the topic of information presentation at the global level and page level and how to present information in a manner that is accessible to users or different cultures and to users with disabilities.

### 3.2.1 Global Site Design

***Simplicity, Straightforwardness, and Easiness*** Information presented in a website should be made simple and straightforward (Nielsen, 2000). This will help the users locate the information they are looking for and will aid in the ease with which they can use the website. Although fancy designs may be aesthetically pleasing, they may distract users or "hide" relevant features or information presented in the site. In the spirit of parsimony, the simplest design that includes all the relevant and important features and information is probably the best.

***Browsers and Other User Agents*** It is important to know what types of browsers users will be using to access the site and to present information in a manner that meets the requirements of the browser. If the information is presented in a manner not supported by the browser (e.g., frames and flash), the user will not be able to see it. The most common browsers should be taken into account when designing the presentation of information. According to the Web analytics firm Net Market Share (2010a), the most commonly used browsers are Internet Explorer, Firefox, Chrome, and Safari. However, browser usage trends are constantly changing and designers should consult the most current data available whenever

they are working on a website. Designers also need to note that the same website may look different when accessed by different browsers (see, e.g., Figure 2). Furthermore, because newer versions of these browsers are released periodically, designers should present the information to accommodate the requirements not only of current versions but of back versions as well. Nielsen (2000) analyzed statistics of browser versions and noted that only about 2% of users per week upgrade or change to a newer version of a browser when moving from version 1 to 2 or 2 to 3 and only 1% from version 3 to 4. Thus, it seems appropriate to keep one year or more behind the latest browser. At the very least, it is important to design for the most used versions of the most popular browser: Internet Explorer. Internet Explorer versions 6.0, 7.0, and 8.0 are each used more often than all other browsers except for Firefox (Net Market Share, 2010b), and each has its own quirks in how it renders Web pages.

It is also important to keep in mind other user agents that users will use to connect to the Web. Users are no longer limited to accessing the Web from their desktops and laptops but can access the Web from many devices, including cell phones, personal data assistants, and ipads. Furthermore, accessing the Internet via mobile devices is becoming more prevalent. A survey by the Pew Internet & American Life Project found that 32% of Americans have at one time used a mobile device to access the Internet (a substantial increase from 24% in 2007), and 19% use the Internet on a mobile device on a daily basis (up from 11% in 2007; Horrigan, 2009). Thus, when designing the format of the Web page, designers should take into account mobile devices in addition to traditional computer monitors. Major factors to consider when designing Web pages to be accessed on mobile devices are that the display is only a small fraction of what is available on computer monitors (which have gotten bigger over the years due to decreases in cost) and the small or special input mechanisms (keypad or touchscreen) text entry, navigation, and information selection (Xu and Fang, 2011).

**Scrolling and Paging** Web designers often struggle with how much information should be presented on a given page of a website. Information that is presented in large amounts becomes lengthy and takes up a lot of space, and users are likely to have to scroll down the page to read or review the entire content of the page. Users do not like to scroll, so some designers recommend placing important or critical information "above the fold," within the space that can be viewed without scrolling (e.g., Pearrow, 2000). To avoid having to scroll, information can be decomposed into smaller chunks and presented on different pages. Users can access the information on the next page by clicking on a link or forward button. This method does not require users to scroll but does require them to page through the information serially (although some sites allow users to "jump" pages by providing links to numbered pages that represent the chronology of the information).

Baker (2003) found that it took participants 19 s longer to read paged text than scrolled text, but there was no significant difference in the level of comprehension between the two groups. He also found that users took 55 s longer to search for information in paged text than in scrolled text. Thus, it may be better to place all the information on a single page and require users to scroll than to break up the information across different pages. However, Tullis et al. (2011) noted that home pages and navigation pages should be kept short. One reason why these two types of pages should be kept short is that they have the goal of pointing users to the information that they need rather than being the source of that information.

The scrolling referred to above involves vertical scrolling or moving down a page by scrolling. However, if the website is designed to be too wide, users may have to scroll horizontally to see the information at the "edges." In general, it is recommended that horizontal scrolling should be avoided unless the task involves item comparison, in which case horizontal scrolling may facilitate the task (see Najjar, 2001).

**Secondary and Pop-Up Windows** *Secondary windows* are windows that open up with the Web page or information that users request, whereas *pop-up windows* are usually Web pages, advertisements, or information that "pops up" or appears without the users' request. Pop-up windows are usually considered annoying by users and should not be implemented in a website. Moreover, Fogg et al. (2003) indicated that "pop-up ads were widely disliked and typically reduced perceptions of site credibility" (p. 7). Secondary windows, on the other hand, can be used appropriately to present information since users are requesting the information and the information is not being imposed on them. Secondary windows are especially helpful for presenting additional information such as online help or detailed product information (see Tullis et al., 2011). When a second window is being used to present additional information, users should be notified of its use by presenting the secondary window in a smaller size that does not occlude the original screen or by prompting the user that another window will be opened to display the requested information.

**Frames** Frames divide the browser screen into distinct parts, each of which can be navigated and scrolled through independently. Frames can be particularly useful for displaying information that needs to be visible while the user scrolls down a page (see Dix and Shabir, 2011). For example, providing navigational tools in a separate frame allow users to access them even when the user scrolls "below the fold" of the content material in another frame (see Figure 3). However, frames have drawbacks. First, not all browsers support the use of frames. Frames can also make the website or pages from the website harder to find by search engines because the page that sets the frame is indexed instead of the content page (see Wisman, 2004). Finally, there can be problems with bookmarking and printing content from websites with frames because the page that sets the frame is the one that is often marked or printed or the last framed accessed is printed.
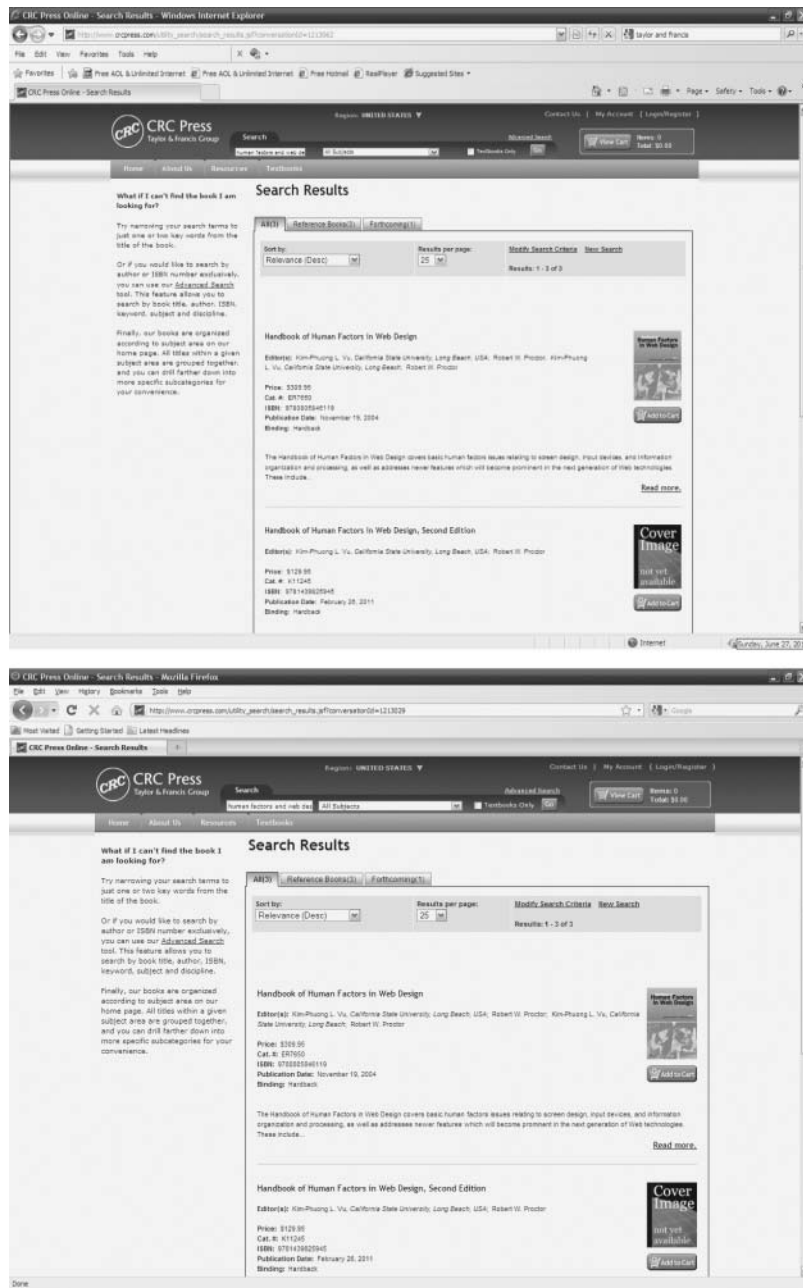
(*a*)

**Figure 2** (*a*) The same website displayed consistently in Internet Explorer 8 (top) and FireFox 3.5.9 (bottom). (*b*) The same website displayed inconsistently in Internet Explorer 8 (top) and FireFox 3.5.9 (bottom).

(*b*)

**Figure 2** (*continued*)

***URL Design*** URLs are Web addresses that direct the users to the desired Web page. The URL of a website starts with the prefix "http://"; however, some Web servers will load the page without the prefix. After the prefix are the letters "www" for World Wide Web, followed by a domain name and specific page titles, if there are any.

Nielsen (2000) noted that the domain name is the most important component of the URL because if users can recall or guess the correct domain name, they can get to the home page and find their way to specific information from there. Because URLs must be specified completely accurately, Nielsen recommends that they be made as
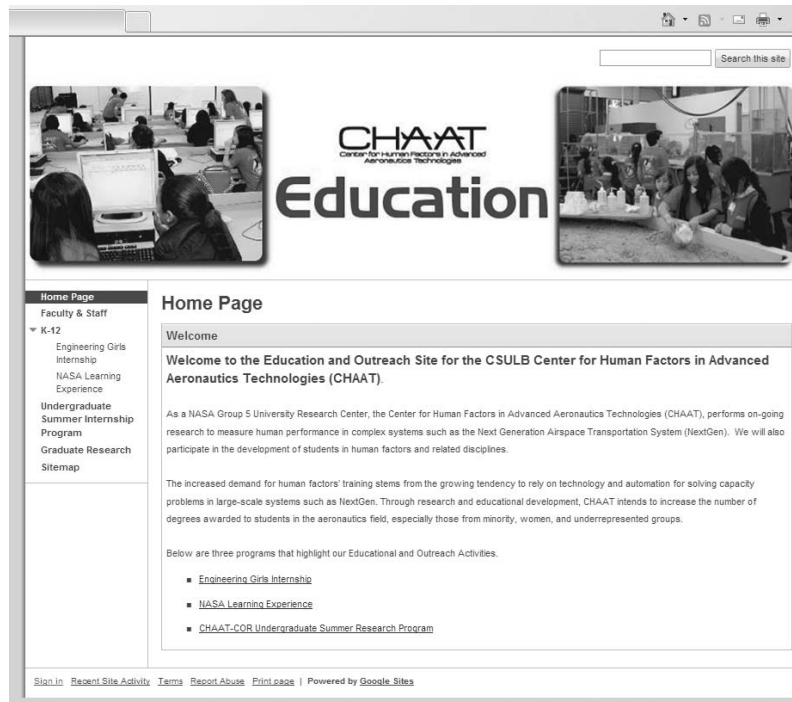
**Figure 3**  Example of a website that uses frames for navigational controls.

short as possible, with common words, all lowercase letters, and without any special characters. Because some users will choose to "cut and paste" URLs into browsers (e.g., from a referral in an e-mail) or link to a specific URL, it is important that the URL not be broken (e.g., by inserting a return break at the end to "wrap" the URL). As mentioned earlier, one problem with a frame is that it breaks the URL, making it difficult or impossible for users to bookmark the specific Web page.

*Error Messages*    There are many possible reasons for errors to occur. Some of these errors may be due to the Web browser that the user is using whereas others can be due to the design of a specific website. Lazar and Huang (2002) examined error messages for Web browsers and noted that many error messages are confusing and difficult for users to understand. Although the designer cannot control the error messages produced by the Web browser, care should be taken in error messages returned by the individual website. For example, if the website asks the user to enter the date in a specific format such as MM\DD\YYYY, but the user enters the date as MM\DD\YY, the error message returned should not be something obscure such as "PAGE ERROR: INVALID REFERENCE, NOT SET TO AN INSTANCE OF THE REFERENCE" but should indicate the error so that the user can fix the problem. A better error message would be: "FORMAT FOR DATE IS NOT VALID, PLEASE ENTER DATE IN THE FORM OF MM\DD\YYYY" (see also Figure 4). This message specifically tells the user that the date field is incorrect

rather than making the user guess what is meant by "INVALID REFERENCE," as in the former error message example. Several Web pages have also been designed to mark the field or entry in which the error was made to help users find where the error occurred.

### 3.2.2   Page Design

*Home Page Design*    The home page is usually, but not always, the first page that a user encounters when entering a website. If users access specific pages in the site (e.g., from a hyperlink on a global search engine's results page or from another website), they should be allowed to have direct access to the home page through a link or "home" button. The home page should communicate the site's purpose, include information about the site or organization hosting the site, and contain navigation functions to major subsections of the site. Nielsen and Tahir (2002) examined the usability of 50 websites' home pages and suggested 113 guidelines for designing usable home pages. From these, Nielsen created a list of the top 10 guidelines for home page usability (Table 1).

*Page Layout*    The layout of the page should be designed in a manner consistent with users' expectations. For example, the home page link or logo linking to the home page is usually located at the top-left corner of the page, navigational links are located at the left column of the page, and the main content is located in the center. Guidelines and standards have been developed to try to make Web pages more consistent, and user testing
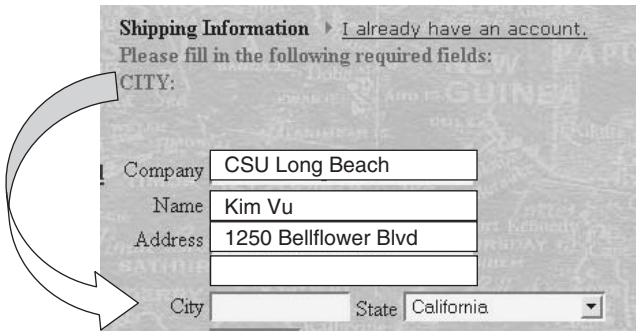
**Figure 4**   Example of a good error message on a website.

**Table 1  Top 10 Guidelines for Home Page Usability**

| Home Page Design Guidelines |
| --- |
| 1.  Include a one-sentence tagline. |
| 2.  Write a window title with good visibility in search engines and bookmark lists. |
| 3.  Group all corporate information in one distinct area. |
| 4.  Emphasize the site's top high-priority tasks. |
| 5.  Include a search input box. |
| 6.  Show examples of real site content. |
| 7.  Begin link names with the most important keyword. |
| 8.  Offer easy access to recent home page features. |
| 9.  Don't overformat critical content, such as navigation areas. |
| 10.  Use meaningful graphics. |

*Source*: Based on Jakob Nielsen's Alertbox, available at http://www.useit.com. Readers should visit useit.com to read more details about the guidelines.

has shown some layouts to be more effective than others. Tullis et al. (2011) conducted a thorough review of research relating to page layout issues and recommended the following based on the literature:

- *Use a fluid layout rather than a fixed layout.* A fixed layout does not change with the size of the browser window, whereas a fluid layout adjusts the elements to the browser window as well as the printed page.
- *Use a medium level of white space.* White space can be used to reduce clutter, although too much white space may increase the time needed for users to locate items if it causes a need to scroll down to see information.
- *Place items in locations where users are likely to expect them.* Because the Web has been around for some time, users have become aware of what a "typical" Web page should look like.

**Links**   Links are important elements of a Web page because users can use the links to go to other Web pages that are connected to the current one. There are two types of links, text-based links and image/icon links.

The mouse cursor changes shape when it is placed over an active text or image link. Text-based links are usually marked by highlighting the text on a website in color and underlining it. Visited and unvisited links should be distinguished clearly to help users know where they have been. Using different colors to mark visited and unvisited links helps reduce search time (Halverson and Hornof, 2004), but care should be taken as to which color to use. Halverson and Hornof showed that search time is increased when red-text links are present on a Web page.

Icon or image-based links are graphics or pictures that link to another page when they are clicked. It is important to make links distinctive from nonlinks so that users know by looking at the image or icon what they can and cannot click on to take them to another Web page, rather than having to place their cursor over each item to find which is an active link. An image link is particularly useful when it provides a visual representation of the information requested or the item or product available from the website. Because text-based links usually download faster than image- or icon-based links since they are smaller in size, thumbnail image links can be used to lead users to an enlarged picture or image to reduce page load times when many images are presented on a page (e.g., a returned search page).

Portals, websites that are intended to provide links to other websites or online resources, often display many links on a single page. Because portals consist of a framework and supporting technologies for integrating multiple systems into a single interface, it has the ability to aggregate content from multiple sources and still provide users with a unifying look and feel to the information (Eisen, 2011). One question that arises is how many links should be displayed on a page. Bernard et al. (2002) had participants locate specific links on a search engine's result page for which the number of links returned was 100, with the number of links present on each page being 10, 50, or 100. Task completion time was shortest for the page with 50 links and longest for the page with 10 links. Users rated the 10- and 50-link-per-page formats roughly equal in terms of how easy it was to find information on the page and in terms of being professional looking but showed a slightly greater preference for the layout of 50 links to a page.

***Text versus Graphics and Images***    Text is the most common method used to convey information. This is especially true since there are some Web browsers that do not support graphics, and skilled users may know how to turn off the graphics mode in browsers that support them. Thus, the question of interest is whether graphics should be used with the text and, if so, what combination of text and graphics is optimal. Sears et al. (2000) found that all participants in their study indicated that websites that contained graphics were more attractive than those that contained only text. Although the use of graphics and animation can distract users, some types of information can be conveyed better by a "picture" than by words (see, e.g., Plaue et al., 2004). Sears et al. conducted an experiment examining a version of the 1997 Microsoft website that was modified to make the site self-contained (i.e., not linked externally) and a simplified version of the site (animations were removed and graphics simplified, graphical links were transformed to text links, etc.). They found that users were able to find desired information more easily with the original site than with the simplified site. Thus, there does not seem to be a problem with including some graphics in a website, as long as ALT tags or labels are used to indicate what the graphics are supposed to be in browsers that do not support them.

Graphics should not be used gratuitously, though, because they tend to be large in size, slowing the download time. As described in the next section, slow download times often frustrate users. The use of graphics and images is important to many e-commerce sites because users often want to see the product. Because search results from an e-commerce site tend to return many products that fit the queried item, including detailed images of each product may not be possible. In this case, small thumbnails of the image should be provided, with the option for users to access a more detailed version of the image if desired.

***Animation***    Animation can be applied to graphics to mimic real-life movement. Animation is frequently used as an attention getter in Web pages (Nielsen, 2000; Lazar, 2003). People are particularly sensitive to movement in the peripheral visual field due to the human sensory and perceptual system (see Chapter 3). Nielsen indicated that the use of animation is good for the purposes of (1) demonstrating continuity, transitions, or changes over time; (2) illustrating three-dimensional structures; and (3) attempting to attract attention. However, it has been shown that having animation present on the screen decreases performance on other tasks (e.g., Zhang, 2000). Zhang had participants perform an identification task in which users had to report the number of times the target information (i.e., strings of letters) appeared in an $8 \times 10$ table of a Web page. On the page containing the search array, animation was presented randomly to the top, sides, or bottom of the table. Zhang found that performance in the identification task decreased in the animation conditions compared to baseline conditions in which no animation was present. Furthermore, the animation particularly hurt performance when it contained information similar to the targeted information that was irrelevant to the task.

Animated banners are typically used as attention getters for advertisements. However, there is at least some evidence suggesting that users do not recall animated banners more than static banners (Bayles, 2002). Furthermore, advertisements, in general, tend to be viewed negatively, especially pop-up ads (Fogg et al., 2003). Thus, designers should take caution when determining whether to include animation. Animation is not always bad, but its presence can hurt performance on other tasks.

***Page Load Times***    The proliferation of high-speed Internet (i.e., broadband) services such as digital subscriber line (DSL) and cable modems during the last decade has drastically reduced the number of households that connect to the Internet via dial-up modems (U.S. Department of Commerce, 2010). The U.S. Department of Commerce's survey of connection types showed that, in 2009, 63.5% (75.8 million) of U.S. households used broadband services (up from 4.4% in 2000), whereas only 4.7% (5.6 million) of U.S. households used dial-up (down from 37.0% in 2000). Though this shift has abated many of the page load time issues for websites accessed from home computers, similar issues are now appearing in the context of mobile devices as more and more users begin to access the Web on-the-go at reduced connection speeds (when compared with broadband). It is important to take the users' connection speed into account because the connection speed may determine some of the users' behaviors. For example, a user may have no qualms about downloading a large media file when he or she is connected to the Web through a cable connection but may hesitate to do so when connected with a mobile device.

Users dislike long Web page load times whether they are operating a traditional computer (e.g., Lightner et al., 1996) or a mobile device (Equation Research on behalf of Gomez, Inc., 2009). There are many reasons why users may react unfavorably to long page load times (e.g., Bouch et al., 2000), including (1) not knowing whether the user made an error, (2) not wanting to wait "forever" to complete a task, (3) thinking that the site is not secure or well designed, and (4) losing his or her train of thought or task set during the delay.

Jacko et al. (2000) conducted an experiment evaluating the effects of network delay on users' perceptions of the quality of Web documents. They manipulated the page load time (mean delay times were 575, 3500, and 6750 ms for the fast, medium, and slow conditions, respectively) of Web pages from a website with only text or one with text and graphics. The content of both websites were the same. Users judged the text and graphic documents to be of higher quality than the text-only documents at short delays, but text-only documents were judged to be of higher quality at long delays. Jacko et al. noted that the lower quality judgments for graphic/text pages with long delays may be due to users attributing the long delay to graphics involved in the document. Sears et al. (2000) also showed that users with slower connections are also less impressed with Web page graphics than those who tend to have access to faster connections. Ten years ago, Nielsen (2000) recommended that Web pages be formatted in a manner

that can load in 10 s or less. Then, in 2004, a study by Galletta et al. suggested that delays should be less than 8 s to promote positive attitudes from users and less than 4 s to encourage users to continue with the task or to revisit the site later. Most recently, a study conducted by Forrester Consulting on behalf of Akamai Technologies, Inc. (2009) found that almost half (47%) of the 1048 online shoppers they surveyed expect to wait no more than 2 s for a Web page to load when they are browsing or searching for a product, and 40% responded that they will actually leave a site if forced to wait 3 s for a page to load. Unfortunately, the average page load times for websites of top shopping, banking, and airline companies are about 4 seconds on a mobile device (Gomez, 2010). Furthermore, a study by Equation Research on behalf of Gomez, Inc. (2009) found that 33% of 1001 mobile device users expect Web pages to load on their mobile device as quickly as (or faster than) on their home computer (another 25% expect the Web page to load almost as quickly on their mobile device as on their home computer). However, in that same study by Equation Research on behalf of Gomez, Inc., only 20% of the users responded that they would wait 5 s (or less) for the page to load before giving up and exiting the page (30% of users reported they would wait for 6–10 s).

Bouch et al. (2000) found that users' tolerance for delays decreased as the time they spent on the task increased, but if the page loaded incrementally, users were more tolerant of longer delays. If it is not possible for a page to download in the recommended time, users should be notified of this in advance. One way to help minimize the download time of Web pages is to reduce graphics to thumbnails and allow users to "zoom" or access larger images separately (perhaps by using an image link).

## 3.3 Designing for Accessibility and Universal Access

One benefit of having a website is that it can be accessed from anywhere and at any time with a variety of computing devices. For a website to be of maximal use, it should be designed proactively to promote accessibility from the early design phases and throughout the development life cycle. Stephanidis and Akoumianakis (2011) consider access to be a contextual issue that is determined by three major parameters: target user, the access terminal or interaction platform, and the task context.

Although a particular website can be viewed by anyone who can arrive at its "address," it is likely that the site is designed to target all not users but rather a subset of users. For example, a website devoted to being a portal to refereed publications in cognitive neuroscience is unlikely to be used by elementary school children. Thus, this site does not have to take into account the cognitive capabilities of elementary school children, but rather, it should be designed to promote easy access of research materials for scientists and medical professionals from all over the world. The website should also be designed to be usable on different computing devices that the target users are likely to use for accessing the information, such as personal computers, laptops, and personal data assistants. Finally,

the design of the website should take into account the context in which the site will be used, such as to browse through the contents to see the latest research in the area or to look up a specific piece of information of interest. In the following sections, Web design issues for different classes of users will be discussed.

### 3.3.1 Cross-Cultural Designs

Cultural differences in the user population relating to time, language, communication style, and social context may affect Web use (Rau et al., 2011). For example, users can be grouped into two general classes in terms of their perception of time: monochronic and polychronic (see, e.g., Rose et al., 2003). Monochronic cultures view time in linear fashion and are very task oriented. Polychronic cultures view time in a more flexible manner in which several tasks can be worked on concurrently and the users switch back and forth between the various tasks. Rose et al. conducted a study with users from two countries whose cultures are primarily monochronic (United States and Finland) and two whose cultures are primarily polychronic (Egypt and Peru). They found that all users had a negative attitude toward delays, but it was less pronounced for the polychronic cultures than for the monochronic cultures.

Web designers should take into account not only cultural issues such as time perception but also the resources to which users are likely to have access. For example, Chung (2008) indicated that between 2000 and 2007 there was tremendous growth in online users from Latin America, the Middle East, and China. Yet, many users from these countries do not have effective means for accessing information on the Web because many search engines and Web portals are designed to primarily serve English-speaking users. Thus, there is a need for Web designers to take into account cross-cultural issues in developing websites such as the representation of meaning through icons, symbols, and colors, the use of graphics, and the differences in page layout when translating images and text from different languages [see Rau et al. (2011) and Chapter 6 for reviews on cross-cultural Web design]. Some of the guidelines summarized by Rau et al. are:

- Use unambiguous language.
- Allow extra space for text.
- Accommodate text reproduction methods.
- Do not embed text in icons.
- Use an appropriate method of sequence and order in lists.
- Take linguistic differences into account.
- Take the direction in which text is read into account.
- Be aware that variations exist within the same language.
- Provide natural layout orientation for information to be scanned.
- Provide layout orientation compatible with the language being presented for menu designs.

- Note that icons designed and tested well in one region may not be accepted by people in other regions.
- Provide a combination of text and picture when designing icons.
- Examine the textual component in the graphics on the Web carefully when it is intended for a global audience.

### 3.3.2   Designing for Older Adults

The population of older adults is large and increasing rapidly. It is estimated that by the year 2050 the number of adults 55 years of age or older will exceed 136 million (U.S. Census Bureau, 2008). Bitterman and Shalev (2004) indicated: "Senior citizens are connected to the Internet for the longest time at a single sitting—more than any other age group" (p. 25). Furthermore, Internet use trends indicate that people who used the Internet when they were younger will continue to use it as they become older (U.S. Department of Commerce, 2010). Thus, websites should be designed to take into account the capabilities of older adults.

Because declines in sensory, cognitive, and motor functioning with age can affect the users' ability to interact with computing devices, it is important to take these considerations into account when designing a website. Below are some design guidelines relating to perceptual and cognitive changes with age that Web designers should take into account [based on recommendations from Bitterman and Shalev (2004) and Fisk et al. (2004)].

Perceptual changes:

- Use larger letters and avoid decorative or cursive fonts (e.g., 12-point sans serif letters).
- Avoid style sheets that prevent users from increasing the font size.
- Use bright colors and avoid combinations of colors of short wavelengths (blue–violet–green).
- Maximize contrast (at least 50:1 contrast, e.g., black text on white background).
- Increase the resolution of elements on the screen.
- Minimize clutter.
- Avoid moving and scrolling text.
- Provide redundant information (e.g., graphic and text or visual and auditory).
- Avoid high-frequency sounds (above 4000 Hz).

Cognitive changes:

- Use a consistent structure for the site and layout of the Web page.
- Convey the system's status clearly.
- Provide clear feedback for errors.
- Minimize demands on working memory.
- Allow for recognition of information, rather than requiring recall.
- Group information.

- Provide a site map that illustrates where the user is.
- Emphasize important and relevant information.
- Provide a "help" section or "support" information.

### 3.3.3   Accessibility for Users with Disabilities

In the United States, approximately 8.5% of the population has at least one of the following types of disability: blind or severe visual impairment, deaf or severe hearing impairment, difficulty walking, difficulty typing, or difficulty leaving home (U.S. Department of Commerce, 2002). Websites should be designed to take this population into account. Designers should not regard having to make the site more accessible to users with disabilities as an extra burden because websites that are designed to improve accessibility for disabled populations often produce benefits to people without disabilities as well (Caldwell and Vanderheiden, 2011). For example, while proper headings are necessary for an accessible design, they also improve usability. Users with vision impairments who use assistive technologies called screen readers are able to jump directly to headings to find content more quickly and easily, and with the headings in place, users without vision impairments are also able to focus on the content they are looking for. Taking headings to the next accessible step, creating an appropriate hierarchical structure with headings will improve usability one step further as well. To create an appropriate hierarchical structure with headings, the level 1 heading (h1; there should be only one level 1 heading per Web page) should be the first heading on a Web page, only level 2 headings (h2) should be directly under the level 1 heading, only level 3 headings (h3) should be directly under level 2 headings, and so on. For example, a possible heading structure is (note: indentation is used for structural demonstration purposes only):

- h1
  - h2
  - h2
    - h3
  - h2
    - h3
    - h3

Providing a proper logical structure such as this is necessary for accessibility because it provides information to a screen reader user about where they are in the flow of information. However, it also assists users without visual impairments because when headings are marked up in this way the same level headings will have the same visual appearance (and different level headings will have a different appearance), and thus the headings will provide a quick outline for the content presented on the screen.

There are many resources to assist the Web designer in creating accessible Web pages. First, there are the Section 508 standards of the Workforce Rehabilitation Act (Access Board, 2000), which federal agencies (and

companies creating products to sell to the government) are required by law to follow in order to make websites accessible. An update of the Section 508 standards is due out in 2010 or 2011. There are also the Web Content Accessibility Guidelines (WCAG) created by the World Wide Web Consortium. There are two versions of the WCAG: 1.0 (Chisholm et al., 2001) and 2.0 (Caldwell et al., 2008). WCAG 2.0 (principle-based) guidelines are often found to be too vague and difficult to implement, so many developers continue to use the earlier version, WCAG 1.0 (technique-based guidelines) for guidance on creating accessible websites. WCAG 1.0 recommends the following to promote the development of Web content that will be accessible to people with disabilities:

1. Provide equivalent alternatives to auditory and visual content.
2. Do not rely on color alone.
3. Use markup and style sheets, and do so properly.
4. Clarify natural language use.
5. Create tables that transform gracefully.
6. Ensure that pages featuring new technologies transform gracefully.
7. Ensure user control of time-sensitive content changes.
8. Ensure direct accessibility of embedded user interfaces.
9. Design for device independence.
10. Use interim solutions.
11. Use W3C technologies and guidelines.
12. Provide context and orientation information.
13. Provide clear navigation mechanisms.
14. Ensure that documents are clear and simple.

Web pages designed following these guidelines (and/or the Section 508 standards) will be more accessible by users with diverse sensory capabilities than if no accessibility guidelines had been consulted but will not be guaranteed usable by people with disabilities. For example, an image used only to create space between two sections on the screen could have an equivalent auditory alternative of "spacer image" for a screen reader to read in order to conform to the first guideline in WCAG 1.0. However, this would make the user experience much more difficult for a person with a visual impairment who navigates the page using only a screen reader because he or she would be exposed to this unnecessary auditory information. The only way to know for sure that users with disabilities can actually succeed in completing the tasks that they have come to a website to complete is by user testing with users with disabilities.

## 3.4 Security and Privacy

The Web allows users to access information, exchange information, and perform transactions online. However, there are well-founded concerns about privacy and

security since personal data collected about users need to be protected. Some users try to avoid websites that ask for personal information because they fear misuse of this information. For example, a user may be hesitant to register his or her e-mail address with a site in order to view its contents because he or she does not want to deal with the possibility of receiving unsolicited emails from the site's organization or one of its affiliates. To date, the most prominent effort for online privacy protection is the W3C's Platform for Privacy Preferences (P3P) project (see http://www.w3.org/P3P/). The goal of the P3P project is to enable websites to encode their data collection and data use practices in a machine-readable XML format and to provide a simple and automated mechanism for users to specify their privacy preferences through a Preference Exchange Language called APPEL. Through the use of P3P and APPEL, a website's privacy policy can be checked against the user's privacy preferences to determine whether the site's data collection and data use practices are acceptable to the user. Although the P3P project is a significant step toward developing privacy protection mechanisms, it is still in development and has its limitations.

In addition to privacy concerns, issues relating to information security have received tremendous attention in recent years because much of our personal information flows through the Web as we receive online services or make online transactions. As a result, it is essential for websites to have reliable security systems that protect the sites against information theft, denial of service, and fraud (see, e.g., Schultz, 2011). The most common method used by websites to identify and authenticate users is the user name–password combination. However, despite its pervasive implementation, it is well known that the user name–password combination is a relatively weak security method because many users fail to adopt crack-resistant passwords (Proctor et al., 2002a). Users also tend to create passwords that are easy to remember (e.g., Riddle et al., 1989; Klein, 1990) or write them down. Furthermore, because different sites have different requirements for acceptable passwords, users have trouble remembering unique passwords for multiple accounts (Vu et al., 2007). One reason why the user name–password method is still popular despite its limitations is that it is relatively easy to implement and is accepted by users (as opposed to more intrusive methods, such as biometrics).

Websites should also be designed to protect against breaches in security such as hacker attacks (unauthorized access to the information stored by the website), denial-of-service attacks (which cause the network to slow down or website to stop working), Web defacements (unauthorized modification of the content of the site), and computer viruses (Schultz, 2011). Techniques to counter security breaches are beyond the scope of this chapter. However, given that the users' trust in the website can be obliterated by security breaches and the potential costs and legal ramifications associated with the breaches can be great, it is important to emphasize that care should be taken when designing the website to incorporate security mechanisms.

## 4  EVALUATING WEB USABILITY

Web usability refers to the ease with which a website can be used, its efficiency and effectiveness, and the satisfaction it provides to users (e.g., Brink et al., 2002). Web usability is extremely important because it is a major determinant for whether or not a particular website will be successful (Nielsen, 2000). Because Web usability is an area within the general domain of computer software usability (see Vu et al., 2011), many issues relevant to software usability discussed in Chapter 46 also apply to Web usability.

Although the topics of Web design and Web evaluation were covered separately in this chapter for ease of presentation, designing and evaluating Web usability often occur in iterative cycles (Mayhew, 2011. Web evaluation should be implemented early in the development process and throughout its life cycle. It is not adequate to conduct usability evaluations on a final product to catch the "bugs." Moreover, usability problems discovered on a final product may not always be fixable. Thus, it is important to emphasize that usability evaluations should be implemented in early, as well as late, phases of design, where usability problems can be identified and more easily fixed.

Web usability evaluation is dependent on the goal or purpose of a website. These goals and purposes were elicited from the organization or users at the start of the design process covered earlier in the chapter and should include consideration of:

1. *Target Users*. Information and descriptions of the target users should include demographics, Web and computer experience, core user tasks, task environment, and preferences. Characteristics of the target users can be obtained by examining user profiles (Fleming, 1998) or personas (Cooper, 1999). *User profiles* are brief summaries of real users' characteristics. *Personas* refer not to real people but to a representation of a type or category of users. Personas are typically used when designers do not have access to user profiles but have a general idea of the characteristics that actual users might have and can be based on the designers' knowledge of who the real users are likely to be.

2. *Core User Tasks*. Core user tasks are those tasks that are frequently performed by the target users. Core user tasks are highly dependent on the nature of the website. Below are examples of some likely core user tasks for each of the major types of websites:

   - *News/Information Dissemination*. Find out the late-breaking news; check the weather or stock market.
   - *Portal*. Find websites relating to Web design; find electronic resources on the topic of usability.
   - *Communication*. Chat with other users with a common interest in e-business; e-mail a friend.

   - *Search*. Find information on a topic such as cancer; find a professor's e-mail contact information in a university's online directory.
   - *E-Commerce*. Purchase Jamaican Blue Mountain coffee; purchase a computer online.
   - *Company/Product Information*. Find the location of a store; track delivery status of a product.
   - *Entertainment*. Watch a movie trailer; play a game online.

The core set of user tasks can be defined at different levels. For example, higher level tasks can include browsing, researching, communicating, searching, purchasing, and accessing entertainment. Each of the higher level tasks can be broken down into subtasks such as locating a product on a Web page, "placing" it into the "shopping cart," providing billing information, and providing shipping information. Once the goals of the website are defined and the core user tasks are established, one of several usability methods can be used to evaluate whether the website is successful in conveying the purpose of the organization and supporting the users in accomplishing their tasks.

### 4.1  Evaluation Methods

There are several general classes of usability tests, although some of the specific methods under each can cut across categories (see Tullis and Albert, 2008):

- Interviews, focus groups, and surveys/ questionnaires
- Naturalistic observation
- Participatory evaluation (ethnographic methods and diary studies)
- Web-based methods (automated sessions, Web logs, and opinion polls)
- Prototyping (paper prototypes and interactive prototypes)
- Usability inspections (heuristic evaluation, cognitive walkthrough, and alternative viewing tools)
- Usability lab testing (usability and performance testing)

The first four classes of usability tests were introduced earlier as methods to help elicit information about users and core user tasks when determining what content to include in the website. However, they are also used as techniques for evaluating usability later in the design process. For example, interviews, focus groups, and surveys/questionnaires can be used to determine the difficulties that users have while interacting with the website and their opinions, attitudes, satisfaction, and preferences for the different features, layout, or structure of the website. Similarly, naturalistic observation can be used to evaluate users' performance on different tasks and facial or bodily expressions while interacting

with a website. Participatory evaluations can be used to evaluate the website in the context of the users' social and task environment or by having users' log usability problems and concerns as they encounter them in daily use of the website.

As mentioned earlier, the Web itself is a tool for performing evaluations on a particular website. When a site is accessed, the Web server records information about the files sent to a browser, which can include the date, time, host domain (the Web address of the requesting browser), file name, referrer (the URL of the page that provided the link), and so on. Not only can Web logs reveal information about the users, such as how many users visit a website and from which domain they are visiting, but Web logs can also be used to evaluate the site by examining the frequency with which specific Web pages in the site are visited, the amount of time spent on a page or task, the success versus bail-out rate, and navigational paths that users take while interacting with the site. Web log data may also provide some clues for usability problems. For example, if a specific Web page is not visited frequently, designers can conduct further investigation to determine whether the lack of access to the page is due to where it is placed within the site or the content that it conveys.

Prototyping, usability inspections, and usability lab tests are used more often in later phases of design, after a conceptual prototype has been developed and can be evaluated. It is important to note, though, that any of these usability methods can be used throughout the design process. For example, a questionnaire can be administered to users at the start of the design phase to determine the types of tasks that users are likely to perform on the website, in the middle of the development life cycle to determine what users like or dislike about the current version of the website and what they would like to see in the final version of it, or near the completion of the site to obtain users' feedback and impressions of the overall functioning of the site.

Instrumented Web browsers can also be used to conduct automated sessions of remote usability tests. Participants recruited for these tests are provided with the instrumented version of the browser, for which they are asked to complete certain tasks by visiting one or more websites. Performance measures such as time on task, successes and failures, links followed, and files opened are recorded. Users can also indicate when they have completed a task (either successfully or unsuccessfully) by clicking on a "finish" button and then be prompted to continue to the next task. Automated sessions may be best for summative usability evaluations (Jacques and Savastano, 2001). In fact, West and Lehman (2006) found no major differences in findings (task success rates and user satisfaction ratings) from an automated usability test and one conducted by a usability engineer. Although observational data from the usability engineer provided more information to the design team about usability issues, the written comments provided by users from the automated sessions were sufficient to identify primary usability issues relating to task failures.

### 4.1.1 Prototyping

Prototyping is a particularly useful tool to evaluate the usability of a Web design, especially at earlier phases of development. Alternative designs for the site can quickly be "mocked up" and tested by usability experts or end users. Prototypes can be as basic as drawn images and features on a piece of paper (low fidelity) or more advanced, to mimic the "look and feel" of a real website (high fidelity). The main distinction between a prototype website and a real one is that the prototype is not fully functioning but, rather, is a representation of the site along with simulations of the functions and features of the site. For example, designers can use a high-fidelity prototype of a website to study the presentation of the search results page by having users click on a search button to take them to a predesigned results page. That is, the results page is not returned by the actual search but is linked to "fake" data.

Low-fidelity prototypes consist of paper mock-ups, storyboards, or paper prototypes. These prototypes are usually hand drawn, without the detail or polished look of a high-fidelity design. The goal of using low-fidelity prototypes is to convey the conceptual design rather than to test its features. As shown by Newman et al.'s (2003) study with expert Web designers, paper prototypes are still commonly used, even though it may not take much time to implement the same design in HTML to be presented on the Web. The Web environment does, however, make it easier to design a "skeleton" website with interactive features (Pearrow, 2000):

- *Paper Prototypes*. Paper prototypes are hand-drawn designs of different screens or static printouts of screen shots. They allow designers to focus on the global properties of the website and extreme flexibility for changing components of the design. Functional aspects of the design can be illustrated via buttons and links and additions to the design (e.g., to mimic a drop-down menu) can be made on Post-It notes.

- *Interactive Prototypes*. As aforementioned, it is easier to build interactive prototypes of a website using HTML. Pearrow (2000) made two distinctions for developing interactive prototypes: horizontal and vertical. The emphasis for a horizontal prototype is to enable all the top-level functions, whereas the emphasis for a vertical prototype is to enable the functions along a particular path (e.g., a task to be completed).

### 4.1.2 Usability Inspection Methods

Usability inspection methods are typically used to evaluate the usability of a system without testing end users (see, e.g., Cockton et al., 2008). Inspection methods cannot replace user testing (see, e.g., Vu et al., 2011) but can reveal many usability problems quickly, before end users have to be recruited and tested. They are also typically considered to be less expensive than many other methods used to evaluate usability because the expert needs to be able to access the website only during his or her evaluation. User-based testing,

on the other hand, often requires the cost associated with setting up a usability lab and payment for the usability expert as well as the end users participating in the test. Inspection methods are generally based on design guidelines or recommendations, best practices, and particular findings derived from research studies and theoretical frameworks. The two best known and most frequently used inspection methods are the heuristic evaluation and cognitive walkthrough. However, since the Web can be accessed from a variety of platforms, alternative viewing tools are often used to ensure the usability of the site when accessed by different computing devices.

***Heuristic Evaluation***    With a heuristic evaluation, usability experts examine the functions, features, layout, and content of a website and determine whether the site's format, structure, and functions are consistent with established guidelines or design recommendations (e.g., Nielsen, 1993). Some of these heuristic guidelines for Web design include (based on Pearrow, 2000):

- Chunk together related information.
- Use the inverted pyramid style of writing.
- Place important information "above the fold."
- Avoid gratuitous use of features.
- Make the pages scannable.
- Keep download and response times low.

Usually, three to five evaluators are needed to find the majority (at least two-thirds) of the usability problems (Nielsen and Molich, 1990). A single evaluator typically finds only about 35% of usability problems (Nielsen, 1993). Furthermore, the more familiar and experienced the evaluators are with human factors and usability engineering, the more effective they become at identifying usability problems. In addition to evaluating whether the website adheres to design principles and recommendations, evaluators can determine whether the site contains common Web design mistakes and remove them before they become usability problems. The top 10 mistakes in Web design from 1996 to 2007 posted by J. Nielsen are listed in Table 2.

***Cognitive   Walkthrough***    With cognitive walkthrough, usability evaluators "walk through" the steps that a user would execute when performing specific tasks on the website. The evaluator tries to perform the task from the user's perspective and identifies any problems that users are likely to encounter. The method focuses on how easy the website is to use and on how easy the functions in the site are to learn and use (e.g., Polson et al., 1992). For each step of the task, the evaluators are encouraged to ask themselves the following questions (e.g., Wharton et al., 1992):

- Will the user form the right goal for the task?
- Will the user notice that the correct action is available?
- Will the user associate the correct action with the correct control or feature?
- Will the user receive feedback about their progress in the task?

Before the evaluators perform the cognitive walkthrough, they should prepare a list of tasks to be performed and questions to be answered when performing each task to guide them through the process. Evaluators need to ensure that the types of tasks they will perform during the evaluation are representative of the tasks that the users typically perform and that a range of difficulty levels be included. The walkthrough itself can sometimes be a tedious and slow process, but there have been some suggested streamlined versions of the cognitive walkthrough that are intended to be more efficient (e.g., Spencer, 2000).

Blackmon et al. (2002, 2003) introduced a variant of the cognitive walkthrough designed especially for the Web. Their Cognitive Walkthrough for the Web (CWW) evaluation process simulates the users' step-by-step interactions with the website in a goal-directed task, such as searching for specific information in the site. CWW uses latent semantic analysis to identify usability problems associated with heading/link titles. More specifically, CWW can identify headings or links that are (1) likely to be confusable with other headings/links in the page or site, (2) unfamiliar to the target audience, and (3) likely to lead to a specific goal (i.e., the goal can be classified under competing headings). Blackmon et al. (2003) used CWW to identify usability problems associated with the headings/link titles of an experimental website designed to present encyclopedia articles. They showed that the performance on the site improved significantly after the site was repaired for problems identified by CWW.

In general, cognitive walkthroughs are particularly beneficial for helping the design team to think along the lines of users' goals and knowledge. Cognitive walkthroughs, though, are less successful than heuristic evaluations at identifying more serious usability problems (see Cockton et al., 2008).

### 4.1.3  Evaluations with Alternative Viewing Tools

Because there are many computing devices that allow users to access the Web from any place at any time,

**Table 2  Top 10 Web Design Mistakes, 1996–2007**

1.  Bad search
2.  PDF files for online reading
3.  Not changing the color of visited links
4.  Nonscannable text
5.  Fixed font size
6.  Page titles with low search engine visibility
7.  Anything that looks like an advertisement
8.  Violating design conventions
9.  Opening new browser windows
10. Not answering users' questions

*Source*: Based on Jakob Nielsen's Alertbox, available at http://www.useit.com. Readers should visit useit.com to read more details about the mistakes.

it is important that designers also consider how the website will be displayed on a variety of browsers (e.g., Internet Explorer vs. Firefox) and platforms (e.g., Web TV, PDAs, mobile phones). Alternative viewing tools are viewers, or software tools, that are intended to show designers how a website would look on different devices and platforms. In addition to the platform issue, Web designers should also keep several earlier versions of popular browsers because it may take up to two or three years before users "upgrade" from one browser version to another (Nielsen, 2000).

### 4.1.4 Usability Lab Tests

Usability lab testing is probably the best method for evaluating the website in terms of its effectiveness in promoting successful interactions with the users. Most usability experts would agree that the methods described earlier are useful tools when evaluating Web usability, but they cannot replace usability lab tests. As a result, many of the techniques described above are used in addition to usability lab testing at various phases of the design life cycle. The basic methodology and procedure on how to conduct a usability test have been described in Chapter 46 (see also Pearrow, 2000; Brink et al., 2002; Dumas and Fox, 2008) and are discussed only briefly in this chapter.

A usability lab test for websites involves observing and analyzing users' task performance (e.g., successful completion rates and failure rates, task completion times, and errors) while interacting with a website as well as their thought processes during the task (e.g., understandings, misunderstandings, and preferences). Ideally, usability lab tests should occur several times throughout the design phase so that iterative evaluations can be made.

During the test session, the evaluators should be separated from the participant by a one-way mirror, so that the participant can be observed but the evaluator does not "get in the way." Communication with the participant is usually done through an intercom system, and the evaluators should allow time for users to adjust to this communication medium. If the designer wants to evaluate the effectiveness of alternative Web designs or page layouts, the users' performance with the various versions of the website can be evaluated. Before bringing users into the lab, designers should determine the goal for the test (e.g., test the navigation paths or the efficiency with which the task can be performed), the population from which the sample users will be recruited, the specific tasks that users are to perform, the procedures that are to be followed, as well as how the data will be coded and analyzed:

1. *Sample of Representative Users*. The users that are recruited to participate in the usability test should be as representative as possible of the targeted user group. Each test participant is usually asked to perform several tasks and to provide detailed information about their interaction with the website. Because of the extensive nature of usability testing, relatively few users are selected and tested. Usually four or five users are needed

to find about 80% of the usability problems (Brink et al., 2002), but a larger sample size (e.g., 8–10 users) can increase the number of usability problems detected as well as providing a more representative sample of the target population.

2. *Sample of Representative Tasks*. The specific tasks that users are asked to perform should be based on a subset of the core user tasks that were identified earlier in the design process. The task should be representative, in terms of both the type of task (search or navigation) and level of difficulty. If the version of the website has not been placed on the Web, the designers should simulate response times for the system to match the connection speed that the end users are likely to encounter.

3. *Procedures to Be Followed*. Typically, a usability test starts with an introductory session in which participants are given an opportunity to familiarize themselves with the lab setup and ask any questions that they might have. Participants are usually given consent forms to sign which include information about the nature of the test, duration of the test, type of compensation, confidentiality of their data, and a statement indicating that their participation is voluntary. Some organizations also require participants to sign a nondisclosure agreement. During the test, the experimenter should try to keep the participants on task but should not have too much interaction with the participants so that the experimenter does not bias the outcome of the test. The experimenter should be empathetic when participants get frustrated during the test. Participants should not leave the test thinking that it was their fault that they had difficulties with the tasks. After the usability test is over, questionnaires or interviews can be administered to collect information about the participants' background and/or their opinions about the tasks.

Although the costs of performing usability tests are much higher than those associated with performing usability inspections, usability testing is better than usability inspection methods at finding general, severe, and recurring problems. It is generally agreed that usability lab testing should not be skipped just because the design was tested previously for usability with inspection methods.

### 4.1.5 Evaluating Accessibility for Users with Disabilities

By definition, accessibility is different from usability because it is focused on providing basic access to the information and functionality needed to accomplish user tasks rather than making that functionality more user friendly. But at its core accessibility is usability for users with disabilities, or at least it should be. This is because the motivation behind creating an accessible design should be "usable accessibility" (Henry, 2002), where users with disabilities not only are able to access information and functionality but also are able to enjoy

a quality user experience. Testing for this quality user experience for users with disabilities must come later though because first we must handle the basic (technical) access to information and functionality on a website.

There are many different tools to help a Web designer evaluate the technical accessibility of a website, but they generally fall into one of two categories (several fall into both), automated testing tools and manual testing toolbars:

- Automated testing tools
  - *Use*. These tools conduct an automated check of a website (or page) by going through the HTML and CSS files to determine if they conform to Section 508 standards, WCAG, or both (depending on selected options) and then typically present a list of accessibility successes, warnings, and failures for the website (or page).
  - *Example*. Deque Ramp (http://www.deque .com/products/ramp).
- Manual testing toolbars
  - *Use*. These toolbars are added on to browsers and contain multiple tools that can be selected to aid in the accessibility evaluation of a Web page (e.g., a tool to highlight all images on the page that do not have alt text). They obviate the need for evaluators to dig through lines of HTML in order to determine whether Web pages have been marked up properly.
  - *Example*. The Web Developer Extension for Mozilla-based Browsers (http://chrispederick .com/work/web-developer/).

Unfortunately, automated testing tools are not able to provide accurate accessibility information about a website's conformance to many standards/guidelines because many of them require human interpretation. For example, one of the WCAG 1.0 guidelines states that a text equivalent should be provided for every nontext element. There is no way for an automated testing tool to determine whether the alternative text used for an image is a text equivalent for that image (i.e., fulfills the same function as the original image). Furthermore, much of the information that automated testing tools do provide needs to be validated with manual testing. Thus, automated testing tools are generally used to find accessibility problem areas within websites with large numbers of pages and to guide testing with manual testing toolbars. However, it is never a good idea to completely skip the manual evaluation of a Web page because the automated testing tool did not report any accessibility problems. It is appropriate, though, to evaluate the accessibility of one type of page and apply the results to other similar pages (e.g., pages that use the same template).

After technical accessibility has been examined (and repaired where necessary), the evaluation can move on to the determination of the usable accessibility of the website. It is important that technical accessibility barriers be repaired before the evaluation of usable accessibility takes place because technical accessibility

allows users to interact with the interface. For example, technical accessibility in the form of allowing a screen reader to work with a website for a user with visual impairments is the equivalent of providing a monitor for a user without visual impairments to work with the website. The actual testing for usable accessibility closely resembles usability testing and many of the same methods for evaluating usability can be used (e.g., interviews, participatory evaluation, heuristic evaluations, walkthroughs, user testing; see, e.g., Henry, 2007). Although creating the most accessible websites for users with disabilities requires their inclusion in the design process, many of the interaction issues they face can be teased out by experts who know how (and with what technology) users with disabilities navigate websites and who can use screening techniques to simulate barriers similar to those faced by users with disabilities. While it may at first seem a daunting prospect to user test both accessibility and usability of a website, it should be remembered that users with disabilities come to websites to complete the same tasks as users without disabilities and so can help to evaluate accessibility and usability simultaneously.

## 4.2  Summary

This section reviewed usability evaluation methods for Web design. Some of the methods were geared toward understanding the user, whereas others were focused on finding design problems. It is important to note that all of these methods have different strengths and weaknesses. Thus, designers need to decide which method would be most appropriate to their goals, which can differ depending on the nature of the website and stage of the design process. From a human factors point of view, the design process should be an iterative one, where improvements are made based on the findings from multiple usability evaluations. We endorse a multi-method approach for evaluating Web usability because different methods should yield converging findings for many relevant design issues but also provide unique insights into other important factors that may influence user trust and use of a website.

In industry, thorough user testing and multiple iterations may be prevented by time and monetary constraints. To overcome this barrier, Medlock et al. (2005) proposed the use of a Rapid Iterative Testing and Evaluation method (RITE method). The RITE method emphasizes rapid changes to the design as soon as a usability problem is identified and a solution is available. With RITE, changes can be made after a single participant identifies a problem rather than after a complete user test with multiple participants. The benefit of incorporating quick changes is that the next participant can evaluate the "improved" design, reducing the iterative design process time. It should be noted that the RITE method relies on traditional usability evaluation methods summarized in this section and is a tool for designers to use when they are under time constraints. This method should not replace traditional methods that are more thorough in nature, as those methods tend to provide the most complete usability findings.

## 5 CONCLUSIONS

Websites should be designed and evaluated for usability from their conception and throughout the development cycle. The impact of designing for Web usability can be seen by examining case studies in which the return of investment is immense. Below are just two examples of usability success stories (see also Bertus and Bertus, 2005; Richeson et al., 2011).

- Dell Computer's investment to improve the usability of its e-commerce website in the fall of 1999 led to a dramatic increase in sales, going from $1 million per day in September 1998 to $34 million per day in March 2000 (Black, 2002).

- The United States Computer Emergency Readiness Team (US-CERT) is a government agency charged with "providing response support and defense against cyber attacks for the Federal Civil Executive Branch (.gov) and information sharing and collaboration with state and local government, industry and international partners" (see http://www.us-cert.gov/aboutus.html). Because the two target audiences of US-CERT consist of technical computer professionals and nontechnical users (e.g., home and business users), they wanted to make sure that their website was usable for both groups. After conducting usability tests and implementing changes to the site that addressed user concerns, they found that both technical and nontechnical user success rates improved over 20%. Moreover, satisfaction with the site improved 16% for technical users and 93% for nontechnical users (US-CERT, 2005).

As illustrated by these examples, it is wise to take usability into account when designing websites. Web resources can change on an hourly, daily, weekly, monthly, or yearly basis. Thus, it is important that the website be maintained so that it provides accurate content and continues to be highly usable. However, maintaining a good website is a challenge for designers because new technologies continue to emerge, making it difficult to keep up with the evolving Web.

## REFERENCES

Access Board, The (2000), "Section 508: Section 508 Standards," available: http://www.section508.gov/index.cfm?FuseAction=Content&ID=12, accessed June 11, 2010.

Altinsoy, E., and Hempel, T. (2011), "Multimodal User Interfaces: Designing Media for the Auditory and Tactile Channels," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, pp. 127–151

Anderson, J. R. (1982), "Acquisition of Cognitive Skill," *Psychological Review*, Vol. 89, pp. 369–406.

Bailey, R. W. (1993), *"Performance Versus Preference," in Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 282–286.

Baker, J. R. (2003), "The Impact of Paging vs. Scrolling on Reading Online Text Passages," available: http://www.surl.org/usabilitynews/51/paging_scrolling.asp, accessed June 27, 2010.

Bayles, M. E. (2002), "Designing Online Banner Advertisements: Should We Animate?" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing our World, Changing Ourselves*, Association for Computing Machinery, Minneapolis, MN, pp. 363–368.

Bernard, M., Baker, R., and Fernandez, M. (2002), "Paging vs. Scrolling: Looking for the Best Way to Present Search Results," *Usability News 4.1*, available: www.usabilitynews.org.

Bertus, E., and Bertus, M. (2005), "Determining the Value of Human Factors in Web Design," in *Handbook of Human Factors in Web Design*, R. W. Proctor and K.-P. L. Vu, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 679–687.

Bidgoli, H., Ed. (2004), *The Internet Encyclopedia*, Wiley, New York.

Bitterman, N., and Shalev, I. (2004), "The Silver Surfer: Making the Internet Usable for Seniors," *Ergonomics in Design*, Vol. 12, pp. 24–28.

Black, J. (2002, December 4), "Usability Is Next to Profitability," *Business Week Online*, available: http://www.businessweek.com/technology/content/dec2002/tc2002124_2181.htm, accessed June 27, 2010.

Blackmon, M. H., Polson, P. G., Kitajima, M., and Lewis, C. (2002), "Cognitive Walkthrough for the Web," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing Our World, Changing Ourselves*, Association for Computing Machinery, Minneapolis, MN, pp. 463–470.

Blackmon, M. H., Kitajima, M., and Polson, P. G. (2003), "Repairing Usability Problems Identified by the Cognitive Walkthrough for the Web," in *Proceedings of the Conference on Human Factors in Computing Systems, CHI 2003*, Association for Computing Machinery, Ft. Lauderdale, FL, pp. 497–504.

Bouch, A., Kuchinsky, A., and Bhatti, N. (2000), "Quality Is in the Eye of the Beholder: Meeting Users' Requirements for Internet Quality of Service," in *Proceedings of CHI 2000*, Association for Computing Machinery, Amsterdam.

Brink, T., Gergle, D, and Wood, S. D. (2002), *Usability for the Web: Designing Web Sites That Work*, Morgan Kaufmann, San Francisco.

Caldwell, B., Cooper, M., Reid, L. G., and Vanderheiden, G., Eds. (2008), "Web Content Accessibility Guidelines (WCAG) 2.0," World Wide Web Consortium (W3C), W3C Recommendation 11 December 2008, available: http://www.w3.org/TR/2008/REC-WCAG20-20081211/, accessed June 11, 2010.

Caldwell, B. B, and Vanderheiden, G. C. (2011), "Access to Web Content by Those with Disabilities and Others Operating under Constrained Conditions," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, pp. 371–401.

Chen, B., Wang, H., Proctor, R. W., and Salvendy, G. (1997), "A Human-Centered Approach for Designing World Wide Web Browsers," *Behavior Research Methods, Instruments and Computers*, Vol. 29, pp. 172–179.

Chiew, T. K. (2009), "Web Page Performance Analysis," Ph.D. Thesis, Department of Computer Science, University of Glasgow, March 24, 2009, available: http://theses.gla.ac.uk/658/01/2009chiewphd.pdf, accessed June 14, 2010.

Chisholm, W., Vanderheiden, G., and Jacobs, I. (2001, July–August), "Web Content Accessibility Guidelines 1.0," *Interactions*, pp. 35–54.

Chou, E. (2002), "Redesigning a Large and Complex Website: How to Begin, and a Method for Success," in *Proceedings of the 30th Annual ACM SIGUCCS Conference on User Services*, Association for Computing Machinery, Providence, RI, pp. 22–28.

Chung, W. (2008), "Web Searching in a Multilingual World," *Communications of the ACM*, Vol. 51, No. 5, pp. 31–40.

Cockton, G., Woolrych, A., and Lavery, D. (2008), "Inspection-Based Evaluations," in *The Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, 2nd ed., A. Sears and J. A. Jacko, Eds., CRC Press, Boca Raton, FL, pp. 1171–1190.

Cooper, A. (1999), *The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity*, Howard W. Sams, Indianapolis, IN.

Dix, A., and Shabir, N. (2011), "Human–Computer Interaction," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, pp. 35–61.

Dorn, B., and Guzdial, M. (2010), "Learning on the Job: Characterizing the Programming Knowledge and Learning Strategies of Web Designers," in *Proceedings of CHI 2010: End-User Programming II,* Atlanta, GA.

Dumas, J. S., and Fox, J. E. (2008), "Usability Testing: Current Practice and Future Directions," in *The Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, 2nd ed., A. Sears and J. A. Jacko, Eds., CRC Press, Boca Raton, FL, pp. 1129–1150.

Eastman, C. M., and Jansen, B. J. (2003), "Coverage, Relevance, and Ranking: The Impact of Query Operators on Web Search Engine Results," *ACM Transactions on Information Systems*, Vol. 21, pp. 383–411.

Eisen, P. (2011). "Design of Portals," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, pp. 303–327.

Equation Research on behalf of Gomez, Inc. (2009), "Why the Mobile Web Is Disappointing End-Users: A Study of Consumers' Mobile Web Experiences," White paper, available: http://www.gomez.com/download/wp_mobile_web_experience_survey.pdf, accessed June 11, 2010.

Fang, X., Chen, P., and Chen, B. (2005), "User Search Strategies and Search Engine Interface Design," in *Handbook of Human Factors in Web Design*, R. W. Proctor and K.-P. L. Vu, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 193–210.

Fisk, A. D., Rogers, W. A., Charness, N., Czaja, S. J., and Sharit, J. (2004), *Designing for Older Adults: Principles and Creative Human Factors Approaches*, CRC Press, Boca Raton, FL.

Flach, J. M. (2000), "Discovering Situated Meaning: An Ecological Approach to Task Analysis," in *Cognitive Task Analysis*, J. M. Shraggen, S. F. Chipman, and V. L. Shalin, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 87–100.

Fleming, J. (1998), *Web Navigation: Designing the User Experience*, O'Reilly & Associates, Sebastopol, CA.

Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., and Tauber, E. R. (2003), "How Do Users Evaluate the Credibility of Web Sites? A Study with over 2,500 Participants," in *Proceedings of the 2003 Conference on Designing for User Experiences*, Association for Computing Machinery, New York.

Forrester Consulting on behalf of Akamai Technologies, Inc. (2009), "eCommerce Web Site Performance Today: An Updated Look at Consumer Reaction to a Poor Online Shopping Experience," White paper, available: http://wwwfp.akamai.com/dl/whitepapers/ecommerce_website_perf_wp.pdf, accessed June 14, 2010.

Galletta, D. F., Henry, R., McCoy, S., and Polak, P. (2004), "Web Site Delays: How Tolerant Are Users?" *Journal of the Association for Information Systems*, Vol. 5, pp. 1–28.

Gomez, Inc. (2010), "Great Expectations: What Users Want from the Mobile Web Experience," White paper, available: http://www.gomez.com/download/wp_mobile_great_expectations.pdf, accessed June 14, 2010.

Greenspan, R. (2004), "Google Gains Overall, Competition Builds Niches," available: http://www.clickz.com/3362591, accessed June 27, 2010.

Hadar, I., and Leron, U. (2008). "How Intuitive Is Object-Oriented Design?" *Communications of the ACM*, Vol. 51, No. 5, pp. 41–46.

Halverson, T., and Hornof, A. J. (2004), "Link Colors Guide a Search," in *Proceedings of CHI 2004*, Association for Computing Machinery, Vienna, pp. 1367–1370.

Henry, S. L. (2002), "Another–Ability: Accessibility Primer for Usability Specialists," *UIAccess,* available: http://uiaccess.com/upa2002a.html, accessed June 14, 2010.

Henry, S. L. (2007), "Accessibility in User-Centered Design," in *Just Ask: Integrating Accessibility Throughout Design*, Madison, WI, Copyright © 2003–2007 Shawn Lawton Henry, All Rights Reserved., available: http://www.uiaccess.com/accessucd/background.html, accessed June 11, 2014.

Horrigan, J. (2009), "Wireless Internet Use," Pew Internet & American Life Project, available: http://www.pewinternet.org/~/media//Files/Reports/2009/Wireless-Internet-Use.pdf, accessed June 28, 2010.

iperceptions (2010), "Retail/E-Commerce Industry Report: Q1 2010, 1 Jan–31 March," available: iperceptions.com/resource-center, downloaded June 22,2010.

Irie, R. L. (2004), "Web Site Design," in *The Internet Encyclopedia*, Vol. 3, H. Bidgoli, Ed., Wiley, New York, pp. 768–775.

Isaacs, E. A. (1997), "Interviewing Customers: Discovering What They Can't Tell You," in *Proceedings of the Conference on Human Factors in Computing Systems, CHI 1997*, Association for Computing Machinery, Atlanta, GA, March 22–27, pp. 180–181.

Jacko, J., Sears, A., and Borella, M. S. (2000), "The Effect of Network Delay and Media on User Perceptions of Web Resources," *Behaviour and Information Technology*, Vol. 19, pp. 427–439.

Jacques, R., and Savastano, H. (2001), "Remote vs. Local Usability Evaluation of Web Sites," in *Proceedings of IHM-HCI 2001*, Association for Computing Machinery, Lille, France, September 10–14, pp. 91–92.

Kammerer, Y., and Gerjets, P. (2011), "Searching and Evaluating Information on the WWW: Cognitive Processes and User Support," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, pp. 283–302.

Kantor, P. L. (2003a), "Information Structures," available: http://academ.hvcc.edu/~kantopet/site_design/index.php?page=info+structures&parent=organizing+info, accessed July 4, 2004.

Kantor, P. L. (2003b), "Using Labels," available: http://academ.hvcc.edu/~kantopet/site_de-sign/index.php?page=using+labels&parent=labeling+ systems, accessed July 4, 2004.

Katsanos, C., Tselios, N., and Avouris, N. (2008), "AutoCard-Sorter: Designing the Information Architecture of a Web Site Using Latent Semantic Analysis," in *Proceedings of CHI 2008: Examining Web Content*, Association for Computing Machinery, Florence, Italy.

Klein, D. (1990), "Foiling the Cracker: A Survey of, and Improvements to, Password Security," in *Proceedings of the USENIX Security Workshop*. The Advanced Computing Systems Association, Boston, MA.

Kobayashi, M., and Takeda, K. (2000), "Information Retrieval on the Web," *ACM Computing Surveys*, Vol. 32, pp. 144–173.

Kuniavsky, M. (2008), "User Experience and HCI," in *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*, 2nd ed., A. Sears and J. A. Jacko, Eds., CRC Press, Boca Raton, FL, pp. 897–916.

Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W., Eds. (2007), *Handbook of Latent Semantic Analysis*, Lawrence Erlbaum Associates, Mahwah, NJ.

Lazar, J. (2003), "The World Wide Web," in *The Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, J. A. Jacko and A. Sears, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 714–730.

Lazar, J., and Huang, Y. (2002), "Designing Improved Error Messages for Web Browsers," in *Human Factors and Web Development*, J. Ratner, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 167–182.

Liao, H., Guo, Y., Savoy, A., and Salvendy, G. (2010), *Content Preparation Guidelines for the Web and Information Appliances: Cross-Cultural Comparisons*, CRC Press, Boca Raton, FL.

Lightner, N. J. (2003), "What Users Want in E-Commerce Design: Effects of Age, Education, and Income," *Ergonomics*, Vol. 46, pp. 153–168.

Lightner, N. J., Bose, I., and Salvendy, G. (1996), "What Is Wrong with the World Wide Web? A Diagnosis of Some Problems and Prescription of Some Remedies," *Ergonomics*, Vol. 39, pp. 995–1004.

Mayhew, D. J. (2011), "The Web UX Design Process—A Case Study," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, pp. 461–480.

Medlock, M., Wixon, D., McGee, M., and Welsh, D. (2005). "The rapid iterative test and evaluation method: Better products in less time" in *Cost-justifying usability: An update for the information age*, R. Bias and D. Mayhew, Eds., San Francisco, CA: Morgan Kaufmann, pp. 489–517.

Michaelis, J. R., Golbeck, J., and Hendler, J. (2011), "Organization and Structure of Information Using Semantic Web Technologies," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, pp. 231–247.

Millen, D. (2000), "Rapid Ethnography: Time Deepening Strategies for HCI Field Research," in *Proceedings of the Conference for Designing Interactive Systems: Process, Practices, Methods and Techniques*, Association for Computing Machinery, New York, pp. 280–286.

Najjar, L. (2001), "E-Commerce User Interface Design for the Web," in *Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents, and Virtual Reality*, Vol. 1, M. J. Smith, G. Salvendy, D. Harris, and R. J. Koubek, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 843–847.

Net Market Share (2010a), "Top Browser Share Trend: June, 2008 to May, 2010," available: http://marketshare.hitslink.com/browser-market-share.aspx?qprid=1&qptimeframe=M&qpsp=113&qpnp=25, accessed June 28, 2010.

Net Market Share (2010b), "Browser Version Market Share: May, 2010," available: http://marketshare.hitslink.com/browser-market-share.aspx?qprid=2, accessed June 28, 2010.

Newman, M. W., and Landay, J. A. (2000), "Sitemaps, Storyboards, and Specifications: A Sketch of Web Site Design Practice," in *Proceedings of the Symposium on Designing Interactive Systems, DIS 2000*, Association for Computing Machinery, New York, pp. 263–274.

Newman, M. W., Lin, J., Hong, J. I., and Landay, J. A. (2003), "DENIM: An Informal Web Site Design Tool Inspired by Observations of Practice," *Human–Computer Interaction*, Vol. 18, pp. 259–324.

Nielsen, J. (1993), *Usability Engineering*, Academic, San Diego, CA.

Nielsen, J. (2000), *Designing Web Usability*, New Riders, Indianapolis, IN.

Nielsen, J. (2001), "First Rule of Usability? Don't Listen to Users," available: http://www.useit.com/alertbox/20010805.html.

Nielsen, J., and Molich, R. (1990), "Heuristic Evaluation of User Interfaces," in *Proceedings of Conference on Human Factors in Computing Systems, CHI 1990*, Association for Computing Machinery, Seattle, WA, April 1–5, pp. 249–256.

Nielsen, J., and Tahir, M. (2002), *Homepage Usability: 50 Websites Deconstructed*, New Riders, Indianapolis, IN.

Nielsen, J., Snyder, C., Molich, R., and Farrell, S. (2000), *E-Commerce User Experience*, Nielsen Norman Group, Fremont, CA.

Pearrow, M. (2000), *Web Site Usability Handbook*, Charles River Media, Hingham, MA.

Pew Internet & American Life Project. (2009a), "Trend Data: Online Activities, Total," available: http://www.pewinternet.org/Trend-Data/Online-Activites-Total.aspx, accessed June 28, 2010.

Pew Internet & American Life Project. (2009b), "Trend Data: Online Activities, Daily," available: http://www.pewinternet.org/Trend-Data/Online-Activities-Daily.aspx, accessed June 28, 2010.

Plaue, C., Miller, T., and Stasko, J. (2004), "Is a Picture Worth a Thousand Words? An Evaluation of Information Awareness Displays," in *Proceedings of Graphics Interface 2004*, Canadian Human–Computer Communications Society, London, Ontario.

Polson, P. G., Lewis, C., Rieman, J., and Wharton, C. (1992), "Cognitive Walkthroughs: A Method for Theory-Based Evaluation of User Interfaces," *International Journal of Man–Machine Studies*, Vol. 36, pp. 741–773.

Proctor, R. W., and Vu, K.-P. L. (2010), "Human Factors and Ergonomics," in *The Handbook of Technology Management*, Vol. 1, H. Bidgoli, Ed., Wiley, Hoboken, NJ, pp. 1057–1067.

Proctor, R. W., Lien, M.-C., Vu, K.-P. L., Schultz, E. E., and Salvendy, G. (2002a), "Influence of Restrictions on Password Generation and Recall," *Behavior Research Methods, Instruments, and Computers*, Vol. 34, pp. 163–169.

Proctor, R. W., et al. (2002b), "Content Preparation and Management for Web Design: Eliciting, Structuring, Searching, and Displaying Information," *International Journal of Human–Computer Interaction*, Vol. 14, pp. 25–92.

Proctor, R. W., Vu, K.-P. L., Najjar, L., Vaughan, M. W., and Salvendy, G. (2003), "Content Preparation and Management for E-Commerce Websites," *Communications of the ACM*, Vol. 46, No. 12, pp. 289–299.

Qin, J. (2004), "Web Content Management," in *The Internet Encyclopedia*, Vol. 3, H. Bidgoli, Ed., Wiley, Hoboken, NJ, pp. 687–698.

Rasmussen, J. (1986), *Information Processing and Human–Machine Interaction: An Approach to Cognitive Engineering*, North-Holland, Amsterdam.

Ratner, J. (2003), *Human Factors and Web Development*, Lawrence Erlbaum Associates, Mahwah, NJ.

Rau. P. L. P., Plocher, T., and Choong, Y.-Y. (2011), "Cross-Cultural Web Design," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, pp. 677–697.

Richeson, A., Bertus, E., Bias, R. G., and Tate, J. (2011), "Determining the Value of Usability in Web Design," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, 753–764.

Riddle, B. L., Miron, M. S., and Semo, J. A. (1989), "Passwords in Use in a University Timesharing Environment," *Computers and Security*, Vol. 8, pp. 569–579.

Rieman, J. (1993), "The Diary Study: A Workplace-Oriented Research Tool to Guide Laboratory Efforts Collecting," in *Proceedings of the Conference on Human Factors in Computing Systems, INTERCHI '93*, Association for Computing Machinery, Amsterdam, April, 24–29, pp. 321–326.

Roberts-Witts, S. L. (2001, September), "A Singular Focus," *PC*, available: http://www.pcmag.com/article2/0,4149,16651,00.asp, accessed June 27, 2010.

Rose, G. M., Evaristo, R., and Straub, D. (2003), "Culture and Consumer Response to Web Download Time: A Four Continent Study of Mono and Polychronism," *IEEE Transactions on Engineering Management*, Vol. 50, pp. 31–44.

Schultz, E. E. (2011), "Web Security, Privacy, and Usability," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, pp. 663–676.

SearchEngineWatch.com (2009), "Top Search Providers for August 2009," available: http://searchenginewatch.com/3634991, accessed June 22, 2010.

Sears, A., Jacko, J. A., and Dubach, E. M. (2000), "International Aspects of World Wide Web Usability and the Role of High-End Graphical Enhancements," *International Journal of Human–Computer Interaction*, Vol. 12, pp. 241–261.

Shadbolt, N., and Burton, M. (1995), "Knowledge Elicitation: A Systematic Approach," in *Evaluation of Human Work*, 2nd ed., J. R. Wilson and E. N. Corlett, Eds., Taylor & Francis, London, pp. 406–440.

Shneiderman, B. (1997), "Designing Information-Abundant Web Sites: Issues and Recommendations," *International Journal of Human–Computer Studies*, Vol. 47, pp. 5–29.

Spencer, R. (2000), "The Streamlined Cognitive Walkthrough Method: Working Around Social Constraints Encountered in a Software Development Company," in *Proceedings of the Conference on Human Factors in Computing Systems, CHI 2000*, Association for Computing Machinery, The Hague, The Netherlands, April 1–6, pp. 353–359.

Stanney, K. M., Maxey, J., and Salvendy, G. (1997), "Socially-Centered Design," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York, pp. 637–656.

Stephanidis, C., and Akoumianakis, D. (2011), "A Design Code of Practice for Universal Access: Methods and Techniques," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, pp. 359–370.

Sullivan, D., and Sherman, C. (2001), "Search Engine Features for Webmasters," available: http://searchendingewatch.com/webmasters/features.html.

Tullis, T., and Albert, B. (2008), *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*, Morgan Kaufmann, Burlington, MA.

Tullis, T. S., Tranquada, F. J., and Siegel, M. J. (2011), "Presentation of Information," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, pp. 153–189.

U.S. Census Bureau, Population Division (2008), "National Population Projections," available: http://www.census.gov/population/www/projections/summarytables.html, accessed June 11, 2010.

U.S. Computer Emergency Readiness Team (US-CERT) (2005), "Usability Lessons Learned," available: http://www.us-cert.gov/usability/, accessed June 23, 2010.

U.S. Department of Commerce, Economics and Statistic Administration and National Telecommunications and Information Administration (2002), "A Nation Online: How Americans Are Expanding Their Use of the Internet," retrieved July 5, 2004, from http://www.ntia.doc.gov/ntiahome/dn/anationonline2.pdf.

U.S. Department of Commerce, Economics and Statistic Administration and National Telecommunications and Information Administration (2010), "Digital Nation: 21st Century America's Progress toward Universal Broadband Internet Access," available: http://www.ntia.doc.gov/reports/2010/NTIA_internet_use_report_Feb2010.pdf, accessed June 11, 2010.

Vaughan, M. W., Candland, K. M., and Wichansky, A. M. (2001), "Information Architecture of a Customer Web Application: Blending Content and Transactions," in *Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents, and Virtual Reality*, Vol. 1, M. J. Smith, G. Salvendy, D. Harris, and R. J. Koubek, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 833–837.

Vicente, K. J., and Rasmussen. J. (1992), "Ecological Interface Design: Theoretical Foundations," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 22, pp. 589–606.

Volk, F., Pappas, F., and Wang, H. (2011), "User Research: User-Centered Methods for Designing Web Interfaces," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, pp. 417–438.

von Ahn, L., and Dabbish, L. (2004), "Labeling Images with a Computer Game," in *Proceedings of CHI 2004,* Association for Computing Machinery, Vienna.

Vu, K.-P. L., and Proctor, R. W. (2003), "Naïve and Experienced Judgments of Stimulus–Response Compatibility: Implications for Interface Design," *Ergonomics*, Vol. 46, pp. 169–187.

Vu, K.-P. L., and Proctor, R. W., Eds. (2011), *Handbook of Human Factors in Web Design*, 2nd ed., CRC Press, Boca Raton, FL.

Vu, K.-P. L., Proctor, R. W., Bhargav-Spantzel, A., Tai, B.-L., Cook, J., and Schultz, E. E. (2007), "Improving Password Security and Memorability to Protect Personal and Organizational Information," *International Journal of Human-Computer Studies*, Vol. 65, pp. 744–757.

Vu, K.-P. L., Zhu, W., and Proctor, R. W. (2011), "Evaluating Web Usability," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, pp. 439–460.

West, R., and Lehman, K. R. (2006), "Automated Summative Usability Studies: An Empirical Evaluation," in *Proceedings of CHI 2006: Automatic Generation & Usability*, Association for Computing Machinery, Montreal, Quebec, Canada.

Wharton, C., Bradford, J., Jeffries, R., and Franzke, M. (1992), "Applying Cognitive Walkthroughs to More Complex User Interfaces: Experiences, Issues, and Recommendations," in *Proceedings of the Conference on Human Factors in Computing Systems, CHI 1992,* Association for Computing Machinery, Monterey, CA, May, 3–7, pp. 381–388.

Wisman, R. (2004), "Web Search Fundamentals," in *The Internet Encyclopedia*, Vol. 3, H. Bidgoli, Ed., Wiley, Hoboken, NJ, pp. 724–737.

Withrow, J., Brink, T., and Speredelozzi, A. (2000), "Comparative Usability Evaluation for an E-Government Portal," Diamond Bullet Design Report U1-00-2, Diamond Bullet Design, Ann Arbor, MI.

Xu, S., and Fang, X. (2011), "Mobile Interface Design for M-Commerce," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL.

Zeng, J. and Duan, J. (2011), "Analyzing and Modeling User Activity for Web Interactions," in *Handbook of Human Factors in Web Design*, 2nd ed., K.-P. L. Vu and R. W. Proctor, Eds., CRC Press, Boca Raton, FL, pp. 645–662.

Zhang, P. (2000), "The Effects of Animation on Information Seeking Performance on the World Wide Web: Securing Attention or Interfering with Primary Tasks?" *Journal of the Association for Information Systems*, Vol. 1, pp. 1–28.

# CHAPTER 49

# HUMAN FACTORS IN AMBIENT INTELLIGENCE ENVIRONMENTS

**Constantine Stephanidis**
**Institute of Computer Science Foundation for Research and Technology—Hellas (FORTH)**
**Heraklion, Crete, Greece**
**Department of Computer Science**
**University of Crete**
**Crete, Greece**

## 1 INTRODUCTION

A vision of ambient intelligence (AmI) is offered in the report of the Information Society Technologies Advisory Group of the European Commission (ISTAG, 2003, p.1): "The concept of Ambient Intelligence provides a vision of the Information Society where the emphasis is on greater user-friendliness, more efficient services support, user-empowerment, and support for human interactions. People are surrounded by intelligent intuitive interfaces that are embedded in all kinds of objects and an environment that is capable of recognising and responding to the presence of different individuals in a seamless, unobtrusive and often invisible way."

As an emerging field of research and development, AmI is rapidly gaining wide attention by an increasing number of researchers and developers worldwide. The notion of AmI is becoming a de facto key dimension of the emerging information society spanning across every human–computer interaction (HCI) research and development domain since next-generation digital products and services are explicitly designed in view of an overall intelligent computational environment. The latter reflects a

fundamentally interdisciplinary research approach necessitating knowledge fusion and technology transfer.

Computer networks, sensor/actuator technologies, user interface software, pervasive computing, artificial intelligence, adaptive systems, robotics, and agent systems are currently seen as the primary domains with a strong impact on AmI research (Nakashima et al., 2010). Computational vision and sensor networks acquire information from the environment regarding events and activities which take place. This information is then elaborated by high-level reasoning modules for behaviour monitoring and for determining the intelligent reaction of applications to what happens in the environment.

While a wide variety of different technologies are involved, the goal of AmI is fundamentally dual: on the one hand to hide the presence of its technological infrastructure from the end users as much as possible and on the other hand to smoothly integrate in everyday objects, thus making it "disappear." The design requirements of an AmI system are: (i) *unobtrusiveness*—devices are distributed in the environment, embedded into different physical objects, becoming invisible to humans unless visibility is needed; (ii) *personalization*—its

behavior can be configured to address individual user requirements; and (iii) *adaptation*—it is capable of automatically modifying its behavior relying on the recognition of user and context characteristics, including individual preferences, without requiring conscious mediation [International Organization for Standardization (ISO), 1999]. To this end, AmI systems support interactivity based on the continuous interpretation and processing of tasks, activities, and contexts (Aghajan et al., 2009).

The AmI environments are anticipated to have a profound impact on the everyday life of citizens in the information society and to potentially permeate a wide variety of human activities. They will affect the type, content, and functionality of the emerging products and services as well as the way people will interact with them (e.g., Emiliani and Stephanidis, 2005; Coroama et al., 2004; Edwards and Grinter, 2001). In the near future, it is expected that the resulting technologies will redefine the way people today understand and use computers. On the other hand, the anticipated omnipresence of technology also raises doubts and fears. For example, continuous monitoring of human presence and activities through sensors may be difficult to accept and raises concerns of privacy and security. Additionally, the question arises of what consequences may be caused by the potential failure of a technological environment which is so deeply intertwined with all sorts of human activities and how people can be protected from negative consequences (Bohn et al., 2005). In this context, as the border between really useful supportive technology and the so-called technological nightmare becomes much thinner, it has been argued that it is of critical importance for AmI environments to develop and evolve from the very beginning with a strongly human-oriented focus towards ensuring that the emerging potential is fully realized and potential pitfalls are avoided. As stated in (Roe, 2007, p. 157), "As all technological innovations, Ambient Intelligence is not good or bad per se, but its impact on people will depend on how it is deployed and used, the time and scale of deployment and the care devoted to involve people in its development."

Therefore, human factors acquire fundamental importance in AmI environments. In this context, *human factors* as defined by Czaja and Nair (2006, p. 32) is "the study of human beings and their interaction with products, environments and equipment in performing tasks and activities. The focus of human factors in on the application of knowledge about human capabilities, limitations, and other characteristics in the design of human-machine systems . . . .The general objectives of human factors are to maximize human and system efficiency and health, safety, comfort and quality of life."

People, their social situation, ranging from individuals to groups, and their corresponding environments and activities (office buildings, homes, public spaces, etc.) become the focus of design considerations and inform the design and interactive behavior of artifacts, applications and services, and ensembles thereof. Besides, human factors knowledge is also instrumental in defining the intelligence of the environment, that is, creating environments capable of "understanding" and fulfilling human

needs and unobtrusively supporting human activities. In order to exhibit more humanlike understanding of human needs and activities and provide contextually appropriate feedback, AmI environments need to acquire sensor information about humans and their functioning as well as to extract adequate knowledge for analysis of such information about human functioning (Aarts and de Ruyter, 2009; Bosse et al., 2010). In other words, human-centered AmI environments can only be achieved through a deep understanding of human activities, interaction, and communication (Nijholt et al., 2009).

On the basis of the above considerations, this chapter discusses the centrality and role of human factors in the emergence and development of AmI environments. In particular, Section 1 analyzes the user-centered design process in the light of the requirements posed by AmI, focusing on emerging problems and potential solutions toward applying and revising existing methods and techniques or developing new ones. Section 2 focuses on some user experience factors which are considered as critical in AmI, including natural interaction, accessibility, cognitive demands, emotions, health, safety and privacy, social aspects, cultural aspects, and aesthetics. For each of them, a brief overview of the main issues involved is provided. Section 3 illustrates three case studies involving the user-centered design of AmI artifacts, discussing for each of them the main adopted design process and the relevant user experience qualities. Finally, Section 4 summarizes the emerging research challenges and puts forward the need for a more systematic approach to the above issues.

## 2 HUMAN-CENTERED DESIGN PROCESS IN AMI ENVIRONMENTS

User centered-design (UCD) has been defined as an approach to designing ease of use into the total user experience with products and systems (Vredenburg et al., 2001) and constitutes a cornerstone of human factors engineering. It involves two fundamental elements: multidisciplinary teamwork and a set of specialized methods for acquiring user input and converting it into design.

The need for UCD has been identified quite early in the HCI field, and numerous articles and books have been written on the subject (Gould, 1988; Hewett and Meadow, 1986; Norman and Draper, 1986). ISO 13407 (1999) provides guidance on human-centered design activities throughout the life cycle of interactive computer-based systems. User-centered design is not simply conducting usability studies or talking to users; it emphasizes the active involvement of users and requires studies to understand users as well as to drive and evaluate the design and the final system. As technology rapidly advances, computer use spreads to almost every domain of everyday life, and new interaction paradigms emerge, UCD remains through the years a fundamental process for the design of every interactive system. In this light, ISO 13407 has been revised to ISO 9241-210 (ISO, 2010) to reflect recent changes and advances which highlight six main principles of UCD: The design is based upon an explicit understanding of users, tasks, and

environments; users are involved throughout design and development; the design is driven and refined by user-centered evaluation; the process is iterative; the design addresses the whole user experience; and the design team includes multidisciplinary skills and perspectives.

UCD includes four iterative design activities, all involving direct user participation, as shown in Figure 1:

1. Understand and specify the context of use, the nature of the users, their goals and tasks, and the environment in which the product will be used

2. Specify the user and organizational requirements in terms of effectiveness, efficiency, and satisfaction and the allocation of function between users and the system

3. Produce designs and prototypes of plausible solutions

4. Carry out user-based assessment

In the context of AmI, the importance of UCD increases even further, but its complexity also increases and the need emerges for rethinking some aspects of the process and of the related methods. In AmI, "focusing on the human" goes beyond previous approaches and truly refers to a new paradigm in developing and using technology. In such a new paradigm, ease of use, unobtrusive design, privacy management, and personalization constitute the main building blocks of the definition of a human-centric interface. AmI environments are by definition capable of adapting to their inhabitants, as they embed two fundamental features, monitoring and reasoning, which make them dynamically reactive to what happens within them.

The next sections discuss each of the UCD activities from the point of view of AmI in an attempt to anticipate the main challenges which will need to be addressed and outline potential directions of investigation.

## 2.1 Context of Use

The context of use in user interface design is usually considered as including the characteristics and roles of users, their goals and tasks, as well as the characteristics of the physical, social, and technological environment where a system is used.

In AmI, the context of use becomes much more complex and articulated as the number and potential impact of relevant factors increase dramatically with respect to conventional computing devices, particularly

regarding the (co-)presence of people, computing devices, and other elements in the surrounding environment. In other words, interaction is no longer a one-to-one relationship between a human, a specific task, and a specific device in a static context. Rather, it becomes a many-to-many relationship between diverse users and a multitude of devices in a dynamically changing environment, where none of the typical context-of-use factors can be assumed as stable over the interaction. It is believed that the key to identifying suitable design approaches in the context of AmI relies precisely in the interplay between the user, the physical and social environment, the available technologies, and the tasks of users. All these factors are highly dynamic and interrelated and vary over time.

Additionally, in AmI, the context of use needs to be not only appropriately captured but also modeled and embedded in the technological infrastructure in such a way as to facilitate the sensing and monitoring of the related parameters.

### 2.1.1 Defining Context of Use in AmI

In defining the AmI context of use, care must be taken to accommodate the complexity deriving from omnipresence, "disappearance," and intelligence of technology.

For example, the very concept of "user" needs to be revisited (Stephanidis, 2001). Different levels of awareness and participation to the communication and computing processes may occur in the AmI environment, according to the characteristics and requirements of different individuals, but also according to the characteristics of the spatial/temporal/technological contexts of human activity. Humans are no longer simply "operators" or "users" but may play different roles in the overall process.

Additionally, a user's abilities may change over short time intervals according to various conditions partly determined by external conditions (e.g., other elements of the environment). Also, as users move in the environment, the current context of use as well as the current interaction technologies may dynamically change (e.g., a different room with different characteristics and technologies). Many users with different needs and requirements may be present at the same time in an environment, and interact concurrently, and this fact may introduce potential conflicts. For example, in smart homes, all inhabitants who are present at the same time are users, but possibly each one is conducting different activities which may interfere with someone else's activities. As a result, *role awareness*, in addition to *user awareness*, becomes
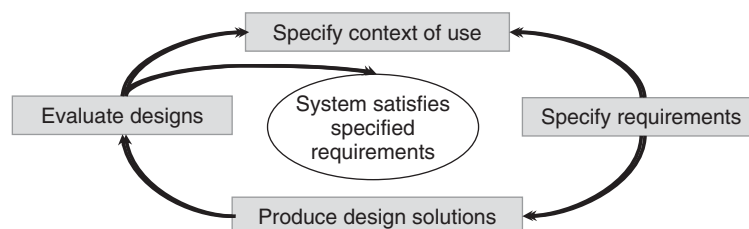


**Figure 1**   Activities involved in UCD (from ISO, 1999).

prominent for adaptation in AmI systems. Additionally, the context may adaptively change due to decisions made by AmI applications and not only according to the current user location (e.g., according to different times of the day of different days). Also, devices can be plugged in or removed from the environment at any time. The dynamic availability of services is also likely to lead to user tasks changing on the fly.

As a consequence, the context of use in AmI should be captured in a dynamic rather than static way, posing various challenges for the methods to be used, the types of information to be gathered, and the representations to be adopted. Indicatively, Table 1 illustrates the context-of-use factors typically included in UCD, whereas Table 2 presents a modified version taking into account factors which are relevant in AmI.

Many of the factors in Table 2 are dynamic, that is, they may depend on other context factors (e.g., available devices may depend on location of the user). As a consequence, the definition of the context of

use in AmI is also strictly related to the issue of monitoring, as it is critical in identifying the elements to be monitored, as well as the conditions and parameters according to which monitoring should take place. In other words, context-of-use factors are not only what the designers should know but also what the system should be aware of. This need is commonly referred to as context awareness, defined as "The environment can determine the context in which certain activities take place, where context relates meaningful information about persons and the environment, such as positioning and identification" (Aarts and de Ruyter, 2009, p. 8).

### 2.1.2 Methods and Techniques for Capturing the Context of Use

Methods commonly adopted in UCD to capture the context of use include stakeholders' identification, surveys, paper-based questionnaires, field study/user observation, diary keeping, and task analysis (Maguire, 2001).

**Table 1  Context-of-Use Factors**

| User Group | Tasks | Technical Environment | Physical Environment | Organizational Environment |
|---|---|---|---|---|
| System skills and experience | Task list Goal | Hardware Software | Auditory environment | Work practices Assistance |
| Task knowledge | Output | Network | Thermal environment | Interruptions |
| Training qualifications | Steps | Reference materials | Visual environment | Management and |
| Language skills | Frequency | Other equipment | Vibration | communications structure |
| Age and gender | Importance | | Space and furniture | Computer use policy |
| Physical and cognitive capabilities | Duration Dependencies | | User posture | Organizational aims |
| Attitudes and motivations | | | Health hazards Protective clothing and equipment | Industrial relations Job characteristics |

*Source*: From Maguire, 2001, p. 595.

**Table 2  Context-of-Use Factors in AmI**

| User Groups | Human Activities | Technological Environments | Physical Environments | Social Environments |
|---|---|---|---|---|
| Individual vs. group | Goals | Available hardware devices, including input/output | Time parameters (day, time, etc) | Activity practices |
| Language skills | Activities | Network | Auditory environment | Roles |
| Age and gender | Output | Software platforms | Thermal environment | Interruptions |
| Physical and cognitive capabilities | Typical location | Middleware | Visual environment | Communication structure |
| Attitudes and motivations | Tasks and steps | Already existing applications and services | Vibration | People copresence |
| Cultural background | Frequency | Sensors and other equipment | Space, furniture | Privacy threats |
| Knowledge of the environment applications, services, and interactive features | Importance Duration Dependencies and intertwining with other activities | Available data from sensors Available reasoning resources | User position and posture Health hazards | |
| Domain knowledge | Required assistance | | | |
| Interaction preferences | Delegation | | | |
| Emotions and psychological conditions | Collaboration | | | |

In principle, all the above methods and techniques can be useful in the context of AmI. However, their application may vary. For example, observation-based methods may evolve due to the monitoring and reasoning facilities offered by the environment itself, which offer the opportunity to gather larger amounts of data regarding various aspects of human activities in an environment over longer time periods. However, the application of such an approach requires the availability of experimental infrastructures or already existing functional environments as well as the presence of real users behaving and acting naturally in the environment. To address this issue, facilities and laboratories have been set up where different AmI technologies are being developed, integrated, and tested in a real-life context, usually simulating a home environment (e.g., Georgia Tech's Aware Home,[*] MIT's House_n,[†] Philips' HomeLab,[‡] Fraunhofer-Gesellschaft's inHaus,[§] and Microsoft Home[¶]).

On the other hand, more traditional methods such as surveys, questionnaires, and diaries can still be very relevant when no facilities are yet available for experimentation.

The identification and modeling of user activities and tasks constitute a different challenge. Typical task analysis techniques have been claimed to be poorly suited to AmI (Verpoorten et al., 2007), as they do not capture the interrelationships of tasks with other contextual factors. For example, traditional task analysis techniques, due to their origin (business environments and application), are oriented toward highly structured tasks with specific goals and steps. However, in AmI environments, everyday human activities do not always have a specific goal and are characterized by a much looser structure, which may not easily decompose into discrete steps. Various extensions of traditional task models have been proposed in this respect. For example, Vredenburg et al. (2001) discuss an approach to generate activity patterns from task structures. Luyten et al. (2005) propose a task-centered approach to design for ambient intelligent environments. The approach is based on visualization and simulation and is targeted to capture, during design, the strong dependency of task execution on the situation or context of use as well as the consequences of a context change on the execution of task specification.

## 2.2  User Requirements

A prerequisite for the successful development of AmI environments is that user requirements are appropriately captured and analyzed. In particular, it is necessary to anticipate future computing needs in everyday life and acquire an in-depth understanding of the factors which will determine the usefulness of diverse interactive artifacts in context. These requirements are likely to be more subjective, complex, and interrelated than in previous generations of technology.

As a starting point toward analyzing user requirements in AmI, studies have been conducted through the use of scenario techniques. Both positive (ISTAG, 2003) and "worst case" (Wright et al., 2010) scenarios have been proposed.

While scenarios provide a useful starting point, they cannot be considered as sufficient to fully capture future human needs and expectations in AmI environments. On the other hand, for the user community it is not easy to express needs and preferences regarding a technological environment which is hard to imagine. In order to contribute to the development, users should be made aware of the technological possibilities and potential approaches to build up the new environment.

It is therefore necessary to bring users in direct contact with AmI technologies and the possibilities they offer. To this purpose, research infrastructure and simulation environments such as those discussed in the previous section can play a fundamental role.

Studies on human requirements in AmI environments have also started appearing, mainly targeted to the home environment. Röcker et al. (2005) reported a multicultural study combining scenario-based techniques, focus groups, and open-ended discussions to identify requirements for the home environment. A set of prioritized requirements was derived, including, besides general issues, support for housekeeping and safety, assistance for personal environment organization and home organization, and support for care and communication with others.

Hellenschmidt and Wichert (2005) reported an experiment targeted to analyze different kinds of ambient assistance in living room environments. In this experiment, 143 subjects interacted with a home entertaining system presenting integrated functions like TV, radio, audio and video playing, telephone, and light controls. Based on the results, seven kinds of assistance were identified with varying levels of user involvement and awareness (from situations where the user is fully informed about all changes in the environment to situations where the environment changes without any direct action of the user).

Zhang et al. (2009) described an experiment aimed at matching user interface intelligence in smart home environments with the type of task to be carried out and the age of the users. Following Rasmussen et al. (1994), tasks are classified as skill-based tasks, rule-based tasks, and knowledge-based tasks.[‖] Interfaces are also distinguished according to three levels of intelligence (low, medium, and high). The results of the study, involving two user groups of young and older people, respectively, show that the level of intelligence of the

---

[‖] In skill-based tasks, performance is governed by patterns of preprogrammed behaviors in human memory, without conscious analysis and deliberation. In rule-based tasks, performance is governed by conditional rules. Finally, at the highest level of knowledge-based tasks, performance is governed by a thorough analysis of the situation and a systematic comparison of alternative means for action.

user interface, as well as the age of the users, significantly affects performance for different types of tasks. Overall, the need arises for highly usable user interfaces allowing direct and accessible user control, avoiding the need to use a variety of different applications to control different devices, and providing an intuitive overview of the overall current state of the intelligent environment.

Schmidt et al. (2007) reported a study involving a combination of contextual inquiries, cultural probes, technology probes, educating the user, participatory design sessions, and creating prototypes from persona-inspired designs. The presentation of physical prototypes has been contextualized in the possible scenarios of everyday life activities (e.g., "when you wake up," "when you brush your teeth"). This proved to be particularly useful for generating design ideas and for understanding the user profile. Indeed, people find it easier to relate to the task-orientated nature of the scenarios, rather than to the abstract and often function-orientated nature of a system specification. The combination of the two, scenarios of everyday life and tangible previews of future technology, proved to be a powerful method to stimulate their creativity. Using the technology probes, people appeared to be less worried about technologies invading their home, thus reducing concerns and fears. Interviews were also used at a later stage, based on ideas generated in the previous phases.

## 2.3 Producing Designs

The design of AmI environments introduces new challenges, primarily due to the embedded interactivity "hidden" in a variety of interconnected, multifunctional computing artifacts (Emiliani and Stephanidis, 2005). This is due to the multifunctional nature of artifacts that must be smoothly integrated in everyday objects. It is therefore crucial to systematically examine how typical interactive functions combine with the rest of functions offered by such artifacts and the way the design of all functions can be optimally accommodated. In particular, typical HCI design processes will likely need to be revisited in terms of their overall scope. For example, while accessibility and usability of every devices in isolation is necessary, it is still not sufficient to guarantee usability in an overall distributed environment where the combined presence of artifacts could possibly introduce new usability issues. This also implies a closer study of the similarities and differences of the physical world and the digital world, as these worlds will no longer be clearly separated but, on the contrary, will be strictly interrelated and fused. Interaction in AmI environments may take place through a mixture of physical and digital elements. Therefore, a convergence emerges between interaction design and industrial design which will need to be addressed through novel methods and techniques (Butz, 2010).

In the context outlined above, a series of new research challenges emerge. At the level of interaction design, new design principles will be required, stemming from the fusion and extension of design principles of everyday objects with principles of user experience design.

Prototyping and related tool support also constitute an issue for AmI environments. Currently available support for user interface prototyping is mainly based on user interface builders incorporated in most integrated development environments. However, such tools do not fully address the need of prototyping in AmI. For example, they are usually bound to specific devices, interaction platforms, and widget toolkits and are not suited to designing a pervasive user experience across devices. A potential step forward to address this issue is the provision of adaptable interaction toolkits, comprising device-aware widgets which know how to display themselves and how to compose an interface layout on different devices (see Chapter 54 of this handbook, "Design for All: Computer Assisted Design of User Interface Adaptation"). Additionally, toolkits to develop dynamically distributed user-interfaces for mobile users, capable of exploiting for interaction purposes ambient computing resources available at the current location, such as the Voyager toolkit (Savidis and Stephanidis, 2005), are very important for reducing the user interface implementation complexity.

Simulation of AmI environments also allows an instrument to bring future users closer to the new technological environment without requiring extensive development. As stated elsewhere (Bandini et al., 2009), traditional design and modeling instruments can provide a suitable support for designing the static properties of AmI environments, for example, through the construction of three-dimensional (3D) mock-ups, but they are not suitable for defining their dynamic behavior and responsiveness. Simulators, on the other hand, allow envisioning both the static features of the ambient intelligence system as well as its dynamic response to the behavior of humans and other relevant entities situated in it.

Another important change in AmI environment is likely to affect who the designers should be and which design knowledge they should possess. Given the omnipresence of technology, the ideal solution would be for users themselves to be able to take decisions about the interactive features of the environment in which they live. Towards this objective, the environment should be able to provide to users very easy means to personalize interaction. The technological environment should empower end users to tailor the environments to their needs, providing methods and tools for editing, interpreting, linking, executing, and rendering of application functions in a user-friendly and intuitive way (Lieberman, 2001).

An example of an approach in this direction is design and play (Kartakis and Stephanidis, 2010), which allows the quick configuration of user interfaces for smart homes ready to be used. Such an approach can be further simplified and automated based on the consideration that the smart home can detect the presence of interactive artifacts and exploit knowledge about their characteristics. In such an environment, designers are likely to undertake new roles, as their main task will be to encode user experience knowledge in the environment rather than directly produce user interface prototypes. Additionally, designers will need to

be aware of the fact that the interaction behavior of the environment will not be statically determined by design but will be more dynamic and will evolve based on initial design and on data gathered through monitoring.

## 2.4   Evaluation

The evaluation of AmI technologies and environments will also need to go beyond traditional usability evaluation in a number of dimensions, concerning both the qualities of the environment to be assessed and assessment methods.

### 2.4.1   Qualities of AmI Environments

Performance-based approaches to usability are not adequate for AmI environments, as they were developed for single-user, desktop applications and are usually applied in laboratory evaluations. Additionally, it is difficult to specify tasks that capture the complexity of everyday activities, and a more subjective view of the user experience is necessary. Qualities which may need to be taken into account in evaluating AmI technologies and environments include highly subjective cognitive and emotional factors (Theofanos and Scholtz, 2004). Therefore, evaluation should target the overall user experience and the emotional response to technology (Hassenzahl and Tractinsky, 2006) rather than traditional usability. However, the concept of user experience needs to be further articulated in order to derive measurable qualities. For example, Gaggioli (2005) presents an attention-based framework for user experience assessment in AmI. Other factors which may need to be assessed include emotional reactions to technology (Herbon et al., 2006), engagement, and fun (Mandryk et al., 2006) as well as trust and user's perception of security and privacy in the environment (Lo Presti et al., 2005).

Adams and Russell (2007) reported a study that is addressing emotional responses to two intelligent ATM prototypes. The two prototypes were similar, but one of them was purposefully designed to be more stressful, showing a lower level of intelligence in the interaction. The results confirmed that the stressfulness of the prototype caused real problems to users going well beyond traditional usability and produced a strong emotional reaction of frustration.

Qualities which characterize AmI environments from a user experience point of view are discussed in more detail in Section 3.

### 2.4.2   Evaluation Methods and Techniques

Ambient intelligence technologies and systems challenge traditional usability evaluation methods because the context of use can be difficult to re-create in a laboratory setting. This suggests that the evaluation of user's experience with AmI technologies should take place in real-world contexts. However, evaluation in real settings also presents difficulties, as there are limited possibilities of continuously monitoring users and their activities (Intille et al., 2006). For example, the experience sampling method, which aims at capturing the user experience in the field, takes advantage of the wide diffusion of mobile devices to request feedback from people during their activities. This approach allows capturing on-time feedback, avoiding to the users the need for recalling situations and related experience at a later stage. Using this approach it is also possible to contextualize questions based on the time of day, the user location, and other parameters (Vastenburg and Romero Herrera, 2010).

Appropriate facilities are required for combining user experience in context with the availability of the necessary technical infrastructure for studying the users' behavior over extended periods of time. It is expected that future testing and evaluation methods and activities will rely to a large extent upon the simulation of the interactive behavior of the technological environment as well as upon continuous monitoring data. For example, human emotions in the environment can be analyzed through face and speech recognition techniques but also physiological monitoring (Herbon et al., 2006).

As in the case of design, evaluation also tends to acquire a more continuous nature in AmI environments. In the long run, the embedded intelligence allows the environment to assess and improve itself, blurring the distinction between evaluation and interaction. Experience research (Aarts and de Ruyter, 2009) aims at developing methods and techniques that allow the validated feedback of users in the process of generating experiences. In experience research, three main dimensions of user experience evaluation are distinguished: (i) eliciting and validating end-user insights in the societal context, (ii) execution of user-centered design cycles in controlled laboratory environments, and (iii) testing of novel concepts and solutions in real-life settings.

## 3   USER EXPERIENCE IN AMBIENT INTELLIGENCE ENVIRONMENTS

This section discusses some of most important aspects of user experience in AmI environments, with the aim of identifying emerging approaches to defining, modelling, and assessing such factors in AmI environments.

### 3.1   Natural Interaction

The pervasiveness of interaction in AmI environments requires the elaboration of new interaction concepts that extend beyond the current user interface concepts such as the desktop metaphor and menu-driven interfaces (Aarts and de Ruyter, 2009). AmI will therefore bring about new interaction techniques as well as novel uses and multimodal combinations of existing advanced techniques, such as, for example, gaze-based interaction (Gepner et al., 2007), gestures (Ferscha et al., 2007), and natural language (Zimmermann et al., 2004). Progress in computer vision approaches largely contributes to the provision of natural interaction in AmI environments, making available, among other things, techniques for facial expression, gaze and gesture recognition, face and body tracking, and activity recognition.

Additionally, interaction will be embedded in everyday objects and smart artifacts. This concept refers to interfaces that use physical artifacts as objects for

representation and interaction, seamlessly integrating the physical and digital worlds.

Such objects serve as specialized input devices that support physical manipulation, and their shape, color, orientation, and size may play a role in the interaction.

The interaction resulting from tangible user interfaces is not mediated and it supports direct engagement of the user with the environment. Consequently, it is considered more intuitive and natural than the current keyboard and mouse-based interaction paradigm (Aarts and de Ruyter, 2009).

Interaction in AmI environments inherently relies on multimodal input, implying that it combines various user input modes, such as speech, pen, touch, manual gestures, gaze and head and body movements, as well as more than one output modes, primarily in the form of visual and auditory feedback. In this context, *adaptive multimodality* is prominent to support natural input in a dynamically changing context of use, adaptively offering to users the most appropriate and effective input forms at the current interaction context. An implementation framework for adaptive multimodal input for cross-media board games is discussed in Savidis and Lilis (2008). Finally, multimodal input is acknowledged for increasing interaction accuracy by reducing uncertainty of information through redundancy (Lopez-Cozar and Callejas, 2010).

## 3.2 Accessibility

Accessibility in the context of AmI is usually intended as inclusive mainstream product design, although a contextual definition is not yet available. However, a gradual transition from assistive technologies to mainstream accessible products is foreseeable (Antona et al., 2007). Given the variety and plurality of devices in AmI environments, different levels of accessibility may be distinguished. A first level concerns accessibility of individual devices. Personal devices will need to be accessible to their owners, but probably basic accessibility should be provided also for other users with potentially different needs. A second level is the accessibility of the environment as a whole, which may be provided through environmental devices and other interactive artifacts. In this case, accessibility can be intended as equivalent access to content and functions for users with diverse characteristics, not necessarily through the same devices, but through a set of dynamic interaction options integrated in the environment.

It is likely that some of the built-in features of AmI environments, such as multimodality, will facilitate the provision of solutions that will be accessible by design (Carbonell, 2006; Richter and Hellenschmidt, 2004). For example, blind users will benefit from the wider availability of voice input and output. Different modalities can be used concurrently, so as to increase the quantity of information made available or present the same information in different contexts or, redundantly, to address different interaction channels, both to reinforce a particular piece of information or to cater for the different abilities of users. A novel aspect is that in AmI environments the accessibility of the physical and the virtual world needs to be combined. For example, for blind, visually impaired, and motor-impaired users, requirements related to interaction need to be combined with requirements related to physical navigation in the interactive environment. Along the same lines, the complexity of the environment and the disappearing of technologies can become insurmountable obstacles for cognitively impaired users if not properly addressed. Age-related factors are also very important, particularly in the light of the fact that a large part of AmI applications will be targeted to supporting independent living and that current understanding of the needs and requirements of users of different age in such a complex environment is limited.

## 3.3 Cognitive Demands

Ambient Intelligence should not introduce increased complexity for its users. As technology "disappears" to humans both physically and mentally, devices are perceived no longer as computers but as smart interactive artifacts of the surrounding environment (Streitz, 2007). The nature of interaction in AmI environments is the result of evolution from human–computer interaction to human–environment interaction (Streitz, 2007) and human–computer confluence (Ferscha et al., 2007). These concepts emphasize the fusion of technologies and environments and the increased involvement of interaction with digital artifacts in all aspects of human life.

Putting strong emphasis on adaptation and usability, AmI environments need not be interaction intensive, despite the fact that humans are surrounded by a wide range of computing devices of different functionality and scale. Therefore, interaction shifts from an explicit paradigm, in which the users' attention is on computing, toward an implicit paradigm, in which interfaces themselves drive human attention only when required or preferred (Schmidt, 2005). Interaction in the emerging environment will be based no longer on a series of discrete steps but on a continuous exchange of information (Faconti and Massink, 2001). Continuous interaction differs from discrete interaction since it takes place over a relatively longer period of time, in which the exchange of information between the user and the system occurs at a relatively high rate in real time. A first implication is that the system must be capable of dealing in real time with the distribution of input and output in the environment. This implies an understanding of the factors which influence the distribution and allocation of input and output resources in different situations for different individuals.

Due to the intrinsic characteristics of the new technological environment, it is likely that interaction will pose different perceptual and cognitive demands on humans compared to currently available technology (Gaggioli, 2005). It is therefore important to investigate fundamental and essential functions of the brain, including perception, thinking, emotion, learning, memory, attention, heuristic search, planning, reasoning, discovery, and creativity (Aarts and de Ruyter, 2009). The main challenge in this respect is to identify and avoid forms of interaction which may lead to negative

consequences such as confusion, cognitive overload, frustration, and so on.

Adams and Russell (2007) presented a cognitive user model targeted at providing a simple framework for the investigation of cognitive factors in AmI. It identifies nine components of human cognition relevant to the use of AmI technologies, namely input and perception, output and responses, feedback, working memory, emotions, mental models, executive functions, complex response skills, and long-term memory.

## 3.4  Emotions

AmI environments should take the user's emotions into account, that is, they should be emotion aware (Zhou et al., 2007) and, when appropriate, realize affective behavior. This implies that the environment should be able to interpret human emotions, generate responses which embed an emotional dimension, and also be able to influence users' emotions. Emotion awareness can help the environment to fine tune its behavior to offer to the users a better user experience, as understanding emotions is essential to the creation and re-creation of experiences (Westerink et al., 2008). Affect computation may constitute the basis for more advanced adaptive behavior, essentially leading to *affective adaptation*. The latter allows for broader adaptation scenarios in AmI environments, where interaction is continuous and users are involved in numerous interaction sessions with varying roles. An implementation of affective adaptation for games, where even affect computation is realized as an adaptive function, is discussed in Savidis and Karouzaki (2009).

An emotion is a physiological response to a situation. Emotions are commonly related to a set of elements, including anger, fear, happiness, sadness, love, surprise, disgust, and shame. Other emotions may be composed from these basic elements. Emotions are usually simply divided into positive and negative, although more articulated scales can be devised. Various approaches have been proposed to enable AmI environments to sense and measure people's emotions. Facial expression and speech characteristics are both reliable indicators of emotional states. Progress in wireless sensors also allows to unobtrusively monitor physiological parameters which show important correlations to various types of emotions, such as heart rate, electrodermal activity, facial muscle activity, and voice (Herbon et al., 2006).

Work has also been conducted in the area of creating means for expressing emotions in AmI technologies, mainly through lifelike avatars. Ortiz et al. (2007) reported a study which confirmed the capability of older users to recognize avatars' emotion expression as well as the positive effect of emotional avatars on the users' experience.

Emotional intelligence is the human ability to be aware of one's own and other people's emotions and react appropriately. An emotion-aware AmI environment should exhibit similar abilities and provide services which respond to emotions (Zhou et al., 2007).

Examples of the emotional behaviors that AmI environments attempt to stimulate in their inhabitants are affective presence, which aims at stimulating participation and creative inspiration (Boehner et al., 2005), laugher (Melder et al., 2006), playfulness into experiencing and understanding information (Eyles and Eglin, 2007), sense of care (Kröse et al., 2003), and wellness (Coughlin et al., 2009). Various means are exploited to generate or control emotions in AmI environments, from avatars and robots capable of emotional expressions to art, music, lighting, and other environment features.

## 3.5  Health, Safety, and Privacy

In a situation in which technology may act on the physical environment and deal with critical situations without the direct intervention of humans, it is also likely that new hazards will emerge for people's health and safety. Possible malfunctions or wrong interpretations of monitored data can lead to unforeseeable consequences, especially for disabled users who will be more dependent on technology than others. Therefore, appropriate strategies for avoiding risks must be elaborated and validated, with emphasis on users' awareness of the involved issues. A related challenge is the consideration of interoperability among different technologies and devices, as the correct functioning of the intelligent environment as a whole needs to be ensured.

Another relevant requirement is privacy, which concerns the effective protection of personal data that are collected through the continuous monitoring of people in AmI environments. Security and avoidance of human errors which may undermine the privacy of data emerge as fundamental requirements for Ami environments (Kraemer and Carayon, 2007). New challenges arise concerning how a person will be able to know what information is recorded, when, by whom, and for which use (Friedewald et al., 2007). Acceptability levels of data availability in the environment may vary depending on many factors and may ultimately be a matter of trade-off between the need for privacy and the options offered by the available information about the users.

In this context, users' acceptance of continuous monitoring becomes a relevant issue. Such acceptance is likely to depend on the perceived functional benefit of AmI environments and on the availability of mechanisms that enable participants to make their own choices in a way that is understandable, transparent, and independent of their comprehension level (Aarts and de Ruyter, 2009). This amounts to users developing trust in the intelligent environment. Wright et al. (2010) provide an in depth discussion of privacy issues in AmI environments, suggesting a series of countermeasures to ensure that AmI technologies develop in such a way to avoid privacy pitfalls and generate trust. These include technological solutions as well as socioeconomic solutions. Technological solutions include minimal data collection, transmission and storage, data and software security, privacy protection in networking, as well as authentication and access control. Regarding the latter, biometric recognition systems appear to be particularly promising in seamlessly enhancing users' protection in AmI environments. Nontechnological measures to address privacy issues include guidelines, standards, codes of practice, legislation, and public awareness.

Although AmI is only moving its first steps, it already very clearly appears that its human-centered and inclusive development raises a variety of ethical and social issues which necessitate appropriate policy, standardization, and legislative intervention (Bohn et al., 2005; Kemppainen et al., 2007). Relevant legislative areas are, for example, personal data protection, consumer protection, accessibility, and telecommunications regulation. In particular, privacy and security are overwhelming discussion topics in relation to AmI.

## 3.6 Social Aspects

In AmI environments, computing is expected to become far more dependent on social factors than it has in the past. Cooperation will be an important aspect, and communication and access to information will be concurrently used to solve common problems in a cooperative manner; moreover, cooperation may be among human users themselves or among user representatives (agents and avatars), to whom variable degrees of trust can be assigned. Access to information and communications will no longer be the task of an individual but will be extended to communities of users who have at their disposal common (sometimes virtual) spaces in which they can interact. Finally, the dynamic reconfiguration of context will be highly dependent on social phenomena (users moving in the environment, meeting, communicating to each other, collaborating, etc). User adaptation may become more difficult when multiple users interact in the same environment (Søraker and Brey, 2007). A typical example is the automatic selection of TV programs based on user habit. The presence of more than one user is likely to make this process more complex, and "group habits" need to be defined. Therefore, users in AmI environments can no longer be studied only at an individual level, and accounts of social behaviour become equally important.

Social connectedness is another essential element of the user experience in AmI environments and concerns the extent to which the environment behaves "socially" toward its inhabitants (Aarts and de Ruyter, 2009). Three main factors can be identified: (i) adoption of communication protocols that are compliant with societal conventions and follow social rules (socialization); (ii) awareness of user's emotions and adaptation of behavior accordingly (empathy, affective adaptation); and (iii) consistent and transparent behavior in interaction with people, which is recognized by the users as conscientious (consciousness).

Another important aspect of AmI environments is social presence, which reflects the degree to which the environment facilitates and supports human social interaction (Biocca et al., 2003). AmI needs to convey both the physical and virtual presence of people interacting and communicating through the environment. People's presence can be detected through sensors and can be represented, for example, through avatars but also through more subtle cues.

## 3.7 Cultural Issues

While AmI develops at a global level, it may be anticipated that cultural factors will become particularly relevant in identifying and reasoning about users' goals and tasks, which may be highly influenced by different cultural backgrounds (Søraker and Brey, 2007).

One of the main characteristics of AmI is natural interaction. However, what is seen as natural behavior and how certain forms of behavior relate to underlying desires depends to some degree on our cultural background. Behavioral indicators such as the range and importance of gesticulation, facial expressions, and body language can differ radically from one culture to another. In this respect cultural issues play a very relevant role in the AmI environment context, and the design of AmI environments needs to be informed by a deep understanding of the different cultures addressed.

Additionally, cultural values and practices may influence the acceptance of AmI technologies and of the environment as a whole. Bick et al. (2009) reported a study investigating the impact of assertiveness, future orientation, gender egalitarianism, uncertainty avoidance, power distance, institutional collectivism, in-group collectivism, performance orientation, and humane orientation on the acceptance of AmI technologies in hospital settings. The obtained results led to the conclusion that national cultural influence factors have a significant effect on the perceived usefulness of ambient intelligence. In particular, the dimensions of humane orientation, uncertainty avoidance, power distance, institutional collectivism, and in-group collectivism appear to play a significant role such that higher scores in these dimensions lead to greater acceptance of ambient technologies.

## 3.8 Aesthetics

Aesthetics is a subject receiving increasing attention in the design of AmI environments, as it is an essential aspect of everyday usage. Another reason is that the design of AmI technologies, as already observed in Section 2.3, needs to combine interaction design with other design traditions engaged in the design of everyday things, thus bringing into play a very different set of perspectives, values, and approaches (Redström, 2007). Aesthetics concerns the form and appearance of artifacts, addressing questions concerning structure and composition, use of material, overall consistency, and so on. Though the area of aesthetics in the design of AmI environments is still very far from presenting a coherent framework, there are attempts to develop related notions, building on industrial design tradition. Overall, the investigation of aesthetic aspects of interactive environments appears to transcend typical human factors studies based on user-oriented statistics and to involve issues such as (Redström, 2007):

- Engagement
- Temporal structures (e.g., interaction patterns and expressions of use that evolve over time)
- Alternative forms of use that challenge expectations on use and user
- Relations to context, for example, cultural references, user identity, traditions, other design domains
- Alternative interface and material combinations

## 4   CASE STUDIES

This section presents three case studies illustrating some aspects of the development of AmI artifacts in the context of the ICS-FORTH AmI Programme. These examples address different application domains and contexts of use, namely the domain of culture in the museum environment, the domain of education and learning in the classroom environment, and the domain of family activities in the home environment. They are small- to medium-size design cases which inevitably address only a limited part of the large number of issues which arise in a human-oriented approach to AmI, regarding both the design life cycle and the user experience aspects taken into account. However, they are illustrative of the implications of putting focus on users and on their contexts of use when designing AmI environments as well as the importance of qualities such as natural interaction, positive user emotions, playfulness, and aesthetic considerations in envisioning AmI artifacts.

### 4.1   iRoom

iRoom is a multimedia system targeted at archaeological and historical museums which supports the exploration of large-scale artifacts in actual size (e.g., a wall painting, a mosaic, a metope) through noninstrumented, location-based, multiuser interaction (Zabulis et al., 2010).

As its name implies, the system is installed in a room ($6 \times 6 \times 2.5$ m³) in which a computer vision subsystem tracks the location of its occupants. This subsystem is comprised of eight calibrated FireWire cameras ($66° \times 51°$) mounted at the corners and at the in-between midwall points, viewing the room peripherally in steps of $45°$. On one of the room walls a dual-projector back-projection screen ($4.88 \times 1.83$ m²) is installed. Behind the screen lies a control room that contains two $1024 \times 768$ short-throw projectors, two stereo speakers, and three workstations. Additionally, in the room there is an information kiosk and a stand with mobile phones. Mobile phones run a custom application that can receive information about their holder's position and render information accordingly.

iRoom (see Figure 2) can present large-scale images of artifacts with which one or more visitors can concurrently interact simply by walking around.

Visitors enter the room from an entrance opposite the wall painting. The vision system assigns a unique ID number to each person entering the room. In the room entrance there is a "barrier" created by four queue posts that guides visitors to move leftward or rightward along a short corridor in order to walk further into the room. As two help signs illustrate, visitors entering the room from the right-hand side of the corridor are considered to be English speaking, while those from the left-hand side, Greek speaking. When at least one person is detected in the room, a soft piece of music starts to play.

The whole room is semantically split into five zones of interest, delimited by different themes presented on the wall painting. These zones cut the room in five vertical slices (with respect to the projection). Furthermore, the room is also split into four horizontal zones that run parallel to the wall painting which are delimited by their distance from it. Thus, a $5 \times 4$ grid is created comprising 20 interaction slots. When a visitor is located over a slot, the respective wall painting part changes, depending on the slot's distance from the wall. For example, when a visitor is in front of a wall piece but standing at the first zone, an outline of the piece is presented along with a descriptive title. If the visitor moves to the zone that is closest to the wall, specific details of the piece are highlighted and related information is provided. All information is presented in the user's selected language. Since users get a unique ID, the system keeps track of the information they have seen as well as the time they have spent on each slot. This allows assigning more than one piece of information to one slot, which may be presented to the visitor when revisiting it. When multiple visitors concurrently use the system, interaction works as follows. If more than one visitor is standing on the same vertical zone of slots, the person who "controls" the respective wall piece (i.e., information is presented according to his or her



**Figure 2**   The iRoom system.

language preferences and position) is the one standing closest to the wall. If this person leaves, the next in line gets control, and so on.

Beyond location sensing, iRoom also supports two other types of interaction: (a) a kiosk and (b) mobile phones. The kiosk offers an overview of the wall painting, an introductory text, and two buttons for changing the user's language. When a visitor stands in front of the kiosk, all information is automatically presented in the visitor's originally set language. Furthermore, the wall piece in front of which the visitor has spent most of the time is highlighted. Mobile phones are used as multimedia guides, presenting images and text (that can also be read aloud) related to the visitor's position. In order to assign a mobile phone to a specific visitor, the visitor has to stand at a spot in the room denoted by a white X and press the phone's selection button.

### 4.1.1 Human Factors in iRoom

iRoom was designed in the context of a wider project targeted to enhance the museum visit experience through interactive edutainment. A formative evaluation of iRoom has been conducted using ethnographic field methods (Blomberg et al., 1993). To this purpose, the exhibit was installed in a dedicated space resembling an exhibition area of a museum. The use of video recording was avoided, since previous experience has shown that users tend to be more reluctant in freely exploring and experimenting with a system when they know they are being recorded. In the case of the dedicated exhibition space, participants were invited on an ad hoc basis, among people of all ages and cultural/educational background visiting (e.g., politicians, scientists, school classes) or working in FORTH facilities (including their families).

Up to now, more than 100 persons have participated. Typically, evaluation sessions involved a facilitator accompanying the visitors, acting as a "guide," and another distant observer discreetly present in the exhibition space. Since there were numerous evaluation sessions, alternative approaches were used. For example, when, at earlier stages, the interactive behavior of iRoom was tested, the facilitator would first provide a short demonstration to the participants and then invite them to try. Alternatively, when ease of use and understandability were assessed, the facilitator would prompt participants to freely explore the artifact without instructions. During and after the sessions, the facilitator held free-form discussions with the participants, eliciting their opinion and experience and identifying usability problems as well as likes and dislikes. The facilitator kept a small notepad for taking notes. After the visitors eft, the two observers would discuss the session, taking additional notes, often reenacting parts of it, in order to clarify or further explore some findings.

Overall, the opinions of all participants about the exhibit ranged from positive to enthusiastic. People of all ages agreed that they would like to find similar systems in museums they visit. Usually, when visitors were first introduced to the exhibit there was a short exclamation and amusement phase, during which they seemed fascinated by the technology and tried to explore its capabilities but, interestingly, after that most of them spent considerable time exploring the exhibit's content. Over different installations it was observed that the large size and luminous intensity contribute to the enhancement of visitor appreciation of the system.

Language selection came to be considered a considerably challenging task for interactive exhibits, rarely addressed by previous efforts. For example, a kiosk was initially used as a means of language selection. During evaluation it was observed that it created both problems of visitor flow (people had to stand in line) and erratic behavior (multiple visitors standing too close during language selection). The current scheme of implicit selection was an improvement in terms of both usability and robustness.

Other identified problems include intended versus unintended actions in location-based interaction and interaction fuzziness. When a visitor crossed the iRoom (e.g., to move closer to a friend), an avalanche of wall changes were triggered. This fact was especially irritating when the visitor was moving in front of other users, momentarily "stealing" their control over a slot. As a solution, a minimum dwell time was adopted in order for a user to gain control over a slot. Another "grey area" was the room's entrance. Since very often people were just peeking in or wanted to watch what was going on before engaging, the only functionality assigned to the zone close to the door was language selection and start of music play (to provide some reactive behavior to the first user stepping in). Also, visitors standing at the boundaries of slots in iRoom would sometimes be in a state of accidentally (or even worse constantly) switching between them. To remedy this problem, the slot's area where the user is has been enlarged.

### 4.2 AmIDesk

AmIDesk (Antona et al., 2010) is an augmented school desk designed in the context of a research project targeted to investigate the potential of AmI technologies in enhancing the learning experience in the classroom through integrating ambient interaction and digital augmentation of physical paper. It consists of an additional piece of furniture designed to fit typical school desks. Such an "add-on" provides a custom plexiglass 27-in.-diagonal wide screen whose inclination can range from $30^\circ$ (with respect to desk surface) to completely horizontal. It embeds almost invisibly all the devices required for the operation of the AmI applications and has a width of 40 cm, thus requiring relatively limited additional space with respect to the standard desk.

The AmIDesk includes:

- An Intel Core 2 Quad Core PC running MS Windows 7
- Two DLP miniprojectors located behind the screen
- One mirror for reducing the projection distance
- Two cameras located behind the screen
- Infrared projectors located behind the screen

- One camera located on top of the screen and capturing images of the conventional desks
- One smart pen and its transmitting device.

To better exploit the custom screen dimensions, a horizontal window manager was developed which includes two application and two menu areas. The screen supports multitouch interaction using the software reported in Michel et al. (2009) (see Figure 3). The smart pen input is captured through the Pegasus SDK.[*] The front vision software, currently under refinement, supports the identification of pages and other features of printed material as well as the recognition of simple gestures. Its combination with the desk's multitouch display allows augmenting physical learning materials with additional context-dependent information.

AmIDesk integrates a set of software applications for enhancing the learning experience of English as a second language. These include a login screen, a welcome screen, an individual personal area summarizing the current delivery status of all assignments, a dashboard where students can temporarily save material, assigned exercises in electronic form with the supported hints and help, a dictionary–thesaurus application, a personal dictionary application, a note-taking application, an application for viewing course-related multimedia, and language-learning games (e.g., hangman). The dictionary–thesaurus presents a short definition of the word, its pronunciation, a button allowing the student to hear the pronunciation, some synonyms, and a few examples of use in a sentence. In addition, the thesaurus offers options for extended descriptions, grammar information, a complete list of synonyms and antonyms, and several examples.

Although AmIDesk is designed for individual use, collaboration is foreseen for some tasks. For example,



**Figure 3**    The AmIDesk touch screen.

the students can exchange materials and the teacher can send materials (e.g., assignments) to the personal areas of all students through simple gestures. The smart environment automatically restricts actions that can be carried out by the students according to the current context. For example, when a reading task is active, the students cannot use some functionality in their system (e.g., multimedia, saving, printing), and when an exercise task or a test is being carried out, students are not allowed to send material to other student's personal areas. The system will also produce statistics regarding the hints that have been requested, per student and for the whole class, so the teacher will be able to monitor the student's progress. The system also monitors the success rate of each exercise and provides statistics. The teacher can then combine this with the hint statistics to review results and adjust the difficulty (remove/alter exercise) in order to improve the learning curve. The system can also measure the time needed by each student to complete a specific task.

### 4.2.1   Human Factors in AmIDesk

The approach followed in the design of AmIDesk was user centered, involving a small group of young learners from the very first steps of the design and targeting to provide intuitive and seamless tools to improve the learning and classroom experience.

The classroom constitutes a challenging context of use for the design of AmI technologies. In practice, there are severe space and layout limits to the introduction of AmI equipment which should be unobtrusive, hidden, or embedded in traditional classroom equipment and furniture. It is very important that such equipment can be installed smoothly and easily moved around in the environment and that space requirements are as limited as possible. This implies several constraints on how the AmI classroom environment can be developed. Additionally, legislation concerning school furniture must be taken into account. In the case of AmIDesk, for example, EU normative regarding standards dimensions of school desks was considered in order to establish the dimensions of the add-on artifact.

The requirements for AmIDesk were originally elaborated through scenarios, with the overall goal of ensuring that the resulting design goes at the heart of the learning process. To this purpose, learning of English as a foreign language was selected as a testing domain, and current practices were examined to identify useful support which can be offered through AmI technologies, focusing on preparation for the first and advanced certificates (thus addressing an adolescent student population). Based on such an analysis, extensive usage scenarios were compiled. The scenarios mainly address activities which take place in the classroom; however, an important consideration is that the software to be developed should be general enough to also support learning at home or elsewhere, independently from the classroom infrastructure and the augmented desk itself. In particular, the scenarios address the seamless context-dependent provision of useful additional information during language-learning activities. Interaction with the provided facilities is based on gestures, either

through the desk screen or directly on paper resources. For example, the learner can indicate a word on the page to view additional information about it or the answer to a fill-the-gap exercise to receive feedback or hints. Text entry during learning activities is based mainly on handwriting through a smart pen. However, an on-screen keyboard is also currently under design to support small text entry tasks on the touch screen. The desk should also be able to identify each user through a personal object (e.g., a school diary, pen, etc). The developed scenarios led to the identification of a set of software applications to complement the desk.

Following the development of the AmIDesk prototype, a formative evaluation experiment was conducted involving five young learners of English as a foreign language (one male and four females aged 11–16, all studying for a first or advanced certificate in English, and all familiar with PCs and mobile phones but not with AmI environments). The experiment was targeted to collect users' opinion regarding:

- The desk itself
- The overall idea of the interactive student desk in the AmI class and how it can assist learning
- The usefulness of each application regarding the English course and in particular the thesaurus, the multimedia application, the personal area for assignments and homework delivery, the myVocabulary area, and the hints and confirmations during exercises
- The UI layout and aspects of the supported gesture interaction

The experiment took place individually for each learner. After a very brief introduction, the learners were driven through a simplified scenario illustrating the main aspects of the desk and the related applications.

During the execution of the experiment, the children were asked questions about various aspects of the scenario, and notes were taken with all their comments. After the completion of the scenario, they were asked to fill out a questionnaire composed of 17 questions. Of the questions, formulated in an informal style to appeal to young learners, 15 used a Likert scale from 1 (sure!) to 5 (no way!). The remaining 2 questions concerned listing aspects of the scenario that the children particularly liked or disliked. Overall, the results were very positive, as all the children involved in the experiment were very interested about the desk and its applications, and some were enthusiastic about having such an artifact available in their classroom.

The preferred features of the desk were the personal area and the dictionary followed by the educational games, the dashboard, the hints and confirmations, touch interaction, pointing at things and viewing info, and the electronic submission of assignments. The disliked features were mainly the desk size and color (the green color had been used for facilitating vision processing) and, with respect to the user-interface (UI) mockups, again the colors and the fonts.

The young learners appreciated the educational support which the desk aims to provide as well as the potential for better organization of work between the classroom and the home environment and collaboration with teachers and other learners. Some of the children also proposed to include a grammar application similar to the thesaurus application, displaying grammar rules, verb tables, and so on, related to the task at hand. On the other hand, the young learners appeared to view the desk as a "trendy gadget" and were very sensitive to aesthetic issues, asking for the possibility of personalizing the desk color, selecting fonts, colors, background images, and avatar images. Regarding interaction, they appreciated the gesture-based applications, but they also asked for more traditional interaction means such as the keyboard.

The results of this investigation are currently being explored in the development of a larger set of applications as well as of a toolkit of pervasive widgets supporting interaction in the classroom environment.

### 4.3 booTable

booTable is an interactive "smart" coffee table prototype, accompanied by a couple of stools, constructed by recycled paper and designed to look as a modern piece of furniture (Grammenos et al., 2010) (see Figure 4). It builds upon the paradigm of surface computing but endeavors to overcome the identified limitations of current design practices. All technological parts (except the inevitable power cord) are carefully concealed in its body through the creation of appropriate design features. booTable fuses alternative, highly versatile input technologies and provides dual visual output, complementing the projection surface with a secondary display channel.

Multitouch input is supported through physical signal measuring. The adopted sensors include four touch buttons, a circular touch, and a vibration sensor. They are placed immediately below the surface in a completely invisible way. A Wiimote above the table's surface is



**Figure 4**  The booTable beta prototype.

also used to track IR-tipped pens and IR-fingertips. Identification of various objects is achieved using a radio-frequency identification (RFID) reader.

booTable is accompanied by two matching stools that also embed sensing technologies and by a set of tabl-e-cloths. Tabl-e-cloths realize an innovative concept. They are RFID-tagged printed canvas sheets that act both as a tangible means for launching software applications and as the basic layer for creating mixed-reality interfaces. booTable stools also embed a wireless sensor for detecting user presence.

booTable comes with a large variety of family-oriented applications, including a book and DVD information retriever, a photo slide show and album, a family notes application, a creative painting application, a chess game, an invaders game, a dual screen storyteller, a mobile phone hub, various e-clocks, and smart light control.

### 4.3.1   Human Factors in booTable

booTable underwent an iterative design process through the realization of two different prototypes with quite different characteristics. The initial design requirements were based on the one hand on the limitations identified in current tabletop interactive systems and on the other hand on the context of use of family activities. Identified requirements include:

- Table functionality: The artifact would be a regular coffee table of the type usually found in the living or sitting room. The surface would be appropriate to support beverages, magazines, books, game boards, and small items such as keys, mobile phones, and remote controls.

- It should have distinctive, attractive, and ideally innovative design. It should be something that people would like to put in their living rooms as is, even if bare from its interactive behavior.

- It should look like a piece of furniture and not an electronic device or gadget. All technological components and their traces should be hidden. When the system is off or inactive, it should still be useful and presentable.

- It should fuse multiple input and output technologies.

- It should be able to somehow change its appearance in order to be personalized and fit in multiple spaces.

- It should target all types of families, with any number of members (even just one), of different ages.

- It should strive for high usability and ease of use by the broadest possible spectrum of user population without requiring previous computer expertise.

- The tabletop should support a large diversity of everyday tasks, including leisure activities.

- The construction cost should be as low as possible.

- It should be as flexible as possible, so that design ideas that did not work out could easily be "removed" and new ones could be tried out without having to reconstruct the prototype from scratch.

Based on the above requirements, a set of usage scenarios was elaborated, and a first wooden prototype of booTable was built. After developing the required software, an informal assessment of the concept and its implementation was conducted during the time course of one week with the voluntary participation of 27 individuals of both genders with ages varying from 15 to 62 years old. Fourteen of them were technology savvy while the rest were random visitors, friends, and so on. The assessment process was short and very flexible. It basically consisted of a brief introduction to the concept, vision and goals of the project, a demonstration of the developed applications during which participants were prompted to interact with the system, and finally a discussion aimed at eliciting impressions from the overall experience, positive and negative aspects of the implementation, ideas for additional applications, and the estimated usefulness and desirability of such a future product. User verbalizations were hand noted by one of the members of the development team.

Overall, the participants' reactions and first impressions were very positive. All of them showed a vivid interest in the concept and stated that they enjoyed interacting with the table. Especially the non- technology-savant ones were rather surprised by the fact that furniture may come equipped with interactive functionality. Most of them regarded it as "useful," "fun," and "impressive" and stated that, at a price similar to a flat-screen TV or a PC, they would probably consider purchasing it. Users were also prompted to help identify as many drawbacks of the prototype as possible. As a result, a number of issues, from simple concerns (e.g., the table was probably too high for a coffee table) to serious usability problems (e.g., RFID tagging and cataloguing books), were identified. In addition, several other problems, mainly related to the selected construction materials and hardware, were raised by the development team. In summary, the conducted evaluation led to abandoning wood as a construction material, as the artifact was too heavy and bulky and needed to be placed against a wall. Additional issues concerned the quality of graphics, the performance of some of the sensors, the low refresh rate of the secondary display (an electronic picture frame), the cumbersome use of RFID tags for retrieving books and DVD information, and the need for more detailed context information for the proper functioning of the smart light application. Based on the above, a beta version of booTable was developed, entirely realized in recycled paper, and the necessary hardware modifications and enhancements were implemented. These included the replacement of the PC to provide better graphics, the replacement of the photo frame with a 7-in. touch screen, a rearrangement of the sensors underneath the table surface, and the installation of a bar code scanner. Various software modules were also modified, and new ones were added.

Upon its completion, the beta prototype of booTable was showcased to a large audience at the exhibition of a major HCI conference. The overall impression regarding both the look and "feel" was very encouraging, while the fact that it was made by recycled paper created a lot of positive reactions. Since many of the people that experienced it were HCI practitioners, many fruitful discussions were held which included interesting ideas about potential improvements and new applications.

## 5 EMERGING CHALLENGES

The issue of human factors in AmI environments is particularity complex, and the current state of the art is still far from offering consolidated practices regarding how to design AmI environments in a human-oriented fashion and how to address the user experience in a more rigorous and scientifically sound way. In this context, a number of new research challenges emerge, including:

- Investigation of human characteristics, abilities, and requirements in the context of AmI
- Suitable accounts and models of the context of use; implies investigating how to model, embed, and reason about user experience qualities in order for AmI environments to exhibit intelligent behavior
- Definition of a user experience framework for AmI environments, taking into account interaction naturalness, accessibility, cognitive demands, emotions, health, safety and privacy, social and cultural aspects, and aesthetics, and elaboration of related assessment criteria and metrics
- Multidisciplinary approaches to defining acceptable levels of safety and privacy risks in AmI environments and establishing related standards, regulations, and technical solutions
- Elaboration of design methods suitable for very complex interactive environments
- Reappraisal of industrial design methods and techniques for integrating interactive and other functional characteristics of artifacts and environments

Given the complexity of AmI technologies and their high level of interdependence with the use context, it is believed that substantial progress toward facing the above issues will be brought about as AmI environments further develop. This is particularly important when taking into account the large amount of usage data that AmI environments will make available, through monitoring, for further analysis and improvement.

This endeavor will be highly multidisciplinary, involving collaboration across multiple domains and building upon several disciplines, including HCI, social sciences, requirements engineering, software quality, human factors and usability engineering, and software engineering. Therefore, it is critical to bring together research teams and diverse user groups so as to start a constructive dialogue and establish a common

vocabulary. The direct and active participation of user representatives in shaping ambient intelligence technologies and applications to reflect and anticipate their needs is also considered a critical factor. Therefore, appropriate research infrastructures are needed to act as test beds and incubators of future technologies.

Towards this end, the Institute of Computer Science (ICS) of FORTH is in the process of creating a large-scale, state-of-the-art AmI facility intended, among other things, to support the establishment and conduct of a line of research targeted to HCI in AmI environments and technologies. This research facility will initially address the application domains of housing, education, work, health, entertainment, commerce, culture, and agriculture. The facility will also encourage international collaboration through hosting visiting scientists from around the world.

It is believed that such a research facility will have a significant role in ensuring that AmI emerges and develops in a way that is acceptable and can be adopted in the long run by all members of the information society as well as facilitating and driving a smooth transition of AmI technologies from research into real-life situations.

## 6 SUMMARY AND CONCLUSIONS

This chapter has discussed the centrality and role of human factors in the emergence and development of AmI environments, focusing on:

- The user-centered design process and how it is affected by the complexity of AmI environments
- Basic user experience qualities which need to inform the design of AmI environments but also be captured and modeled so as to enhance interaction, responsiveness, and intelligent behavior of the environment

To illustrate the above, Section 1 discussed the UCD process in light of the requirements posed by AmI, focusing on emerging problems and potential solutions for applying and revising existing methods and techniques or developing new ones. Overall, the UCD process as practiced today appears to be a more than valid starting point for investigating how to put users at the center of AmI development. Various well-known methods and techniques have been shown to be useful in this respect. However, UCD in AmI needs to face the challenges and exploit the opportunities posed by the extended context of use and its inextricable fusion with the interactive environment. A very important aspect in this respect is the availability of monitoring data over extended period of times, which can be exploited when adapting interaction and environmental behavior on the fly but also in continuously reshaping design as well as proposing new methods and techniques for the various UCD activities.

An additional element to take into account is the fusion of technology with the human living space, which brings about the requirements of combining

interaction design with industrial and architectural design. Clearly, as it happened decades ago with the emergence of HCI, an opportunity exists here for the foundation of a new design discipline deeply rooted in human factors but characterized by its own processes, methods, rules, and practices.

Section 2 focused user experience factors considered critical in AmI, including natural interaction, accessibility, cognitive demands, emotions, health, safety and privacy, social aspects, cultural aspects, and aesthetics. For each of them, a brief overview of the main issues involved has been provided, focusing on existing or emerging approaches. Obviously, the list is not complete, and existing approaches far from offer a comprehensive framework.

Section 3 presented three case studies of the UCD of AmI artifacts developed in the ICS-FORTH AmI Programme, namely an interactive wall for the display of museum artifacts, an augmented school desk, and a smart coffee table. Obviously, the scope of these case studies is limited. However, each illustrated in practical terms some aspects of the adopted design process and of the user experience qualities relevant to the specific projects.

Finally, Section 4 put forward the need for a more systematic approach to the above issues. To this end, the availability of appropriate research infrastructures and multidisciplinary collaborations is critical.

## ACKNOWLEDGMENT

## REFERENCES

Aarts, E., and de Ruyter, B. (2009), "New Research Perspectives on Ambient Intelligence," *Journal of Ambient Intelligence and Smart Environments*, Vol. 1, No. 1, pp. 5–14.

Adams, R., and Russell, C. (2007), "Lessons from Ambient Intelligence Prototypes for Universal Access and the User Experience," in *Proceedings of the 9th Conference on User Interfaces for All*, Königswinter, Germany, September 27–28, 2006, C. Stephanidis and M. Pieper, Eds., Lecture Notes in Computer Science, Springer-Verlag, Berlin and Heidelberg, pp. 229–243.

Aghajan, H., Augusto, J., and Delgado, R. (2009), *Human-Centric Interfaces for Ambient Intelligence*, Academic Press, Burlington, MA.

Antona, M., Burzagli, L., Emiliani, P.-L., and Stephanidis, C. (2007), "Ambient Intelligence and Implications for People with Disabilities" in *Towards an Inclusive Future: Impact and Wider Potential of Information and Communication Technologies*, P. R. W. Roe, Ed., COST219ter, Brussels, pp. 158–187.

Antona, M., Margetis, G., Ntoa, S., Leonidis, A., Korozi, M., Paparoulis, G., and Stephanidis, C. (2010), "Ambient Intelligence in the Classroom: An Augmented School Desk," in Proceedings of the 2010 AHFE International Conference (3rd International Conference on Applied Human Factors and Ergonomics), W. Karwowski and

G. Salvendy. Eds., Miami, FL, July 17–20, CRC Press, Boca Raton, FL.

Bandini, S., Bonomi, A., and Vizzari, G. (2009), "Simulation Supporting the Design of Self-Organizing Ambient Intelligent Systems," in *Proceedings of the 2009 ACM Symposium on Applied Computing (SAC)*, Honolulu, Hawaii, March 9–12, 2009, Sung Y. Shin and Sascha Ossowski, Eds., ACM, New York, pp. 2082–2086.

Bick, M., Kummery, T. F., and Maletzkyz, M. (2009), "Towards a Research Agenda on Cultural Influences on the Acceptance of Ambient Intelligence in Medical Environments," in *Proceedings of the Fifteenth Americas Conference on Information Systems*, San Francisco, CA, August 6–9, 2009.

Biocca, F., Harms, C., and Burgoon, J. K. (2003), "Toward a More Robust Theory and Measure of Social Presence; Review and Suggested Criteria," *Presence*, Vol. 12, No. 5, pp. 456–480.

Blomberg, J., Giacomi, J., Mosher, A., and Swenton-Wall, P. (1993), "Ethnographic Field Methods and Their Relation to Design," in *Participatory Design: Principles and Practices*, D. Schuler and A. Namioka, Eds., Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 123–155.

Boehner, K., Sengers, P., and Gay, G. (2005), "Affective Presence In Museums: Ambient Systems for Creative Expression," *Digital Creativity*, Vol. 16, No. 2, pp. 79–89.

Bohn, J., Coroamà, V., Langheinrich, M., Mattern, F., and Rohs, M. (2005), "Social, Economic, and Ethical Implications of Ambient Intelligence and Ubiquitous Computing," in *Ambient Intelligence*, W. Weber, J. M. Rabaey, and E. H. L. Aarts, Eds., Springer, Berlin, pp. 5–28.

Bosse, T., Callaghan, V., and Lukowicz, P. (2010), "On Computational Modeling of Human-Oriented Knowledge in Ambient Intelligence," *Journal of Ambient Intelligence and Smart Environments*, Vol. 2, pp. 3–4.

Butz, A. (2010), "User Interfaces and HCI for Ambient Intelligence and Smart Environments," in *Handbook of Ambient Intelligence and Smart Environments*, H. Nakashima, H. Aghajan, and J. C. Augusto, Eds., Springer, Berlin, pp. 535–558.

Carbonell, N. (2006), "Ambient Multimodality: Towards Advancing Computer Accessibility and Assisted Living," *Universal Access in the Information Society*, Vol. 5, No. 1, pp. 96–104.

Coroama, V., Bohn, J., Mattern, F., (2004), "Living in a Smart Environment—Implications for the Coming Ubiquitous Information Society," in *Proceedings of the International Conference on Systems, Man and Cybernetics 2004*, October 10–13, 2004, IEEE SMC 2004, Vol. 6, The Hague, The Netherlands, pp. 5633–5638.

Coughlin, J. F., Reimer, B., and Mehler, B. (2009), "Driver Wellness, Safety & the Development of an AwareCar," MIT AgeLab White Paper, available: http://agelab.mit.edu/system/files/file/Driver_Wellness.pdf.

Czaia, S. J., and Nair, S. N. (2006), "Human Factors Engineering and Systems Design," in *Handbook of Human Factors and Ergonomics*, 3rd ed., G. Salvendy, Ed., Wiley, New York, pp. 32–49.

Edwards, W. K., and Grinter, R. E. (2001), "At Home with Ubiquitous Computing: Seven Challenges," in *Proceedings of the 3rd International Conference on Ubiquitous Computing*, Atlanta, GA, September 30–October 02, 2001, Lecture Notes in Computer Science, Vol. 2201. Springer-Verlag, London, pp. 256–272.

Emiliani, P. L., and Stephanidis, C. (2005), "Universal Access to Ambient Intelligence Environments: Opportunities and

Challenges for People with Disabilities," *IBM Systems Journal*, Special Issue on Accessibility, Vol. 44, No. 3, pp. 605–619.

Eyles, M., and Eglin, R. (2007), "Entering an Age of Playfulness Where Persistent, Pervasive Ambient Games Create Moods and Modify Behaviour," in *Proceedings of the 3rd International Conference on Games Research and Development* (*Cybergames '07*), Manchester, United Kingdom, September 2007.

Faconti, G., and Massink, M. (2001), "Continuous Interaction with Computers: Issues and Requirements," in *Universal Access in HCI: Towards an Information Society for All*, Vol. 3 of *Proceedings of HCI International 2001*, New Orleans, LA, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 301–304.

Ferscha, A., Resmerita, S., and Holzmann, C. (2007), "Human Computer Confluence," in *Universal Access in Ambient Intelligence Environments, Proceedings of the 9th ERCIM Workshop on User Interfaces for All*, C. Stephanidis and M. Pieper, Eds., Königswinter, Germany, September 27–28, 2006, pp. 14–27.

Friedewald, M., Vildjiounaite, E., Punie, Y., and Wright, D. (2007, February), "Privacy, Identity and Security in Ambient Intelligence: A Scenario Analysis," *Telematics and Informatics archive*, Vol. 24, No. 1, pp. 15–29.

Gaggioli, A. (2005), "Optimal Experience in Ambient Intelligence," in *Ambient Intelligence*, G. Riva, F. Vatalaro, F. Davide, and M. Alcañiz, Eds., IOS Press, Amsterdam, The Netherlands, pp 35–43.

Gepner, D., Simonin, J., and Carbonell, N. (2007), "Gaze as a Supplementary Modality for Interacting with Ambient Intelligence Environments," in *Proceedings of 12th International Conference on Human-Computer Interaction* (*HCI International 2007*), Beijing, China, July 22–27, 2007, Lecture Notes in Computer Science Series, Springer.

Gould, J. D. (1988), "How to Design Usable Systems," in *Human–Computer Interaction*, M. Hellender, Ed., North-Holland, Amsterdam, The Netherlands, pp. 35–41.

Grammenos, D., Georgalis, Y., Kazepis, N., Drossis, G., and Ftylitakis, N. (2010), "The bootTable Experience: Iterative Design and Prototyping of an Alternative Interactive Tabletop," in *Proceedings of ACM DIS 2010*, Aarhus, Denmark, August 16–20, 2010, K. Halskov and M. Graves Petersen, Eds., ACM Press, New York, pp. 272–281.

Hassenzahl, M., and Tractinsky, N. (2006), "User Experience—A Research Agenda," *Behaviour & Information Technology*, Vol. 25, No. 2, pp. 91–97.

Herbon, A., Oehme, A., and Zentsch, E. (2006), "Emotions in Ambient Intelligence—An Experiment on How to Measure Affective States," paper presented at the Workshop on the Role of Emotion in Human-Computer Interaction, held in the frame of the 20th British HCI Group Annual Conference, University of London, September 12, 2006, available: http://www.zmms.tu-berlin.de/prometei/download/publikationen/ahe/Herbon_etal_2006.pdf.

Hellenschmidt, M., and Wichert, R. (2005), "Goal-Oriented Assistance in Ambient Intelligence," Technical Report, Fraunhofer-IGD, Darmstadt, available: http://www.igd.fhg.de/igd-a1/publications/publ/ERAmI_2005.pdf.

Hewett, T. T., and Meadow, C. T. (1986), "On Designing for Usability: An Application of Four Key Principles," in *CHI '86 Proceedings*, pp. 247–251.

Information Society Technologies Advisory Group (ISTAG) (2003), "Ambient Intelligence: From Vision to Reality," available: ftp://ftp.cordis.lu/pub/ist/docs/istag-ist2003_consolidated_report.pdf.

Intille, S. S., Larson, K., Munguia Tapia, E., Beaudin, J., Kaushik, P., Nawyn, J., and Rockinson, R. (2006), "Using a Live-In Laboratory for Ubiquitous Computing Research," in *Proceedings of PERVASIVE 2006*, K. P. Fishkin, B. Schiele, P. Nixon, and A. Quigley, Eds., Lecture Notes in Computer Science, Vol. 3968, Springer-Verlag, Berlin and Heidelberg, pp. 349–365.

International Organization for Standardization (ISO) (1999), "Human-Centred Design Processes for Interactive Systems," ISO 13407:1999, ISO, Geneva.

International Organization for Standardization (ISO) (2010), "Ergonomics of Human-System Interaction—Part 210: Human-Centred Design for Interactive Systems," ISO 9241-210:2010, ISO, Geneva.

Kartakis, S., and Stephanidis, C. (2010), "A Design-and-Play Approach to Accessible User Interface Development in Ambient Intelligence Environments," *Computers in Industry*, special issue on "Human-Centered Computing Systems in Industry," Vol. 61, No. 4, pp. 318–328.

Kemppainen, E., Abascal, J., Allen, B., Delaitre, S., Giovannini, C., and Soede, M. (2007), "Ethical and Legislative Issues with Regard to Ambient Intelligence," in *Impact and Wider Potential of Information and Communication Technologies*, P. R. W. Roe, Ed., COST, Brussels, pp. 188–205.

Kraemer, S., and Carayon, P. (2007), "Human Errors and Violations in Computer and Information Security: The Viewpoint of Network Administrators and Security Specialists," *Applied Ergonomics*, Vol. 38, No. 2, pp. 143–154.

Kröse, B., Porta, J., van Breemen, A., Crucq, K., Nuttin, M., and Demeester, E. (2003), "Lino, the User-Interface Robot," in *Ambient Intelligence*, G. Goos, J. Hartmanis, and J. van Leeuwen, Eds., Lecture Notes in Computer Science, Springer, Berlin, pp. 264–274.

Lieberman, H., Ed. (2001), *Your Wish Is my Command*, Morgan Kaufmann, San Francisco, CA.

Lopez-Cozar, R., and Callejas, Z. (2010), "Multimodal Dialogue for Ambient Intelligence and Smart Environments," in *Handbook of Ambient Intelligence and Smart Environments*, H. Nakashima, H. Aghajan, and J. C. Augusto, Eds., Springer, Berlin, pp. 559–579.

Lo Presti, S., Butler, M., Leuschel, M., and Booth, C. (2005), "A Trust Analysis Methodology for Pervasive Computing Systems," in *Trusting Agents for Trusting Electronic Societies*, R. Falcone, S. Barber, J. Sabater-Mir, and M. P. Singh, Eds., Lecture Notes in Computer Science. Springer, Berlin, pp. 129–143.

Luyten, K., Vandervelpen, C., and Coninx, K. (2005), "Task Modeling for Ambient Intelligent Environments: Design Support for Situated Task Executions," in *Proceedings of the 4th international Workshop on Task Models and Diagrams*, Gdansk, Poland, September 26–27, 2005, TAMODIA '05, Vol. 127, Association for Computing Machinery, New York, pp. 87–94.

Maguire, M. (2001), "Methods to Support Human-Centred Design," *International Journal of Human-Computer Studies*, Vol. 55, pp. 587–634.

Mandryk, R. L., Inkpen, K. M., and Calvert, T. W. (2006), "Using Psychophysiological Techniques to Measure User Experience with Entertainment Technologies," *Behaviour & Information Technology*, Vol. 25, No. 2, pp. 141–158.

Melder, W. A., Truong, K. P., Uyl, M. D., Van Leeuwen, D. A., Neerincx, M. A., Loos, L. R., and Plum, B. S. (2006), "Affective Multimodal Mirror: Sensing and Eliciting Laughter," in *Proceedings of the International Workshop on Human-Centered Multimedia*, Augsburg, Bavaria, Germany, September 28–28, 2007, HCM '07, Association for Computing Machinery, New York, pp. 31–40.

Michel, D., Argyros, A. A., Grammenos, D., Zabulis, X., and Sarmis, T. (2009), "Building a Multi-Touch Display Based on Computer Vision Techniques," in *Proceedings of the IAPR Conference on Machine Vision and Applications (MVA'09)*, Hiyoshi Campus, Keio University, Japan, May 20–22, 2009, pp. 74–77.

Nakashima, H., Aghajan, H., and Augusto, J. C. (2010), "Ambient Intelligence and Smart Environments: A State of the Art," in *Handbook of Ambient Intelligence and Smart Environments*, H. Nakashima, H. Aghajan, and J. C. Augusto, Eds., Springer, Berlin, pp. 3–31.

Nijholt, A. Stock, O., and Nishida, T. (2009), "Social Intelligence Design in Ambient Intelligence. AI and Society," *Journal of Human-Centred Systems*, Vol. 24, No. 1, pp. 1–3.

Norman, D., A., and Draper, S. W., Eds. (1986), *User Centered System Design: New Perspectives on Human-Computer Interaction*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Ortiz, A., del Puy Carretero, M., Oyarzun, D., Yanguas, J. J., Buiza, C., Gonzalez, M. F., and Etxeberria, I. (2007), "Elderly Users in Ambient Intelligence: Does an Avatar Improve the Interaction?" in *? ERCIM UI4ALL Ws 2006*, C. Stephanidis and M. Pieper, Eds., Lecture Notes in Computer Science, Vol. 4397, Springer, Berlin, pp. 99–114.

Rasmussen, J., Pejtersen, A. M., and Goodstein, L. P. (1994), *Cognitive Systems Engineering*, Wiley, New York.

Redström, J. (2007), "Aesthetic Concerns in Pervasive Information Systems," in *Pervasive Information Systems*, G. Giaglis and P. Kourouthanassis, Eds., M. E. Sharpe, Armonk, NY, pp. 197–209.

Richter, K., and Hellenschmidt, M. (2004), "Interacting with the Ambience: Multimodal Interaction and Ambient Intelligence," paper presented at the W3C Workshop on Multimodal Interaction, Sophia Antipolis, France, July 2004, available: http://www.w3.org/2004/02/mmi-workshop/richter.pdf.

Röcker, C., Janse, M. D., Portolan, N., and Streitz, N. (2005), "User Requirements for Intelligent Home Environments: A Scenario-Driven Approach and Empirical Cross-Cultural Study," in *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient intelligence: Innovative Context-Aware Services: Usages and Technologies*, Grenoble, France, October 12–14, 2005, Gerard Bailly, James L. Crowley and Gilles Privat, Eds., Association for Computing Machinery, New York, pp. 111–116.

Roe, P. R. W., Ed. (2007), *Towards an Inclusive Future: Impact and Wider Potential of Information and Communication Technologies*, COST219ter, Brussels.

Savidis, A., and Karouzaki, E. (2009), "Artifical Game Presenter Avatars," paper presented at the ACM Conference on Advances in Computer Entertainment Technology, Athens, Greece, October 29–31, Association for Computing Machinery Digital Library, New York.

Savidis, A., and Lilis, Y. (2008), "Adaptable Pluggable Multimodal Input with Extensible Accessible Soft Dialogues for Games," paper presented at the ACM Conference on Advances in Computer Entertainment Technology, Yokohama, Japan, December 3-5, Association for Computing Machinery, New York, pp. 155–158.

Savidis, A., and Stephanidis, C. (2005), "Distributed Interface Bits: Dynamic Dialogue Composition from Ambient Computing Resources," *Personal and Ubiquitous Computing Journal*, Vol. 9, No. 3, pp. 142–168.

Schmidt, A. (2005), "Interactive Context-Aware Systems Interacting with Ambient Intelligence," in *Ambient Intelligence*, G. Riva, F. Vatalaro, F. Davide, and M. Alcañiz, Eds., IOS Press, 159–178, available: http://www.ambientintelligence.org.

Schmidt, A., Terrenghi, L., and Holleis, P. (2007), "Methods and Guidelines for the Design and Development of Domestic Ubiquitous Computing Applications," *Pervasive and Mobile Computing*, Vol. 3, No. 6, pp. 721–738.

Søraker, J. H., and Brey, P. (2007), "Ambient Intelligence and Problems with Inferring Desires from Behaviour," *International Review of Information Ethics*, Vol. 8, No. 1, pp. 7–12.

Stephanidis, C. (2001), "Human-Computer Interaction in the Age of the Disappearing Computer," in *Advances in Human-Computer Interaction I, Proceedings of the Panhellenic Conference with International Participation on Human-Computer Interaction (PC-HCI 2001)*, N. Avouris and N. Fakotakis, Eds., Patras, Greece, December 7–9, Typorama Publications, Patras, Greece, pp. 15–22.

Streitz, N. A. (2007), "From Human–Computer Interaction to Human–Environment Interaction: Ambient Intelligence and the Disappearing Computer," in *Universal Access in Ambient Intelligence Environments, Proceedings of the 9th ERCIM Workshop on User Interfaces for All*, Königswinter, Germany, September 27–28, 2006, C. Stephanidis and M. Pieper, Eds., Springer, Berlin, pp. 3–13.

Theofanos, M., and Scholtz, J. (2004), "Towards a Framework for Evaluation of Ubicomp Applications," *IEEE Pervasive Computing*, Vol. 3, No. 2, pp. 82–88.

Vastenburg, M. J., and Romero Herrera, N. (2010), "Adaptive Experience Sampling: Addressing the Dynamic Nature of In-Situ User Studies," in *ISAmI 2010*, J. C. Augusto et al., Eds., AISCS 72, Springer, Berlin, pp. 197–200.

Verpoorten, K., Luyten, K., and Coninx, K. (2007), "Task-Based Prediction of Interaction Patterns for Ambient Intelligence Environments," in *Human-Computer Interaction: Interaction Design and Usability*, J. Jacko, Ed., Springer, Berlin, pp. 1216–1225.

Vredenburg, K., Isensee, S., and Righi, C. (2001), *User-Centered Design: An Integrated Approach*, Software quality institute series, Prentice Hall, Upper Saddle River, NJ.

Westerink, J., Ouwerkerk, M., Overbeek, T. J. M., Pasveer, F., and de Ruyter, B. (2008), *Probing Experiences: From Assessment of User Emotions and Behaviour to Development of Products*, Philips Research Book Series, Springer, Berlin.

Wright, D., Gutwirth, S., Friedewald, M., Vildjiounaite, E., and Punie, Y., Eds. (2010) *Safeguards in a World of Ambient Intelligence: The International Library of Ethics, Law and Technology*, Vol. 1. Springer, Berlin.

Zabulis, X., Grammenos, D., Sarmis, T., Tzevanidis, K., and Argyros, A. A. (2010), "Exploration of Large-Scale Museum Artifacts Through Non-Instrumented, Location-Based, Multi-User Interaction," in *Proceedings of the*

*11th VAST International Symposium on Virtual Reality, Archaeology and Cultural Heritage, VAST'2010*, Palais du Louvre, Paris, France, 21–24 September 2010.

Zhang, B., Rau, P. L. P., and Salvendy, G. (2010), "Design and Evaluation of Smart Home User Interface: Effects of Age, Tasks and Intelligence Level," *Behaviour and Information Technology*, Vol. 28, No. 2, pp. 239–249.

Zhou, J., Yu, C., Riekki, J., and Kärkkäinen, E. (2007), "AmE Framework: A Model for Emotion-Aware Ambient Intelligence," paper presented at the Second International Conference on Affective Computing and Intelligent Interaction (ACII2007): Doctoral Consortium, Lisbon, Portugal, September 12–14, 2007, pp. 163–169.

Zimmermann, G., Vanderheiden, G., Ma, M., Gandy, M., Trewin, S., Laskowski, S., and Walker, M. (2004), "Universal Remote Console Standard: Toward Natural User Interaction in Ambient Intelligence," in *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, Vienna, Austria, April 24–29, 2004, Association for Computing Machinery, New York, pp. 1608–1609.

# CHAPTER 50

# INTERACTIVITY: EVOLUTION AND EMERGING TRENDS

**Constantine Stephanidis**
**Institute of Computer Science Foundation for Research and Technology—Hellas (FORTH)**
**Heraklion, Crete, Greece**
**Department of Computer Science**
**University of Crete**
**Crete, Greece**

**Vassilis Kouroumalis and Margherita Antona**
**Institute of Computer Science Foundation for Research and Technology—Hellas (FORTH)**
**Heraklion, Crete, Greece**

## 1 INTRODUCTION

In its short history, human–computer interaction (HCI) has been characterized by a trend toward elaborating, designing, and establishing more human-oriented, natural, and intelligent forms of interaction, progressively addressing the needs and requirements of a wider, less experienced, and more naive user basis (Stephanidis et al., 1998). This path is intrinsically linked with (i) the progressive emergence of new, more general and systematic framework for studying and designing interaction and (ii) technological evolution supporting the establishment of richer alternative interaction techniques and user interface styles.

The concept of interactivity between humans and computers, contextually defined in this chapter as *the extent to which the characteristics of an interactive system affect the communication behavior of both the user(s) and the system itself*, plays a crucial role in this respect. Interactivity is not a new concept, as it has been investigated in the literature with respect to both human communication and various types of interactive systems. In this chapter, the evolution of interactivity as it has manifested itself in HCI is considered.

Although it is difficult to identify the forces that shape the evolution of interactivity, two main mechanisms can be considered, on the one hand technological advancements and breakthroughs offering new possibilities, and on the other hand the social impact of these technologies, starting from specializsed user communities and gradually expanding to the society at large. Such impact could take many forms, ranging from the commercial success and user adoption of graphical operating systems to the shaping of the new information society in the age of the Internet.

To shed light on interactiviy in the context of HCI, understand its evolution, and outline emerging challenges, this chapter firstly briefly reviews existing accounts of interactivity, focusing on identifying dimensions of the concept which can be meaningfully used for analyzing and explaining its various instantiations in HCI.

Second, the chapter looks at different theoretical frameworks of design and evaluation that influenced the evolution of interactivity.

Third, it looks into the evolution of interaction, starting from the 1950s until today, by concentrating on technological and research advancements that have led to the current interaction models and styles, as well as at application domains that fostered significant steps in the evolutionary path of interactivity. A brief outline of the most important interaction paradigms is presented, outlining the interactivity dimensions addressed in each of them.

Finally, it attempts to present the latest developments and emerging trends in HCI and address several issues that concern them.

## 2 CONCEPT OF INTERACTIVITY

### 2.1 Interactivity in Human Communication

The concept of interactivity originates in the context of human communication and has been addressed in various related disciplines, such as philosophy of language, linguistics, semiotics, and communication psychology. While a review of such accounts is beyond the scope of this chapter, some basic aspects of interactivity in human communication naturally lend themselves to provide terms of comparison in analyzing interactivity in HCI.

In the famous work on communication theory by Shannon and Weaver (1949), communication has been defined as a process whereby information is enclosed in a package and is channeled and imparted by a sender to a receiver via some medium. The receiver then decodes the message and provides feedback to the sender. All forms of communication require a sender, a message, and an intended recipient. However, the receiver does not need to be present or aware of the sender's intent to communicate at the time of communication in order for the act of communication to occur. Communication also requires that all parties share a common code or language for message exchange.

Interactive communication is commonly defined as a process involving at least two participants, where the content of a particular message is determined in part by the content of the prior messages from all participants, that is, by the communication context (Chapanis, 1988). Interactive communication can take place through a symbolic system, notably natural language in spoken or written form, and complementary through gestures, facial expressions, and actions. Natural language is a unique symbolic system. Some of the most important distinguishing characteristics of human language are (Hockett, 1960):

- *Vocal–Auditory Channel*. Standard human language occurs as vocal communication (i.e., producing sounds with the mouth), which is perceived by hearing it. Exceptions are writing and sign language, which are examples of communication in the visual and manual channels, respectively.

- *Rapid Fading* (*Transitoriness*). The human language signal does not persist over time. Speech waveforms fade rapidly and cannot be heard after they fade. Writing and audio-recordings can be used to record human language, so that it can be re-created at a later time.

- *Interchangeability*. The speaker can both receive and broadcast the same signal. This is distinctive from some forms of animal communications.

- *Total Feedback*. Speakers can hear themselves speak and can monitor their language performance. This differs from some other simple communication systems, such as traffic signals, which are not normally capable of monitoring their own functions.

- *Semanticity*. This is a fundamental aspect of all communication systems, implying that specific signals can be matched with specific meanings. Speakers of a language recognize the meaning to which a signal is associated.

- *Arbitrariness*. There is no necessary connection between the form of the signal and its meaning.

- *Discreteness*. The basic units of speech (such as sounds) can be categorized as belonging to distinct categories. There is no gradual, continuous shading from one sound to another in the linguistics system, although there may be a continuum in the real physical world.

- *Displacement*. The speaker can talk about things which are not present, either spatially or temporally. For example, human language allows speakers to talk about the past and the future as well as the present. Speakers can also talk about things that are physically distant (e.g., other countries, the moon) or even refer to things and events that do not actually exist.

- *Productivity*. Human languages allow speakers to create novel, never-before-heard utterances that others can understand. Human beings are unrestricted in what they can talk about, and no area of experience is accepted as necessarily incommunicable. This includes language and communication themselves. Thus human language allows metalinguistic discourse.

- *Learning*. Human language is not something inborn. Although humans are probably born with an ability to do language, they must learn, or acquire, their native language from other speakers. This is different from many animal communication systems where the animal is born knowing their entire system, for example, bees.

Another inherent characteristic of natural language, which distinguishes it from formal languages, such as programming languages and command languages, is underspecification of meaning, which may take two forms, namely ambiguity and vagueness. Ambiguity refers to the fact that natural language words and utterances may be interpreted in different ways depending on context. Vagueness refers to the fact that natural language may refer to events and entities at an

abstract level, omitting details that are not relevant in a specific context. Underspecification is often mentioned as a "defect" of natural language, which constitutes an obstacle to precise communication. On the other hand, it can be seen as an economy mechanism which allows human communication to be specific enough for a particular purpose with the minimum necessary effort (Wasow et al., 2005).

Apart from the characteristics of language, human spoken dialogue can be analyzed along a number of dimensions which appear to be relevant also in the wider context of HCI. These include (Petukhova and Bunt, 2009):

- *Dialogue Purpose and Domain of Discourse.* Dialogues are usually motivated by goals, tasks, or activities which are noncommunicative in nature, for example, to obtain certain information, to solve a problem, and to act in a game.
- *Contact, Presence, and Attention,* A basic requirement of communication is that the parties are in contact and stay so. For some types of dialogue this aspect is of a particular importance, namely when there is no or limited visual contact between the participants. For example, telephone conversations are dependent on the quality of the communication channel. But also, when dialogue participants have direct visual contact, they tend to permanently check the attention of their interlocutors and their readiness to continue the conversation. To this purpose, they utilize both their bodies and facial expressions (e.g., gaze) and a variety of vocal phenomena to show attention as well as the type of reaction they expect from others.
- *Grounding and Feedback.* Successful dialogue is based on shared knowledge and beliefs (Clark, 1996). To establish such a common communication basis, speakers and addressees during dialogue attempt to confirm that each of them has understood what is uttered. This process is called grounding. Grounding includes feedback (Traum, 1999), that is, the speaker during dialogue provides information on his or her own processing of the partner's previous utterance(s).
- *Taking Turns.* Turn management, another essential aspect of any interactive conversation, is defined as the distribution of the right to occupy the sender's role in dialogue. Turn taking is usually understood as obeying normative rules, depending on the speaker's needs or motivations and beliefs, and on the rights and obligations in a conversational situation.
- *Social Obligations and Politeness.* Participating in a dialogue is a social activity, where one is supposed to do certain things and not others and to act in accordance with the norms and conventions regulating social behavior. Each participant in dialogue not only has functional but also ethical tasks and obligations and performs social obligation acts to fulfill these. Such obligations

include politeness rules, such as not imposing anything to the communication partner, offering alternative options, and encouraging positive feelings (Lakoff, 1973).
- *Dialogue Structure.* Dialogue participants may at several dialogue stages indicate their view of the state of the dialogue and make the hearer acquainted with his plans for the continuation of the conversation. The speaker can give indications that he or she is going to close the discussion of certain topic(s) or wants to concentrate the hearer's attention on a new topic. Dialogue structuring is based on the speaker's view of the present linguistic context, on his or her plan for continuing the dialogue, and on the assumed need to structure the discourse for a partner.
- *Handling Errors.* Speakers continuously monitor the utterance that is currently being produced or prepared to produce (Clark and Krych, 2004), and when problems or mistakes are discovered, they stop the flow of the speech and signal to the addressee that there is trouble and that a repair follows (error signaling). Human conversations contain large numbers of phenomena such as disfluencies, interruptions, confirmations, anaphora, and ellipses (Glass, 1999).
- *Timing.* Another aspect of communication which is concerned with disfluent speech production is time management, where the speaker suspends the dialogue for one of several reasons and resumes it after minor (stalling) or prolonged (pause) delay. Delays take place at all major levels of planning—from retrieving a word to deciding what to talk about next (Clark and Fox Tree, 2002).
- *Adaptation.* One of the most robust findings of studies of human–human dialogue is that people adapt their interactions to match their conversational partners' needs and behaviors (Pennebaker and King, 1999). People adapt the content, the syntactic structures of their utterances, as well as their lexical choices to match their partners'. They also adapt their speaking rate, amplitude, and clarity of pronunciation (Walker at al., 2007). Adaptation is also a crucial aspect of intercultural communication, that is, people adjust their communication styles toward or away from each other during cross-cultural interactions (Cai and Rodriguez, 1996).

Besides speech dialogue, other aspects of human communication are also important, For example, recent phenomenological views on language and communication emphasize action associated with speech (Tripathi, 2005). These actions into which language is woven are inseparable from communicative meaning. Thus, language has an extra dimension associated with social conventions and actions, such as gestures, pointing, and body language. Human beings have the ability to utilize their entire bodies for the purpose of communication (rather than simply voice or writing), thus implying

multimodality. The tone of the voice, body language, and gaze all constitute communicative meaning, either consciously or subconsciously (Bunt and Beun, 1998).

Human communication is also supported through semiotic systems other than natural language, namely iconic languages. Icons are semiotic signs which directly resemble the objects they refer to. In contrast to natural language, iconic languages are not arbitray. Because of their communicative power, which transcends different languages and cultures, icons are used in a variety of real-life situations to inform people about particular conditions or give instructions. Typical examples appear in public information spaces, trains, airplanes, cars, and printed books (Barker, 2000). The human ability to communicate through action and iconic languages is at the center of the notion of direct manipulation (see Section 4.3).

Finally, emotion plays a central role in human communication, especially when disagreement between participants emerges. Emotional reactions represent an important type of feedback on the effects of utterances on dialogue participants. In dialogue, emotional reactions can be signaled by response speed, reiteration of claims, lexical choice, response avoidance, sentence length, and so on. Likely emotional reactions are defensiveness, indignation, frustration, anger, regret, guilt, and enthusiasm (Anderson and Guerrero, 1998).

## 2.2 Interactivity in HCI

Many definitions of interactivity have been provided in the HCI literature, especially with reference to Web services, computer-supported communication, computer-supported work, electronic advertising, e-learning, interactive TV, electronic games, and virtual reality.

User–machine interaction was the focus of early definitions of interactivity, in which the emphasis was on human interaction with computers. To be interactive, a computer system must be responsive to users' actions. In this context, interactivity has generally been measured in terms of input or output devices, for instance, the number of "point-and-click" opportunities on a computer screen (Shneiderman, 1998). Norman (1990) suggested that the interactive process is a repeated loop of decision sequences of a user's action and the environment's reaction.

However, though user–machine interaction is an important aspect of interactivity, it is not adequate to fully capture the concept, especially since the emergence of more advanced technology such as the Internet. As a result, researchers have started investigating interaction in a technological context more broadly, also considering other types of interaction, such as user–user interaction and user–message interaction.

User–user interaction is usually discussed from an interpersonal communication perspective. In this respect, the more communication in a computer-mediated environment resembles human interpersonal communication, the more interactive such an environment is considered (Ha and James, 1998). However, a medium such as the Internet offers many possibilities to break the boundaries of traditional interpersonal communication. Through the Internet, people no longer

need to be at the same place or communicate at the same time. Research has shown that computer-mediated communication and face-to-face communication are not functional alternatives (Flaherty et al., 1998), as each has distinctive characteristics and addresses different needs.

From a user–message interaction perspective, interactivity is defined as the ability of the user to control and modify messages (Steuer, 1992). Whereas people have little control over messages in traditional media, the Internet gives users much more freedom in controlling the messages they receive and allows messages to be customized according to the users' own needs. Based on the above, Liu and Shrum (2002) define interactivity as the degree to which two or more communication parties can act on each other, on the communication medium, and on the messages and the degree to which such influences are synchronized.

Other definitions have focused on two distinct aspects of interactivity: reciprocal communication and control (Liu, 2003). Reciprocity implies that interaction should allow two-way flow of information, and the information being exchanged in a sequence should closely relate to each other (Rafaeli and Sudweeks, 1997). Additionally, the exchange of information should happen in real time. The control dimension implies that participants should be able to exert control on both sent and received information (Jensen, 1999) as well as over the communication medium. Both control and reciprocal communication are important aspects of interactivity. Control helps ensure a reciprocal exchange that satisfies the needs of all communicating parties, while reciprocal communication provides an effective channel for exerting control. Melding the two aspects, Liu and Shrum (2002) proposed three dimensions of interactivity: active control, which describes a user's ability to voluntarily participate in and instrumentally influence a communication; two-way communication, which captures the bidirectional flow of information; and synchronicity, which corresponds to the speed of the interaction. Based on the above, Liu (2003) defines a framework for measuring interactivity on websites.

Interactivity has been discussed also in relation to new media and educational technologies. Rice (1984) defined "new media" as consisting of communication technologies that allow or facilitate interactivity among users or between users and information. Heeter (1989) describes six dimensions of interactivity in new media: (i) complexity of available choice, meaning the amount and variety of user choices; (ii) the effort that any user of a media system must exert to access information; (iii) responsiveness: interactivity is a continuous variable measuring how "actively responsive a medium is to users"; (iv) information use monitoring, that is, how well information selection can be monitored across a population of users; (v) ease of adding information, meaning the degree to which users can add information for access by the audience; and (vi) interpersonal communication facilitation, which comes in at least two forms: asynchronous (allowing users to respond to messages at their convenience) and synchronous (allowing for concurrent participation).

Downes and McMillan (2000), following an interview-based study, propose a conceptual definition of interactivity identifying six main dimensions: direction of communication, time flexibility, sense of place, level of control, responsiveness, and perceived purpose of communication.

In the domain of e-learning, Chou (2003) investigates a technical framework for interactivity following an empirical methodology. The framework includes the following dimensions: (i) choice, that is, the amount and multimedia type of information users (learners and instructors) have access to as well as other types of user options; (ii) nonsequential access, that is, users can access information in a nonlinear way; (iii) responsiveness to learner, tha is, the system responds to a user's request without delay; (iv) monitoring information use, that is, the system can collect data on the users themselves, their selections, their use of information, and so on, and the users can monitor personal information which is collected; (v) personal-choice helper, that is, information is available to help learners choose instructional content; (vi) adaptability, that is, the interaction process and the exchange of information are adapted to individuals; (vii) playfulness, that is, information stimulates users' curiosity and fun; (viii) facilitation of interpersonal communication, that is, users can communicate asynchronously and/or synchronously; and (ix) ease of adding, that is, users can add information and content to the system.

Based on the instructional quality of the interaction, Schwier and Misanchuk (1993) identified three levels of interaction, namely reactive, proactive, and mutual interactions. A reactive interaction is a response to a given question. Proactive interaction involves learner's construction and generation activities during the learning process. In a mutual interactive environment, the learner and system are mutually adaptive in reactions with each other. The relationships among the three levels of interaction are hierarchical in terms of quality of interaction. Therefore, the quality of a mutual-level interaction is higher than that of a proactive-level interaction, and the quality of a proactive-level interaction is higher than that of a reactive-level interaction. Consequently, the higher levels of interaction provide a greater opportunity for mental engagements and learner involvements than the lower ones.

In the context of virtual reality, Steuer (1992) defines interactivity in telepresence as a concept which "refers to the degree to which users of a medium can influence the form or content of the mediated environment" (p. 80) and can be further understood in terms of degrees of speed, range, and mapping. Speed relates to how responsive the system is to the user's actions. Range refers to how many possibilities for manipulation there are in the mediated environment, including intensity (loudness, brightness, etc.), spatial organization (where objects appear, etc.), and temporal ordering. Mapping in this context refers to how closely actions taken on the mediated environment are mapped to corresponding "natural" actions in the human physical environment, thus contributing to a sense of telepresence.

New media are usually intended to include electronic games. In such a context, interactivity is considered a very important design dimension. Friedl (2002) distinguishes three levels of interactivity in games: player to player, player to game, and player to computer. By understanding these different types of interactivity, game designers can identify the factors that affect each and use appropriate techniques and methods in their own game designs. These factors also provide a basis for continuous evaluation during the development process and can be used to classify and analyze online game designs.

The above variations on the theme of interactivity lead to two observations. First the concept of interactivity in HCI, while somehow commonly understood, is still subject to research in order to achieve consensus on its constituting dimensions across perspectives and application domains. Second, many of the aspects of interactivity which have been taken into account in existing work in the HCI field are closely interrelated with aspects of human communication as briefly outlined in Section 2.1.

The next section will present an analysis of popular design frameworks as they have evolved in the HCI field over the years, attempting to highlight how emphasis on different design concerns impact the interactivity of the design outcomes.

## 3   THEORETICAL FRAMEWORKS

This section looks into the theoretical frameworks that influenced interaction design and consequently the development process, the degree of human-centeredness of the proposed interaction paradigms, and eventually the evolution of how users use and perceive interactive technology. In the overall evolutionary history of interactivity, these theoretical frameworks can be thought of as driving forces that influence and guide interactive technology development, with the underlying assumption that the more human-centered the design process, the more interactivity assumes a fundamental role.

### 3.1   Human Factors and Ergonomics

Frederick Taylor's (1911) *The Principles of Scientific Management* was arguably the first publication aimed at improving the work practice using the new technologies of the time in the industry. The prime motivation was improving efficiency, and although this was more of a management-oriented work than actual human factors, it served as the starting point for looking scientifically at the work process and the people who executed the necessary tasks.

Efficiency and cutting down time and costs was the primary motivation, but during World War II there was a shift of priorities. It was important to improve the safety and efficiency of aircraft cockpits and weapons systems to cut down on human loss. Here is where the actual foundations of human factors and ergonomics lie. In the work done to improve aircraft equipment and controls, the focus was on the human side, and instead of trying to suit people to the work task and

$$T = a + b \log_2\left(1 + \frac{D}{W}\right)$$

**Figure 1** Fitts's Law, where $T$ is the average time taken to complete the movement; $a$ represents the start/stop time of the device (intercept) and $b$ stands for the inherent speed of the device (slope). These constants can be determined experimentally by fitting a straight line to measured data. The distance $D$ is from the starting point to the center of the target and $W$ is the width of the target measured along the axis of motion, where $W$ can also be thought of as the allowed error tolerance in the final position, since the final point of the motion must fall within $\pm W/2$ of the target's center.*

equipment, the opposite was now the case; the U.S. army employed psychologists who studied the behavior of pilots and made measurements to determine human capabilities and limitations. An obvious example that was derived from this research domain was the work of Fitts, made famous by his law (see Figure 1), which was part of a study aimed at finding optimal designs and solutions for cockpit controls that would be easier to use by pilots. From the point of view of interactivity, early work in human factors and ergonomics can be seen as a first attempt to match the system's characteristics to the human user and design systems which are better suited to human abilities.

After World War II, in 1957, the aviation psychologists formed the Human Factors Society (HFS, later named the Human Factors and Ergonomics Society) in the United States. This new discipline thrived in the following years (Grudin, 2008).

### 3.2 Model Human Processor/GOMS

The most influential work that emerged from the marriage between cognitive psychology and human factors, as well as computer science, was the model human processor (MHP; see Figure 2), first presented in detail in the seminal book *The Psychology of Human-Computer Interaction* by Card et al. (1983). It is interesting to note the background of the authors. Card, with a background in psychology, was working in Human Factors at Xerox, and Newell, already a well-respected researcher in artificial intelligence, took an interest in studying human behavior (McCarthy, 1988).

The model is based on looking at the interaction between the human and the computer fundamentally as an information-processing task, that is, treating the two parties of the interaction, the human and the computer, as two information-processing systems, each with its own properties, performance capabilities, and limitations. The human user is the party that has specific goals and attempts to accomplish them by feeding commands to the computer system. Output from the computer is processed and reviewed, and the cycle continues until the user goal is accomplished. In the MHP, the human information-processing system is treated in terms

analogous to those of the computer system; the human cognitive architecture consists of 3 processors (perceptual, cognitive, and motor), the memory (working and long term), 19 parameters, and 10 principles of operation.

The presentation of visual information on the computer display is perceived by the perceptual processor (basically the eyes and ears). The cognitive processor processes information chunks from memory that was put there by the perceptual processor. The motor processor acts after those chunks have been processed and evaluated and a decision toward accomplishing a goal has been taken. An action may be performed by the motor processor. All these events can be isolated and ultimately analyzed to find the optimal and most efficient solutions.

The authors did not imply that human users do actually operate fundamentally as computers do; in fact they explicitly stressed that this is not the case. However, the model provided tangible and quantifiable measures of performance for some of the tasks and consequently allowed measuring (in some cases at least) how different presentations and parameters of the interaction affect the efficiency and usability of the communication between the two systems and how the user may benefit in accomplishing his or her goals.

In the same book, the GOMS (goals, operators, methods, and selection rules) analysis framework was also presented. A GOMS analysis of a task describes the hierarchical procedural knowledge a person must have to successfully complete that task. Based on that and the sequence of operators that must be executed, it is possible to make quantitative predictions about the execution time for a particular task. Other analyses, such as predictions of error, functionality coverage, and learning time, are also sometimes possible. Since the original formulation presented in Card et al. (1983), a number of different forms of GOMS analysis have been developed, each with slightly different strengths and weaknesses (John and Kieras, 1996; Byrne, 2008)

The above approaches have led to further analysis of humans as information-processing systems. The fundamental assumption and identification of the human user as a system that can be divided into further subsystems that can be analyzed have produced a significant volume of work (mainly from the human factors side of the community), which in large part has been applied successfully in HCI. For example, under the human information-processing system approach, extensive work has been conducted on issues of human attention, distraction, memory performance, problem solving, response times, and more, all of which are related to interactivity phenomena. While the hard science behind this work is irrefutable, there are however differing opinions on the success of the application of such models among HCI specialists, specifically regarding the overall interaction experience from the perspective of the user and the general context of use. A complementary approach to the human information-processing model that partially addresses this issue is the ecological approach (Wickens and Carswell, 2006), which takes into account the environment and views the information flow, not in distinct stages, but as an integrated flow.

---

* All figures in this chapter are in the public domain and have been retrieved through Google.
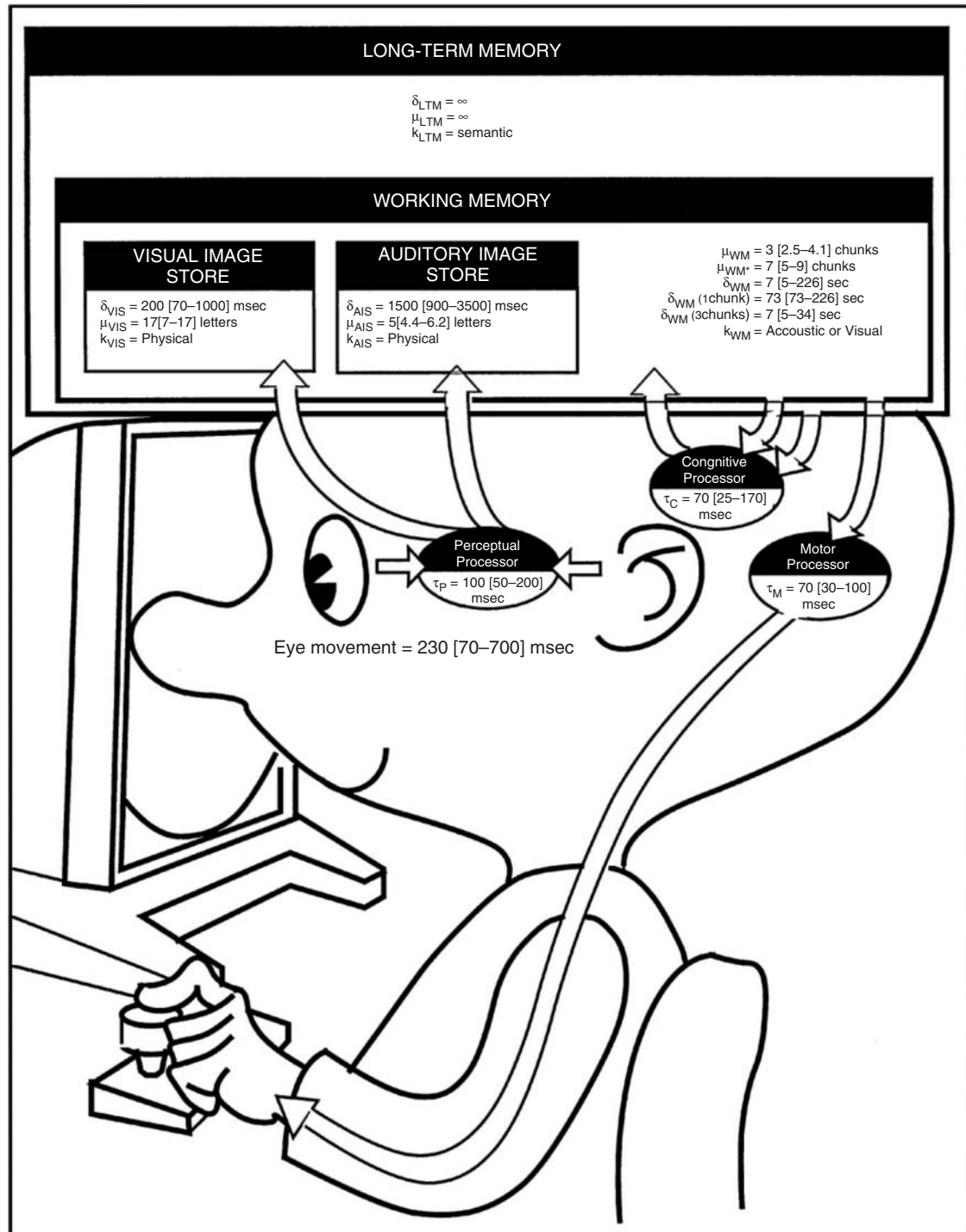
**Figure 2**    Model human processor.

Three years after the publication of *The Psychology of Human-Computer Interaction* (Card et al., 1983), another approach came into focus from the cognitive science discipline which tried to address the lack of accounting for the personal experience of the user in the above models, named user-centered design, which is addressed in the next section.

## 3.3 User-Centered Design

User-centered design is among the most influential design philosophies in the HCI community and had a major impact in the design and development of information systems. Traditionally, the software development process was seen as a series of distinct stages of activity that reminisce a waterfall, in which each activity naturally leads into the next. This is generally known as the waterfall model (Royce, 1970).

Inherited from the traditional engineering industry, the waterfall model divided the process into neat and manageable sections, ideal for setting and monitoring deadlines, as well as producing a rich amount of tangible deliverables. However, its monolithic nature is not really suited for the development of software, especially when usability issues are taken seriously. User-centered design became the alternative that emerged from the HCI community, shifting the focus from a technology-driven approach to the user being the center of each development phase. In addition, it responded to the unrealistic distinction between each stage of development by calling for the blending of each development phase and the need for iteration in the various stages of the process life cycle, with each iteration loop ending with an evaluation of the outcome based on user feedback.

As a result of this approach, every stage in the life cycle is characterized by the strong involvement of users, in the beginning as the main source of requirements specification, later as the main providers of feedback, to the final evaluation of the product with user testing. Each stage of the development cycle has been a subject of research and the past years have seen the introduction of many different techniques to enhance the quality of the outcome. For example, in the requirements specification stage, analysts employ the use of field observation, focus groups, personas, diary keeping, and more. In the design stage, rapid prototyping and evaluation techniques with user testing or expert evaluation techniques are adopted.

In 1986 Norman and Draper were the editors of a collection of papers under the title *User Centred System Design* (UCSD), which they described as "the design of computers, but from the user's point of view." In his chapter in the book, entitled "Cognitive Engineering," Norman presented his model of HCI, based on cognitive science. This model decomposes human action into seven distinct stages, starting from establishing a goal to evaluating the outcome from the computer in relation to this goal. The precise number of stages in the model can vary; the author nevertheless claimed that any theory of action involved a continuum of stages in the action/execution side and similarly in the perception/evaluation side of the full interaction cycle.

In the same chapter he also mentions mental models, a concept that he also discussed, among others, in the book *Mental Models* (Gentner and Stevens, 1983): "people form internal, mental models of themselves and of the things and people with which they are interacting. These models provide predictive and explanatory power for understanding the interaction" (p. 7). Mental models were another cognitive psychology construct that was primarily used to explain how people perceived the world around them and how these affect cognition and reasoning in interaction.

## 3.4 UX, POET, and Emotional Design

The term *user experience* (UX) was made popular by Norman, Miller, and Henderson when they were working for Apple in the 1990s. In 1995, they published a paper which dealt with the cross-organizational process that Apple used in interface research and development. The overall process was called user experience. The defining feature of the UX process was to view the user's interaction, not just in terms of hands-on experience with the company's product, but, more broadly, encompassing all interaction with the company itself, including marketing, retail, support, and services. In practice, this meant the bridging of various departments within the company, keeping them inside the loop of development and emphasizing intercommunication and collaboration.

In the beginning the term did not have a well-defined meaning and was subject to diverse interpretations. For example, as it coincided with the explosion of the Web and the dot.com bubble (see Section 4.5), companies used the term as "user-centered design for the web" (Morville, 2010). In time and with adequate clarifications given by its influential originators, user experience has been established as a well-understood concept. ISO 9241-210 gives this definition: "a person's perceptions and responses that result from the use or anticipated use of a product, system or service (International Organization for Standardization, 2010)." The definition's notes explain that user experience includes the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviors, and accomplishments that occur before, during, and after use. Three factors are listed that influence UX: system, user, and the context of use.

The Nielsen-Norman Group's definition is given from a company's perspective:

> "User experience" encompasses all aspects of the end-user's interaction with the company, its services, and its products. The first requirement for an exemplary user experience is to meet the exact needs of the customer, without fuss or bother. Next comes simplicity and elegance that produce products that are a joy to own, a joy to use. True user experience goes far beyond giving customers what they say they want, or providing checklist features. In order to achieve high-quality user experience in a company's offerings there must be a seamless merging of the services of multiple disciplines, including engineering, marketing, graphical and industrial design, and interface design.[*]

---

[*] http://www.nngroup.com/about/userexperience.html

UX views HCI in its broader context. When an organization creates technology, it embodies in a product its idea for a solution to an end-user problem, with the goal that this will ultimately help the organization itself. HCI is how the end user interacts with the product, but this symbiotic relationship between the end users and the organization lies at the core of how that interaction is structured. Understanding the user experience, therefore, is the process of understanding the end-user needs and the organization needs with the goal of maximizing the benefit to both.

In his *The Psychology of Everyday Things* (POET), a book highly influential among the design community, expanding on the ideas presented in UCSD, Norman (1990) emphasized usability and making products easier to use. Design aesthetics were not really considered, in fact there were remarks about designers winning awards for products that lacked usability. However, aesthetics and beauty are always major factors while considering a product. The book is highly influential as noted, but it did receive criticism regarding this. Norman included the gist of this criticism in his 2002 essay called "Emotion and Design: Attractive Things Work Better":

> If we were to follow Norman's prescription, our designs would all be usable, but they would also be ugly.

Seeing that this critique was in fact valid, Norman looked into how emotion and affect influence the user experience, acceptance, and ultimately preferences. The 2002 essay became the preface for his 2004 book *Emotional Design*, which again proved to be influential in the design community. In the book, the author presented three aspects of design that deal with human response towards it, based on psychology's ABC model of attitudes. The ABC model stands for affect, behavior, and cognition. Translated to design aspects these became visceral, behavioral, and reflective. The behavioral aspect can be thought of as "traditional" HCI territory; effectiveness of use, how well the design fulfills its purpose. The other two aspects were the so-to-speak missing pieces for a more holistic approach to design, namely emotion and rationalization. The visceral aspect deals with the design's appearance and beauty and how those affect users. It is part of human nature, a system to make rapid judgments of what is good or bad, safe or dangerous, and of course if something is beautiful and desirable. The reflective aspect is the rationalization and intellectualization of a product. A product can be totally unusable but still be desirable, to the point that users will forgive its shortcomings in usability. A telling example was the teapot for the masochist (see Figure 3), famous from the cover of POET. Not particularly beautiful (visceral), certainly not useful but scores highly on the reflective aspect; it can become an object of discussion, it tells a story, and it is unique and therefore desirable. What was more interesting, however, was that all these three aspects are influencing each other and a designer may actually take advantage of this.

The point of these aspects regarding design is that it was not enough to focus on usability alone. A user



**Figure 3**   The masochist's teapot.

will begin evaluating a product (including interfaces) and form an opinion about it from the moment he or she looks at it. The response of the user on the visceral and reflective levels will influence the experience of usage. The design must therefore be appealing aesthetically, which will make the user invest more time in interaction with it to learn how to use it. The same applies for the reflective aspect. If the user perceives that the design appeals to his or her self-image, it will enhance the visceral appeal or more surprisingly the behavioral appeal (see Figure 4).

Don Norman's work was not the first to underline these aspects of design, although it certainly helped bring them into the spotlight. Research into emotions and affect and their influence on cognition was being done years before. The essay, for example, cites the experiments conducted by Kurosu and Kashimura (1995) and then duplicated by Tractinsky (1997), which proved that indeed aesthetics and affect play a role in the usability of interfaces.

### 3.5   Universal Access and Design for All

The emergence of the Web and of the so-called information society in the 1990s brought about radical changes in the way people work and interact with each other and with information. In this context, the "typical" computer user can no longer be identified. In the past, the typical computer user was often considered as a professional, capable and willing to
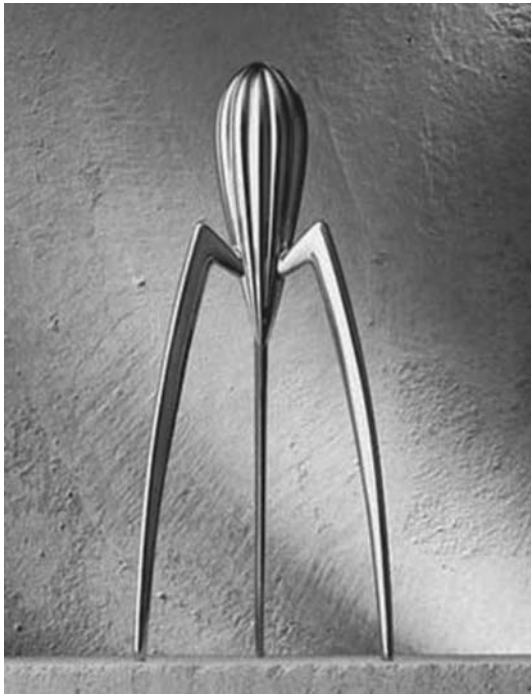
**Figure 4**   Juicy Salif — an example of emotional design.

use technology in the work environment in order to increase productivity and performance. In the new environment, interactive artifacts are used by diverse user groups, including people with different cultural, educational, training, and employment background, novice and experienced computer users, the very young and the elderly, and people with different types of disabilities.

Accordingly, while in the past computer-mediated human activities were mainly oriented toward the business application domain, in the context of the information society existing applications undergo fundamental changes, and new ones appear. The latter include access to online information, e-communication, digital libraries, e-business, online health services, e-learning, online communities, online public and administrative services, e-democracy, telework and telepresence, and online entertainment.

Finally, technological proliferation increases the range of systems or devices facilitating access to information resources. These devices include personal computers, but also standard telephones, cellular telephones with built-in displays, television sets, information kiosks, and various types of information appliances. Depending on the context of use, users may employ any of the above to review or browse, manipulate, and configure information artifacts at any time.

The above radical changes brought about the need to revise HCI frameworks and approaches to cater for a much larger and diversified user base and context of use,

leading to the concepts of universal access and design for all (see also Chapter 54).

In Stephanidis et al. (1998) universal acccess is defined as follows: "Universal access in the Information Society signifies the right of all citizens to obtain equitable access to, and maintain effective interaction with, a community-wide pool of information resources and artifacts" (p. 6). Accessibility has been a term traditionally associated with elderly individuals, individuals with disabilities, and more in general individuals with functional limitations (Stephanidis et al., 1999). However, because of the current influx of new technologies into the market, the population of users who may have particular interaction requirements is growing. As a result, accessibility has taken on a more comprehensive connotation. This connotation implies that all individuals with varying levels of abilities, skills, requirements, and preferences be able to access information technologies (Stephanidis et al., 1999). Universal access also implies more than just adding features to existing technologies. Rather, the concept of universal access emphasizes that accessibility be incorporated directly into the design (Stephanidis et al., 1998).

The term *design for all* denotes an effort to unfold and reveal the challenges of accessibility and usability as well as to provide insights and instrument appropriate solutions in the HCI field (Stephanidis et al., 1998). The fundamental vision is to offer an approach for developing computational environments that cater for the broadest possible range of human abilities, skills, requirements, and preferences.

Design for all in the information society is the conscious and systematic effort to proactively apply principles and methods and employ appropriate tools in order to develop information technology and telecommunication (IT&T) products and services which are accessible and usable by all citizens, thus avoiding the need for a posteriori adaptations or specialized design. Design for all in HCI recognizes, respects, values, and attempts to accommodate the broadest possible range of human abilities, requirements, and preferences, eliminates the need for "special features" and fosters individualization and end-user acceptability.

Design for all fosters a proactive strategy, postulating that accessibility and quality of interaction need to be embedded into a product at design time. This entails a purposeful effort to build access features into a product as early as possible (e.g., from its conception to design and release). In the context of HCI, a proactive paradigm is advocated for the development of systems accommodating the broadest possible end-user population. In other words, design approaches are required that seek to minimize the need for a posteriori adaptations and deliver products that can be adapted for use by the widest possible end-user population (adaptable user interfaces).

This implies the provision of alternative interface manifestations depending on the abilities, requirements, and preferences of the target user groups. The main objective in such a context is to ensure that each end user is provided with the most appropriate interactive experience at run time. Producing and enumerating

distinct interface designs through the conduct of multiple design processes would be an impractical solution, since the overall cost for managing in parallel such a large number of independent design processes, and for separately implementing each interface version, would be unacceptable (Stephanidis, 2001).

The scope of design for diversity is broad and complex, since it involves issues pertaining to context-oriented design, diverse user requirements, as well as adaptable and adaptive interactive behaviors. This complexity arises from the numerous dimensions that are involved and the multiplicity of aspects in each dimension. In this context, designers should be prepared to cope with large design spaces to accommodate design constraints posed by diversity in the target user population and the emerging contexts of use in the information society. Moreover, user adaptation must be carefully planned, designed, and accommodated into the life cycle of an interactive system, from the early exploratory phases of design through to evaluation, implementation, and deployment. Additionally, design for diversity is anticipated to be an incremental process in which designers need to invest effort in anticipating new as well as changing requirements, accommodating them explicitly in design through continuous updates.

In terms of interactivity, universal access and design for all introduce two important dimensions previously overlooked in HCI: the individual diversity of users as well as the need to adapt interaction behavior to such diversity. This implies that there is no best interaction style, but different interaction styles may be appropriate in different circumstances depending on the involved users and the context. For example, universal access fosters the view that both visual and nonvisual (e.g., speech-based) rendering of an interface dialogue can be provided (either alternatively or multimodally) to cater for the interaction requirements of sighted and blind users.

## 4  EVOLUTION OF INTERACTION

The evolutionary history of interaction began with the introduction of computer technologies in the 1940s and 1950s. The primary force behind this evolution was technological research. Over time, when new technologies became mature enough to allow for different approaches to interactivity, the theoretical frameworks briefly outlined in the previous section came into play, along with business and user adoption of the various interactive products. This section looks into the most widespread paradigms of interactivity. This includes describing the evolutionary history of the interactive technologies involved but also, where appropriate, the evolution of specific application domains, such as the World Wide Web, where the key aspect of interactivity is not the technology involved (in this case network servers and clients, the infrastructure, and the various protocols) but the context of use and the new information space that is available to humans.

The above-mentioned evolution is punctuated by the debate among approaches based on the conversation metaphor, which tries to emulate human dialogue, and the model world metaphor, which emphasizes the user's direct action mediated through a visual language (Hutchins et al., 1986). The conversational metaphor privileges textual language, whereas the model world paradigm relies more on iconic languages (although text may also be present).

The conversation paradigm lies at the basis of interaction paradigms such as command-based interfaces and speech interfaces, whereas the model world metaphor informs direct manipulation interaction and all its subsequent evolutions. Whereas the conversation paradigm addresses interactivity by progressively developing methods and tools to better understand the communication context, the model world paradigm adopts a radically different approach, whereby the context is reconstructed visually to ground communication.

In more recent years, these two paradigms appear to merge in multimodal interfaces, virtual environments, and ambient intelligence (AmI) environments.

### 4.1  Early Stages

HCI started in the 1940s with the construction of the first computers. At that time, interaction was very cumbersome and limited to trained scientists. ENIAC (see Figure 5), arguably the first general purpose electronic computer, was a massive machine that occupied a large room and needed weeks to program via punch cards. Such were the machines for about a decade. While the main focus was to keep the machines working correctly, that is, functionality, there was also an effort to make the interaction easier for the operators by formatting printouts and reducing the programming tasks by creating machines that could store programs, usually on tape (Grudin, 2008). In the strictest sense, the interaction was limited to basic operation of the machine. The process of programming and using the output was a separate work done away from the computer itself. In 1955 transistors replaced tubes in computer hardware, but the basic mode of interaction was basically the same. Until the mid-1960s this was the norm; the only people who had hands-on access were operators and data entry personnel who dealt with switches, knobs, and dials. It was not until the invention of the microchip and later the microprocessor that the way people interact with computers would change radically, making HCI a discipline that is relevant and essential to the design of computers.

This has been called the first wave of computing (Weiser and Brown, 1997), characterized by many people making use of a single computer. Computers were large mainframe machines that people booked time on to run their programs and get back their results, without actually dealing with the machine itself.

The second wave of computing is the current one, the era of the personal computer (PC), when advancement in technology made it possible to have small PCs for individual discrete use. During this period, there was an explosion in the evolution of interaction, starting from the command-based interfaces, to graphical user interfaces (GUIs) and direct manipulation, to Web-based and mobile interaction.
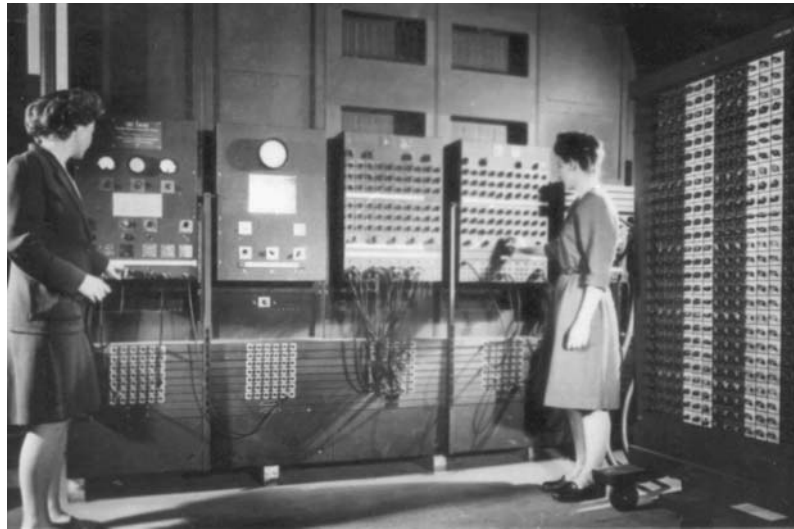
**Figure 5**  The ENIAC computer.

The third wave has been given many names, such as ubiquitous computing, the invisible computer, AmI, and calm technology, each name pointing to the presence of computers in the environment, in the background, where one person now has access and use of many computers distributed in the surrounding environment.

### 4.2  Command-Based Interaction

Command-based interaction was typical of command line interfaces. These were the de facto interfaces that people used in the 1970s and a significant part of the 1980s. Command line interfaces are based on the conversation model and provide a means of passing instructions to the computer by typing them in with the keyboard in the form of commands, which could be whole words or abbreviations. An example is the Windows Command Prompt, where the user can type in DOS-type commands.

Their advantages are that they provide direct access to system functionality and can be very efficiently used by expert users. They can also be very flexible with the use of parameters that allow the user to perform complex tasks with one command.

Their first incarnation was in the form of teletypes, where the operator would type in commands and receive one-line output from the computer, such as feedback or status messages, on scrolling paper. These were then replaced by glass teletypes or video display units, available to very few at the beginning until the CRT (cathode ray tube) monitors that became more widespread (see Figure 6).

On the other hand, command line interface users complained about the slower speed and lack of flexibility in entering multiple commands. But the most serious problem with command-based interaction is the slow learning curve, as the user must learn the various, often arbitrary commands. Indeed, one task of HCI at the time

was the design of command names in order to make them more easily memorable. So, even if expert users tend to prefer working with command line interfaces, most people viewed this type of interaction baffling and the sight of a blinking cursor on a screen did not give any clues to the proper use of the system. Command-based interfaces, although they attempt to establish a simple form of dialogue, present very limited interactivity, as there is no conversational context, and exchanges of command feedback are independent from each other.

Until the 1980s, computers were reserved for use in work by trained staff performing often tedious data entry tasks and for the computer enthusiasts. It would not be until the commercial success of the Mac in 1985, and more prominently of the Windows 3.0 environment in 1990, that command line interfaces would gradually disappear or start being integrated into GUIs. In fact, many applications currently still offer a dual mode of interaction by incorporating a command line interface complementary to the GUI.

### 4.3  Direct Manipulation

Direct manipulation (Schneiderman, 1983) was the next major step in the way people interacted with computers. Related research started in the 1960s and brought out novelties such as the mouse or the GUI (Engelbart, 1963; Engelbart and English, 1968; Sutherland, 1963), was further developed into prototypes in the 1970s until finally Xerox's Palo Alto Research Center completed the Alto and then the Star system. It would take a few more improvements and additions by Apple in the Lisa computer and more importantly in the Macintosh, which became a commercial success to bring the Desktop Metaphor and Direct manipulation to the mainstream. This shift was solidified with the global success of Microsoft's Windows, starting in 1990 (Windows 3.0) and exploding by the time Windows 95 was released.

**Figure 6**  Command line interfaces.

This new means of interaction offered easiness of use, more intuitive interfaces, and a richer experience. It marked the beginning of the second wave of computing, where PCs were available to the wider audience for discrete use, and with the addition of the more friendly interaction that GUIs and direct manipulation interaction offered, suddenly computer use was becoming the norm instead of a sophisticated work tool that needed extensive training to operate (see Figure 7).

In direct manipulation interaction users perform actions directly on visible objects with the use of a pointing device, most usually a mouse. Objects include window controls, menus, icons, buttons, and other elements in the so-called WIMP (windows, icons, mouse, pointer) interfaces (such as those mentioned above), but there are other types of direct manipulation interfaces, such as 3D interfaces [in virtual reality (VR) environments] or haptic interfaces.

The advantages were obvious compared to command line interfaces. The new interfaces were easier to learn and remember; the user received immediate visual feedback and more accurate representations of what he was working on, that is, WYSIWYG (what you see is what you get), and it was easier to reverse actions (undo), which meant that the new interfaces were also less error prone. Finally, these visual-rich interfaces exploited more fully the human use of visual–spatial cues.

On the other hand, the new paradigm was a much more challenging job for the designer; instead of formatting screens of text, the designer should consider color choices, layout choices, GUI control choices, and a lot more. Even many years after the establishment of the WIMP style of interfaces and extensive HCI research on the subject, it is not an easy job.

Direct manipulation interfaces can be distinguished in two main categories, namely WIMP and post-WIMP.

WIMP are the classic interfaces that have been the standard GUIs for most computers in the past 25 years. The name comes from the elements that characterize the interface, that is, the windows and icons on the screen manipulated with the mouse pointer. As described above, this style of interaction originated in the late 1970s, when the GUI was combined with the mouse in the Alto machine and then the Star. The

**Figure 7**  Graphical user interfaces.

Macintosh would make the match a commercial success, introducing the now familiar Desktop Metaphor to the public, where the screen is metaphorically viewed as the top of a desk and various objects that lie on it may be manipulated by the mouse pointer.

While WIMP-style interfaces still dominate as the interaction style, there are several other examples of direct manipulation interfaces that either expand on the classic WIMP interfaces or are entirely different. An example of the former would be the recent multitouch interfaces developed by Apple,[*] where the use of a touchpad (on the macbook) or the screen itself (as in the iPad) allows the user to manipulate objects via different gestures and movements.

---

[*] http://www.apple.com/magictrackpad/

An example of the latter is Google Earth's* Zooming User Interface (ZUI), where the interaction is focused in the zooming in and out of a 3D photorealistic representation of Earth and clicking points of interest.

These two examples also serve to underline another important distinction between direct manipulation interfaces which concerns the hardware available. It is more common to visualize a typical personal computer with a keyboard and a mouse for input and a monitor for the outpu,t but the example of Apple's latest hardware points to new input techniques because of the novel hardware. There are more similar examples, for example, surface computing hardware, which are typically large surfaces that act as touch screens, where the user directly manipulates objects by touching them. Another example would be VR environments, where the hardware could include a VR helmet and a data glove or data wand to manipulate virtual objects (see Section 4.8).

Although they radically differ from the conversational paradigm of interaction, direct manipulation interfaces introduce several elements of interactivity. First, they establish a visual context, which plays the role of grounding the human–computer communication process. Second, they provide more structured dialogues and more articulate feedback with respect to command-based interfaces. An important aspect of direct manipulation is its reliance on metaphors. In this respect, visual languages used in direct manipulation interfaces fundamentally differ from natural language. However, it is exactly this metaphoric value which allows the direct manipulation paradigm to ground communication and enrich dialogue, by offering a real-world context in which users' actions can be rooted.

### 4.4 Conversational Interfaces

During the late 1970s and 1980s, when significant advances in the fields of artificial intelligence, natural language processing (NLP) appeared to be maturing, a trend toward the explicit provision of humanlike communication in user interfaces emerged. In the context of HCI, NLP applications range from various speech recognition systems to natural language interfaces to database, expert, and operating systems (Manaris, 1998).

As the use of computers expanded throughout society, affecting various aspects of human life, it became clear that the number and heterogeneity of computer users was were dramatically increasing and that that many of these users were not computer experts. The provision of interaction means exploiting natural language was conceived as a potential path towards increasing the user friendliness of interactive computer systems and their eventual acceptability to users.

NLP offers mechanisms for incorporating natural language knowledge and modalities into user interfaces. As NLP tools started becoming more powerful in terms of functionality and communicative capabilities, their contribution to HCI also became more significant.

The history of NLP can be very briefly summarized into three main phases. The first phase started in the mid-1940s and lasted until the early 1960s. It is characterized by an emphasis on algorithms relying mostly on dictionary-lookup techniques used in conjunction with empirical and stochastic methods. During this phase, some NLP application areas began to emerge. An example is speech recognition, employing speaker-dependent, template-based architectures.

The second phase in NLP spanned approximately from the early 1960s until the late 1980s. It was characterized by (a) a strong emphasis on language theory, including the lexicon, syntax, semantics, and pragmatics; (b) the construction of "toy" systems that demonstrated particular principles; and (c) the development of an NLP industry which commercialized many of the achieved results. In terms of applications, this phase is mainly characterized by question-answering systems and database interfaces (1960s) as well as interfaces to other interactive systems. During this phase, it became apparent that symbolic approaches to NLP problems were not adequate when attempting to widen linguistic coverage or apply the developed systems to a different domain. This realization motivated the development of nonsymbolic approaches, mainly based on statistics, connectionism, or the analysis of language corpora.

In the late 1980s, NLP entered an empirical and more "user-centered" phase. Major advances and tangible results from the last 50 years of NLP research were reinvestigated and applied to a wider spectrum of "real-life" tasks, including, for example, spelling checkers, grammar checkers, and limited-domain, speaker-independent, continuous-speech recognizers for various computer and telephony applications.

During this phase, HCI entered the mainstream of computer science. This was a result of the major advances in graphical user interfaces during the 1980s and early 1990s (see Section 4.3), the proliferation of computers, and the emergence of the World Wide Web (see Section 4.5). Accordingly, the evolution of NLP reflected the continued growth of research and development efforts directed toward performance support, user-centered design, and usability testing. Emphasis was mainly placed on systems integrating speech recognition and traditional NLP models as well as hybrid systems—systems combining results from symbolic, stochastic, and connectionist NLP approaches.

A conversational agent is the human—computer dialogue system that interacts with the user turn by turn using natural language. Efforts were initiated by Alan Turing (1950) in his famous paper "Computing Machinery and Intelligence." He suggested that within 50 years a computer would pass a comparison test if it is a human or a machine. Historically the first conversational agent was the ELIZA system (Weizenbaum, 1966). ELIZA featured the dialogue between a human user and a computer program representing a psychotherapist. ELIZA is based on simple stimulus–response architecture (i.e., patterns and related responses).

LUNAR (Woods et al., 1972), developed in late the 1960s, was the first natural language interface to a database (NLIDB). The LUNAR database contained chemical analyses of moon rocks and had a significant influence on subsequent computational approaches to natural language. Subsequently several NLIDB appeared

exploiting different approaches to handling natural language. By the mid-1980s, natural language interfaces to databases became a very popular research area, and numerous research prototypes were implemented.

These initial attempts identified a series of practical difficulties in the development of NLIDB, including extensibility of lexica and grammars and portability of systems across different application domains. Other criticisms came from the HCI community, since natural language was considered too ambiguous to provide effective communication in user interfaces. On the other hand, when restricted to some sublanguage to limit ambiguity, natural language loses its distinctive feature of allowing free expression and becomes more similar to a command or formal language which needs to be learned to be used (Hill, 1983).

Additionally, natural language interfaces were criticized for leading users to anthropomorphize the computer, or at least to attribute more intelligence than is warranted to it. This leads to unrealistic users' expectations regarding the capabilities of the system, and in turn such expectations lead to disappointment when the system fails to perform accordingly (Shneiderman, 1998). Various experiments were conducted to find out how users can adapt to system's restrictions in vocabulary and syntax, and the results appeared to confirm that humans are keener to learn a command language than to restrict their use of natural language to conform to the system's limited abilities (Slator et al., 1986; Ogden and Bernick, 1997).

In subsequent years, natural language interfaces have been applied to operating systems (Manaris, 1994) and information retrieval (Jacob and Rau, 1988). The focus in more recent approaches has been more on using speech as well as on achieving more natural forms of dialogue communication. Systems that use speech interfaces range from call routing to navigation systems to VoiceXML-type applications which enable speech interfaces on the Web (Jokinen, 2009). The common technology is based on recognizing keywords in the user utterance and then linking these to appropriate user goals and further to system actions. Speech-based conversational interfaces, besides recognizing speech input, also provide speech output (Zue and Glass, 2000).

Besides the capability to understand and generate linguistic expressions, some systems include cooperation and planning of complex actions on the basis of observations of the communicative context, that is, communicative competence (Jokinen, 2009). A well-known example of plan-based system is TRAINS (Allen et al., 1995), a train route planner, where a human manager and the system must cooperate to develop and execute plans.

The notion of natural interaction in this context refers to the spoken dialogue system's ability to support functionality that the user finds intuitive and easy. The challenge that speech and language technology faces in this context is not only in producing tools and systems that enable interaction in the first place but also to design and build systems that allow interaction in a natural way, that is, to provide models and concepts that enable experimentation with complex natural language interactive systems and to test hypotheses for flexible HCI.

In addition to improved interaction strategies, natural language interfaces are also required to be extended in their knowledge management and reasoning capabilities, so as to support inferences concerning the user's intentions and beliefs behind the observed utterances. The goal of building natural interactive systems thus comes close to studying intelligent interaction in general. Toward this end, research efforts have attempted to achieve context understanding (Jokinen, 2009) as well as to exploit the notion of grounding (see Section 2.1) which is inherent in human communication (Traum, 1999).

Further evolution of research toward natural dialogue–based communication has led on the one hand to developing the concept of multimodality (see Section 4.7) and on the other hand to embodying conversational systems in anthropomorphic representations (Cassell, 2001). Such anthropomorphism, implemented in the form of animated avatars, is targeted to make explicit the system's intelligent behavior and allows representing the system's knowledge to humans in multiple ways on multiple modalities (e.g., speech and hand gestures).

## 4.5 Web-Based Interaction

The Internet materialized in late 1969, when ARPANET was deployed, connecting four academic institutions in the United States. Soon, with the implementation of several network protocols, islands of networked computers appeared, leading to the introduction of services that had been developed in previous years, such as hypertext, email, and eventually the World Wide Web (Hafner, 1998). At the beginning the latter suffered from the same problems of command line interfaces compared to GUIs. It was not until 1994, when Mosaic was released, the first graphical web browser, very similar to the browsers we use today, that Internet use exploded beyond academia and government. Three years earlier the Internet had been made available for unrestricted commericial use, but it would take the graphical Web browser to provide the missing piece to exploit the full implications that the Internet presented.

From an interaction perspective, in its first years the Web consisted of Web pages that contained primarily text, images, and links to other Web pages. The Web pages themselves were not highly interactive; apart from clicking on links and navigating through Web pages, the primary interaction style was the familiar form-filling paradigm (Grudin, 2008), including elements such as entering text, selection from menus, checkboxes and option buttons, and submitting the forms via command buttons. Content was primarily static, but a major difference in the context of use was the discretionary nature of interaction. This had already begun with the introduction of home computers in the early 1980s, where users could choose to use a computer instead of the alternative "traditional" method (e.g., consider the typewriter–word processor options), but soon, as computers became the standard tool in the workplace, their use was not a matter of discretion anymore. The Web brought discretionary use back into focus and the vast number of competitive content choices and the notorious impatience with slow speeds meant that

user discretion and preference became a major research theme in HCI.

Perhaps the most striking change brought about by the introduction of the Web into the wide audience was precisely its penetration into everyday households. The computer was technology used by people with specific tasks and discrete needs, whether it was a work tool or an entertainment device. With the Internet and the Web, the computer became an interactive communication window to the world, as revolutionary a change as the television, only with much more potential and with much richer interaction.

This journey into the household began with the introduction of affordable microcomputers. The Macintosh and Windows provided the appropriate interface to make the technology more accessible to the novice user, and finally the Internet and the Web made it indispensable to the citizen of the information age. The new avenue of e-commerce made possible by the Web also meant that virtually all business environments had to employ computers, if anything for communication and coordination purposes. From a sociological perspective, the Web also gave rise to social computing, the creation of virtual communities in the form of forums and newsgroups, and more recently social networking sites such as Facebook or LinkedIn. Because of the Web, people simply spend much more time interacting with computers; for example, it was reported in 2005 that 75% of Americans use the Internet and spend an average of 3 h a day online (Stone, 2005).

If interactivity between computers and humans was a secondary issue before, its research dependent on technology developments, the Web definitely brought it to the forefront of hot topics, a fact evident by the sudden flourish of activity in the HCI discipline.

Naturally, a major proportion of this activity revolved around Web design practices and evaluation techniques. The Web brought many new interaction design issues to the table. This was a natural consequence for two reasons. First, because the Web was a new technology, it suffered from a lack of concrete design guidelines and practices. Second, creating a Web page was considerably easier compared to programming an application, so content was created by many people outside the computing community. A lot of people started publishing Web pages, many of which suffered from serious, amateur design issues, such as excessive use of colors and images, unreadable content, and inconsistency.

Matters were complicated by the inconsistencies across Web browsers, the platform-independent nature of the Web, the lack of control from the part of the designer in the way pages render on the user's end (e.g., browser window size, font settings, laptops vs. desktop PCs) (Nielsen, 1997), the lack of common standards, and the introduction of new Web technologies that were developed to provide richer interaction. The latter, as all new technologies, suffered from misuse and abuse of featurism over usability.

Some of these problems were addressed by the formation of the the World Wide Web Consortium (W3C), founded in 1994. The task of the consortium was to develop international standards for the World Wide Web, including its primary language, Hypertext Markup Language (HTML) or the Cascading Style Sheets (CSS) language,[*] which seeks to separate content from presentation. Another important activity of the W3C was the development of accessibility guidelines, aiming to reduce as much as possible the exclusion of users with disabilities from accessing the Web.[†]

Less formally but perhaps more influentially, Jakob Nielsen's work on Web design through his articles on correct practices and criticizing bad practices (such as his famous article on the top 10 Web design mistake), along with a number of books on usability for the Web, played (and still play) a major role in reducing design mistakes which were so frequent in the early Web, such as poor layout, bad choice of colors, abuse of distracting and annoying animations, poor content management, inappropriate writing style and typography, excessive size of Web pages, and consequent slow loading times and more.

HTML was problematic for the creation of good Web pages. It was simply inadequate from an interaction perspective, offering very little beyond simple form controls. In response to this, CGI-scripts, mostly written in Perl, "the duct-tape of the Internet," according to Hassan Schroeder Sun's first webmaster,[‡] were employed at the beginning to add more programming power to developers. By the end of the century, languages had been introduced specifically for the Web, such as Java, PHP, ASP, or Flash. This finally allowed developers to create applications for the Web that could match, at least to a certain extent, the interactive richness of standard software applications.

HTML was also inappropriate as a layout tool, having been implemented as a hypertext language, aiming to link content, not present it. In response to this problem, Web developers relied excessively on elements such as tables or graphics hacks to realize their designs, leading to accessibility issues. The development of CSS by the W3C answered that problem to an extent.

Regardless of specific interaction issues of the Web, the main challenge users faced on the Web was locating the specific content they were looking for. The Web is a vast space of information and it was soon apparent that one of the most important issues was searching. The names of the first popular browsers were a clear indication of this problem. Navigator and Explorer both hinted at an ocean of information that the user could surf through. Search engines therefore became a critical application and usually the starting point of the Web interaction experience.

### 4.5.1 The .com Bubble and Web 2.0

One of the most important events in the history of the Web was the so-called .com bubble and burst (Cassidy, 2003). It is important not only as a socioeconomic phenomenon but also as a case study to gain

---

[*] http://www.w3.org/standards/webdesign/htmlcss

[†] http://www.w3.org/standards/webdesign/accessibility

[‡] http://www.oreillynet.com/pub/a/oreilly/perl/news/importance_0498.html

lessons about the nature of the Web and what actually worked or not. Not to be underestimated also is the significant exposure the new technology received through mainstream media, which also played a part in the growth of the Internet and computer usage.

The period covered between the beginning of the bubble until the burst is generally considered to be between 1995 and 2001. In that time, Internet use exploded in numbers and a lot of companies tried to exploit this new exciting medium. Everyone was certain that the Web was changing society and that a new huge potential market was made available but there was definite uncertainty on how exactly to exploit this market. Many companies were founded without a specific business model and consequently went bust without ever making any actual profit, erroneously thinking that traffic would somehow generate revenue through advertising or that the elimination of the traditional brick-and-mortar model would translate to profit. The hype in the stock market and the soaring of stock values of these so-called .coms sustained this illusionary impression until roughly the end of the decade/century/millennium, but at the end the bubble burst and only those who had understood the nature of the Web survived, indeed thrived.

One explanation for the success of these companies, such as Google, eBay, and Amazon, was that they matched and took advantage of the specifications of the so-called Web 2.0. This concept, first articulated in 2001 after the .com burst, sought to explain the common factor between the aforementioned companies and propose a new approach for understanding the Web. The underlying principle was that the Web should be seen as a platform, as opposed to a medium for which standard desktop applications should be developed. The example cited by the originators of the concept is Netscape versus Google (O'Reilly, 2005). Netscape began by trying to replace the desktop with the "webtop" (their browser) and planned to populate that webtop with information updates and applets pushed to the webtop by information providers who would purchase Netscape servers. In reality, value was transferred up the stack to services delivered over the Web platform. Google on the other hand was such a service, and customers were paying directly or indirectly for its use. There were no issues of new software releases or operating system (OS)/hardware-specific editions. Netscape relied on the classic software paradigm, Google on the concept of the service running between the two computers (the Google server and the user's computer), in the Web space. The real challenge was managing the data and turning it into useful information for the end user. Similarly, Amazon did not offer a particularly different catalogue of products than its competitors, but it invested heavily in the management of all sort of data, so it could make it into useful information for its customers that would lead them to a purchase.

This focus on the data and its processing into useful information also pointed to another significant factor, namely users add value. Google and Amazon do not actually produce any of the data that they serve to their customers. Google simply collects data from the Web

and indexes it. Amazon keeps track of user behavior. eBay offers a platform for users to conduct transactions. Wikipedia does not generate its content, it only offers a means to manage it. It is the users that provide the actual content and this ultimately means that if an application can provide a useful service with a critical mass of users, then it can be successful. The underlying principle is in essence a paraphrase of the well-known Open Source mantra "given enough eyeballs, all bugs are shallow" (Raymond, 2000); given enough users, the content is valuable. Typical examples of this principle, apart from those mentioned above, are blogs, wikis, media-sharing sites (such as YouTube), social networking sites (such as Facebook or Myspace), and so on.

Collective user-generated information also proved to be the best solution regarding the Web's primary challenge, that of searching. Google became the de facto search engine by exploiting user sponsorship of websites by considering links as sponsorships of approval, with great success. The same principle also applies to Amazon, as it exploited its users' selections as an indication of what is the most probable content they were looking for, in the form of suggestions for related content. A key feature of a successful Web service is to provide the easiest route to desirable content, and managing collective user data has proved to be a very efficient way to achieve this.

In terms of interactivity, it can be argued that not much has changed in the way users physically interact with the computer or in what type of interaction controls are used. The input and output devices remain the same and only the link is a Web-specific interaction element in the interface. However, there is undoubtedly a major transformation in the context(s) of use as well as the environment space that the user moves through, resulting in an overall different user experience that has a profound effect on the way humans perceive computers. The Web 2.0 paradigm of harnessing collective intelligence and user-generated content is one example of how computer services have changed significantly in the course of a decade. The concurrent developments in communication technologies, specifically wireless and mobile communications, and the widespread availability of high-end mobile devices have provided a set of emerging tools which are starting to be used in augmented-reality and AmI applications.

## 4.6 Mobile Interaction

Mobile interaction is a relatively new field of research in the HCI community, but it has become one of the most interesting, since the current generation of mobile devices has reached a technological maturity that allows much more sophisticated interaction than when they first appeared. Furthermore, the inherent property of mobility and their personal nature, coupled with the advanced processing power and multimedia capabilities, makes them a good candidate for playing a major part in the way people generally interact with computers in the future (see Section 4.8). Before examining the characteristics of mobile interaction and the design challenges that stem from the device properties, this section looks into the historical context of mobile

devices, focusing on mobile telephony and personal digital assistants (PDAs), the combination of which (the smartphone), will be the primary focus of the chapter.

In 1979 Sony released the Walkman. It was an instant success because it managed to take one activity, listening to music, which was confined to the home, and take it anywhere. Its obvious advantage and appeal were that it was small and easy to carry around and offered the same service, more or less, with another much larger cassete player. Interestingly, before the product was launched, critics thought it would be a commercial failure because it did not offer a recording function. It turned out of course that most people did not need that specific function but were very happy with being able to listen to music anywhere. The Walkman also marked a trend toward miniaturization and portability.[*]

PDAs are basically hand-held minicomputers. The term was first used by Apple in 1992 to describe the Apple Newton,[†] the company's first attempt at creating a mobile device which also featured a touch screen. Before the Newton, the line was blurred between small hand-held electronic organizers (such as the very minimal Psion Electronic Organizer and the quite sophisticated Sharp's Wizard series) and portable PCs, which were closer to the size of what we now refer to as notebooks. The latter trace their roots back to 1972, when Alan Kay proposed the design of Dynabook (Kay, 1972), which however was never built into a working prototype. The Dynabook is considered the ancestor of the laptop or the tablet PC and was a huge influence on the Palo Alto by Xerox, which Kay had joined in 1970.

PDAs were quite popular in the 1990s, but the market was fragmented and the devices never really caught on as more than electronic organizers in the market, offering calendar support, note taking, and so on, despite efforts from major players such as Microsoft to support the medium by releasing a PDA-specific OS, CE Windows. However, PDAs were used extensively in business and health care.

Mobile telephony had a similar evolution (see Figure 8). Telephony started as a fixed-location service, then moved into the car (although mostly in the United States), until the mobile phone in the mid-1990s. Mobile telephony took off in the mid-1990s and the first-generation cellphones were large and rather cumbersome devices, with a minimal screen. The primary use of mobile phones at the time was limited to calling and answering as well as text messaging. The standard interface was the keypad plus some buttons, without any standard configuration across the various devices. Technological advancements led to devices with much better processing power and screens to the point that these devices were matching the capabilities of PDAs. Roughly by the early 2000s, mobile phones offered color screens, better GUIs, and smoother interaction than first-generation phones. Usability matured in getting the right design for the handsets as well as establishing design guidelines for the specific challenges posed by the nature of the devices, in regards to interface styles, text entry, and so on.

---

[*] http://en.wikipedia.org/wiki/Sony_Walkman
[†] http://en.wikipedia.org/wiki/Apple_Newton



**Figure 8**  Evolution of mobile phones.

Appropriately designing the hardware controls was not a trivial task either. Nokia took four years to reduce the number of buttons (besides the standard numeric keypad) from eight to four, from 1994 to 1998. It took extensive research and user feedback to realize that the mobile phone user at the time primarily did two fundamental tasks: dialing from the phone book and answering the phone. Therefore, the elegant and successful solution they came with, having simplicity in mind, was to use one big prominent button (the Navi-Key) directly below the screen. The button was used to answer and hang the phone as well as confirming a selection (Jenson, 2002).

New-generation mobile phones are also referred to as smartphones, since the next logical step was to combine the functionality of the mobile phone with that of the PDA (see Figure 9). The first official smartphone is arguably the Simon, a device designed by IBM in 1992 and released in 1993 by BellSouth. It featured a stylus-operated touch screen and combined a mobile phone with many PDA capabilities, including email. Its initial price was too high to penetrate the market. In 1996 Nokia released the Communicator, a mobile phone literally combined with a PDA device, as the first prototypes were a Nokia phone hinged together with a HP PDA. The 9000 Communicator, as it was called,

**Figure 9** From PDAs to smartphones.

effectively marked the beginning of the smartphones. It was a very cumbersome device, and it did not affect the PDA market much, but eventually smartphones rendered stand-alone PDAs almost obsolete. The term smartphone itself was probably used for the first time in 1997 when Ericsson released the GS88 phone.

Mobile interaction is radically different from PC interaction. The reasons for this can be grouped into two categories, device characteristics and context of use. The most obvious device characteristic is of course its size. This includes the size of the screen, as well as the size of the controls, usually the keypad and navigation buttons of the mobile phone. The fact that there is no standard screen size or hardware controls further complicates matters. Furthermore, the input/output interface consists of the screen and the keypad, that is, no mouse or keyboard is available. These two characteristics are the main factors influencing interface design in mobile devices.

Because of the lack of a pointing device, although modern mobile devices are capable of rendering quality graphical user interfaces, albeit in a small scale, the majority of the devices do not offer direct manipulation. Instead, the major interaction paradigm is scroll and select, where the interface is presented most commonly in the form of a list-based layout. This layout also takes advantage of the fact that most mobile screens are portrait oriented (height bigger than width). Several variations or features of list-based layouts can be found according to the task at hand. For example, a common implementation is fish-eye lists, where the item selected expands to reveal more information. Such a solution works well for long lists, such as contacts or email messages. Another helpful solution to long lists is circular scrolling, where the list loops after reaching

its end. This is helpful because users of mobile phones cannot use the scrollbar as they would on a normal PC with a mouse.

The second challenge in mobile interaction design is data entry, specifically text entry. Text entry with a keypad is a notoriously tedious process and users avoid it as much as they can. Solutions to this problem were the use of autocomplete functions, predictive text, either dictionary based or pure predictive algorithms (which seem to perform better overall) (MacKenzie et al., 2001), and other novel solutions, such as gesture recognition, shapewriting (both solutions for touch screen– equipped devices), and voice recognition.

It should be noted however that, although these paradigms and styles regarding mobile interaction concern approximately 79% of devices (Entner, 2010), the new-generation smartphones featuring touch screens overcome some of those problems by allowing direct manipulation interaction. They are still affected by the obvious size differences, and many of the guidelines certainly apply, but the difference in the overall user experience is very notable. This is important because the market share of these devices is growing very rapidly, and it is not a wild speculation to assume that this trend will continue in the following years (Entner, 2010), especially since all major players, including Google, Microsoft, Sony Ericsson, and Nokia, as well as Apple of course, the first that made a huge market impact with the iPhone, are focusing on the development of such devices.

In essence, smartphones appear to be creating a paradigm shift in mobile interaction that can be compared to the paradigm shift of command line interfaces to the direct manipulation of GUIs. The old

paradigm is scroll and select using hardware buttons, and the new one is direct manipulation interfaces through touch screens.

The context of using a mobile device is inherently different from the use of a static personal computer. Mobile phones are primarily communication devices but also, since the merging of PDA capabilities, cameras and media players, as well as the introduction of mobile Web technologies such as 3G, they have become part of a larger ecosystem of networked devices and an important personal appliance for the user. Simply put, today the owner of a modern smartphone does a lot more than simply call or text another person. The smartphone is also used as a music player, a Web browser (which involves many of the issues we discussed in the Web interaction section), a global positioning system (GPS), a gaming device, and more, usually on the go.

Evaluating mobile device usage is particularly challenging because it is not possible to conduct studies in the laboratory as the context of use is entirely different. Mobile users are by definition usually on the move and, more importantly, interaction occurs infrequently and in bursts. Observing the interaction in real circumstances however is equally difficult (if not more), since there are too numerous environments and circumstances to take into account and there is no way to predict when the user will actually use a device. Furthermore, interaction is mostly private, so the act of observation would contaminate the results as there is no way to determine if the user has altered his or her behavior (Jones and Marsden, 2006).

Regarding the context of use of mobile Web interaction, it merits a closer examination because modern smartphones have rapidly changed its use.

### 4.6.1 Evolution of Mobile Web Interaction

Access to the Internet through a mobile phone has been available since 1996, when Nokia released the 9000 Communicator model. However, the technical limitations of mobile devices made browsing the Internet almost impossible, as Web pages did not render gracefully on the small screen or at all, making them unreadable. As an attempt to address this problem, the Wireless Application Protocol (WAP) was developed in 1998 as an open international standard, based on which WAP browsers could provide the basic services of a desktop computer browser but in a simplified form to overcome device limitations. Slow speeds, pricing, and the notably poorer experience compared to the familiar desktop browsing were the reasons why the WAP-based mobile Web did not really catch on, with the notable exception of Japan. In the latter, a rival system to WAP was developed and released in 1999 by NTT DoCoMo, the i-mode, which became a huge commercial success. Soon afterward, its two major rivals in the mobile market in Japan offered WAP-based services, with considerable success as well. Mobile Web use in Japan has been far more widespread than the rest of the world until very recently. There are several reasons for this, primarily favorable flat-rate plans, extremely high 3G handset penetration and excellent network quality and signal coverage, and the carriers' approach of "open garden" as opposed to Western carriers who try to keep consumers on their portals (Billich, 2010).

3G mobile Web speeds are considerably better, but still the experience is not even remotely similar to the desktop counterpart. The main reason is that, instead of trying to fit everything in the tiny space of a mobile device screen, designers began to customize and reduce the functionalities offered through the mobile sites to make the experience more appropriate with the context of use of mobile users. This also followed the specifications by the W3C Mobile Web Initiative,[*] which addressed the issue of balancing the homogenous nature of the Web and the specific circumstances of the mobile context of use.

Regarding the latter, Google identified three primary contexts of use for mobile consumers of its services:

- *The Casual Surfer or "Bored Now" User*. Users who find themselves with spare time (such as waiting in lines, while traveling by train, sitting in cafés, etc). These users resemble the casual Web surfers. Since mobile phones cannot match the robust user input of a desktop PC, applications for these users should be tailored.

- *The Repeat Visitor or "Repetitive Now" User*. Users who seek the same information on a regular basis, such as stock market prices, weather reports, and sports scores. Catering to their needs would be ensuring that repetitive steps or search queries can be eliminated by "remembering" each user's preferences, in the same manner that cookies work in desktop browsers.

- *The "Urgent, Now!" Visitor*. Users who seek specific information fast, such as directions to the airport or the nearest ATM. The key issue in this case being location, mobile services catering to such situations should emphasize location awareness.

## 4.7 Multimodal Interfaces

The multimodal systems process combined natural input modes, such as speech, pen, touch, hand gestures, eye gaze, and head and body movements, in a coordinated manner with multimedia system output (Oviatt, 1999). Multimodal systems represent a paradigm shift from conventional WIMP interfaces toward providing users with greater expressive power and naturalness. The goals are twofold: to achieve an interaction closer to natural human–human communication and to increase the robustness of the interaction by using redundant or complementary information in different modalities (Reeves et al., 2004).

One of the first multimodal systems was Bolt's Put That There System (Bolt, 1980), where the users interacted with the world through its projection on the wall using speech and pointing gestures. Subsequent attempts in the domain of multimodal interaction rely on advances in speech and natural language processing,

---

[*] http://www.w3.org/Mobile/

computer vision, and gesture analysis. Major progress has occurred in both the hardware and software for component technologies like speech, pen, and vision, as well as in the development architectural components and design frameworks (Oviatt and Cohen, 2000). Additionally, applications have been built that range from map-based and virtual reality systems for simulation and training, to field medic systems for mobile use in noisy environments, to Web-based transactions and standard text-editing applications. An overview of architectures and applications for multimodal interfaces is provided elsewhere (Oviatt et al., 2000)

Multimodal systems integrate complementary modalities to yield a highly synergistic blend in which the strengths of each mode are capitalized upon and used to overcome weaknesses in the other. Whereas traditional interfaces support sequential and unambiguous input from devices such as keyboard and conventional pointing devices (e.g., mouse, trackpad), multimodal interfaces relax these constraints. For example, they can support asynchronous, ambiguous, and inexact input by applying more sophisticated analysis of input. They can also detect and correct errors utilizing models of the media, user, discourse, and task (Maybury, 1999).

Systems that process multimodal input also aim to give users better tools for controlling embedded visualization and multimedia output capabilities, as opposed to the limited possibilities offered by keyboard and mouse input, in particular when dealing with complex environments (Oviatt, 1999).

In the context of multimodal interfaces, as the center of HCI shifts toward natural multimodal behavior, human communication patterns are used to control computers in a more transparent interface experience than ever before. Such interface designs become more conversational in style, rather than limited to command and control, because many of the modes being processed are language-oriented (speech, manual gestures, pen input) or involve communication broadly defined (gaze patterns, body movement) (Oviatt and Cohen, 2000).

Achieving natural patterns of multimodal input is however not as straightforward as it would appear. A dominant issue in this respect concerns the integration and synchronization requirements for combining different modalities into a system. Oviatt (1999), in the seminal paper "Ten Myths of Multimodal Interaction," on the basis of empirical evidence as well as experience, analyzes common pitfalls which may negatively impact multimodal interaction design. Examples of such assumptions are that users will always interact multimodally if they have the possibility to do so, that speak and point is the dominant modality integration pattern, that speech is the primary input mode in multimodal systems, and multimodal languages do not differ from unimodal languages. Regarding this last point, which presents particular interest from the point of view of interactivity, Oviatt claims that multimodal languages are briefer, syntactically simpler, and less disfluent than natural unimodal speech, as multimodality allows to eliminate linguistic complexity resulting from the need to express verbally elements to which the user can

refer deictically (through gestures) using a different modality. A related issue concerns redundancy between modalities.

Whereas it is commonly claimed that the content conveyed through different modalities in multimodal communication contains a high degree of redundancy, it appears that users tend to use different modalities in a complementary rather than redundant way.

To optimize human performance in multimodal systems, principles regarding how to integrate multiple modalities or how to support multiple user inputs (e.g., voice and gesture) have been elaborated on based on cognitive science literature on intersensory perception and intermodal coordination (Reeves et al., 2004).

Perceptual user interfaces (PUIs), introduced by Turk and Robertson (2000), are multimodal user interfaces which combine active input, such as speech, pen-based gestures, or other manual input with a passive input mode that requires no explicit user command, such as vision-based tracking that unobtrusively monitors user behavior and senses a user's presence, gaze, and/or body position. Through perceptual interfaces, multimodality permeates into intelligent interaction environments where interaction is to a large extent implicit and continuous (see Section 4.10).

## 4.8  Virtual and Augmented Reality

Virtual and augmented reality (VR and AR, respectively) are two areas that share many aspects. Their difference lies in the degree to which they replace (in the case of VR systems) or enhance/augment (AR) the real world. The similarity is that this is performed with the use of computer hardware and software. Practically, virtual and augmented reality systems have common research backgrounds, but the applications of each field are distinctly different. Both are in the stage where mass use of these technologies is not widespread and research is ongoing. Technological developments and costs are major factors for this, and the fluid nature of computing developments in general will determine the role these technologies will play in everyday life. Below, a brief historical account of both technologies is presented, describing the major features of each, in terms of hardware, software, user experience, and applications.

Virtual reality is a term that encompasses a broad range of research directions and applications. In the popular mind, virtual reality is the classic futuristic technology, being showcased in science fiction works such as Star Trek, where crew members of a spaceship can experience realistic worlds simulated by a sophisticated computer. A technical definition is offered in Heim (1998, p. 6):

> Virtual Reality is an immersive, interactive system based on computable information.

The term itself was first used by Jaron Lanier in 1986 (Behr, 2002), replacing various descriptions such as virtual worlds. The latter was used by Ivan Sutherland, who built what is considered the first head-mounted display (HMD), an early VR prototype called the Sword

of Damocles attempting to realize what he had called earlier "the ultimate display" (Sutherland, 1965). Other terms included synthetic enviroments, tele-existence, artificial reality, and immersive computing, the latter emphasizing a key feature of VR systems, namely immersion. Immersion refers to the feeling of being present in another reality apart from the real world. According to Heim, this goes beyond physical input and output because it involves psychological components, but it also surpasses purely mental imagination because of the sensory input involved.

A successful VR system should at least partially offer this user experience of being someplace else, in a reality created artificially. Total immersion is what most researchers refer to as strong VR.

Virtual reality differs significantly from the "regular" computer interaction in terms of input and output devices (see Figure 10). As the whole paradigm revolves around the immersive experience of being present in an artificial world, the way users interact with this world depends on these devices. The most important device is the medium that provides (primarily) the visual information that attempts to replace the physical world perceived through the eyes. There are two basic ways that are employed for this: HMDs and CAVE-type environments.

As the name suggests, HMDs are displays that fit on the user's head, often as a helmet, which block the view of the physical world and present the computer-generated images to replace it. The displays themselves are miniaturized and the technology used is usually CRT or liquid crystal display (LCD) types of screens. Issues that affect the quality of the user experience are the weight of the device, the resolution of the displays, and the field of view (FOV). The more sophisticated devices include a head-tracking system, which sends data to the computer system about the positioning of the user's head (and consequently gaze), so the image displayed is refreshed accordingly. This is essential to create an immersive experience, as the user freely moves his or her head about and views the scenery changing appropriately around him or her. This feature makes the HMD a significant input device as well. Some HMDs also include headphones to provide the audio. HMDs are also known as goggles and, together with data gloves, form the goggles-and-glove VR paradigm. The problems with HMDs, despite significant progress since the Sword of Damocles system, are still their cumbersome nature, suboptimal resolution, and limited FOV (Stanney and Cohn, 2008).

The data glove is an input device that is worn like a glove that tracks finger and hand movements and gestures and allows its wearer to manipulate virtual objects, thus bringing direct manipulation interactivity in a virtual environment. Its history can be traced back to 1977 with the Sayre Glove, which was an inexpensive, lightweight glove that was developed to track hand movements. Over the next years, this technology was enriched to include more sensors for tracking finger flexure and introducing tactile feedback to the fingertips, making the data glove an output device as well. Notable incarnations were the Digital Data Entry glove (1983), which was developed to recognize the Single Hand Manual Alphabet for the American Deaf and the 1987 Nintendo Power Glove, which was a crude data glove in terms of precision but it was the first that was widely available to the public (Sturman and Zeltzer, 1994). Data suits are the whole-body equivalent of the data glove. The user wears a suit that contains sensors that track the entire body movement. The data are then sent to the computer, which appropriately updates the virtual equivalent of the user, sometimes called a cyberbody.

CAVE-type environments are systems that are based on a configuration of large displays. They are rooms where usually projectors are employed to combine images to create a large scene. CAVE stands for Cave Automatic Virtual Environment, which was the name of the first such system, developed by the University of Illinois in 1992, the name also pointing to Plato's allegorical cave from *The Republic*. CAVE-type systems offer some significant advantages over HMDs. They allow multiuser immersion and collaboration, without



**Figure 10**   VR input and output devices.

cumbersome wearable devices, user mobility (which is only achievable through expensive omnidirectional treadmills in the case of HMD-based VR), high resolution, and wider FOV (Cruz-Neira et al., 1992).

While vision and hearing are the two senses that dominate interaction with computer systems, there are however efforts to exploit another human sense, touch. This is an area that interests particularly VR researches, since touch is a channel that provides rich information in real life, particularly in situations where VR is being used, such as health care. Doctors often feel with their hands bodily areas to make a diagnosis. To realize these sorts of tasks in a virtual space, tactile interfaces or haptics have been developed. These devices use pins or small electrical currents to stimulate the nerves that cause the sensation of touching. Another method used is by using motor mechanisms to provide resistance to the user's probings (Iwata, 2008).

Perhaps VR has failed to fullfill the expectations of enthusiastic fans as well as more moderate (as they seemed at the time) expectations of widespread VR use for the average user. VR is still very costly, so its main users remain limited to large business organizations or the military. But it has found a place in many applications (Heim, 1998).

Training is the most established application area for VR (Stanney and Cohn, 2008). The first VR applications were for simulation and training, especially in aviation and the military, and still today flight or combat simulations are typical examples of VR technology use.

Entertainment has always been a major force behind VR research and development. Arcade rooms often have a VR system that provides a much more exciting gaming experience than regular gaming systems. Disney is another example. Mine (2003) described work in creating VR attractions, where guests to Disney World can enjoy 4–5-min rides in CAVE-type systems, allowing a whole family to be simultaneously immersed in the game.

Another major consumer of VR technology is the building industry. It is quite common these days to construct 3D computer representations of building projects that are used not only to determine the design but also for showcasing and presenting the project to prospective customers, who can have a virtual walkthrough or fly-by in their future home. They can then offer feedback and express their preferences before any costly construction begins. One construction company specializing in luxury homes called their VR presentations "our greatest marketing tool."* The automotive, aerospace, naval, and other industries also have been exploiting VR in product design and manufacturing. Virtual wind tunnels are favored over their physical expensive counterparts. NASA was the first to combine the HMD and data glove technology in their VIEW (Virtual Interactive Environment Workstation) system, used among other things to develop a virtual wind tunnel and, according to Mark Pesce (2000), it was then that "Virtual Reality was

born," since it was one of the first instances that the users commented they actually felt being inside the virtual environment (tele-presence).

A recent VR application that has been utilized is in treating psychological conditions, namely phobias (Stanney and Cohn, 2008) and posttraumatic stress disorder (Rothbaum et al., 1999). While these treatments are fairly new and expensive, it is expected that they will become more common. Other areas of health care that benefit from VR is the treatment of patients who undergo painful procedures, such as skin scraping in burn victims. The immersive quality of VR actually distracts patients from the pain when they become absorbed in the VR program (Mueller, 2002).

Finally, VR has been used successfully in surgery education, where the surgeon examines and "operates" on 3D holographic images of a patient, obtained through normal magnetic resonance imaging (MRI) scans and computer tomographic (CT) images, before performing the procedure on the real patient (Versweyveld, 2001). This technique also takes advantage of force feedback, where the controls relate the pressure needed to be applied by the surgeon. VR has also been used in education and learning in various ways. One common example is in the reconstruction of places from history (Acevedo et al., 2001), where users can walk through sites that are either far away or have been destroyed. The exciting and entertaining aspect of VR definitely contributes to the learning experience.

## 4.9 Augmented Reality

Augmented reality began at roughly the same time as VR research, sharing a common origin and diverging in the application domain. As mentioned, AR does not seek to replace reality like VR does; instead it enhances (or augments) reality by superimposing computer-generated information over the primary visual field. In some cases, the physical view is processed to remove irrelevant data with the goal of enhancing not reality itself but the user's perception of it. The term itself is relatively recent, dating back to 1990. It was coined by Tom Caudell while he was working at Boeing. He used the term to describe a system utilizing a HMD to assist workers assembling cables into aircrafts (Mizell, 2001).

HMDs have been described above as the classic devices used in VR applications. In fact, HMDs are used extensively in AR applications, with the difference being that the displays employ semitransparent mirrors that do not block the physical view but allow computer-generated imagery to be projected on them. Alternatively, non-see-through displays can be used but they use real-time video of the outside world with superimposed graphics, effectively achieving the same result. They were first developed for use in fighter jet pilot helmets (and the H in the acronym actually stands for "helmet" instead of "head"), to replace heads-up displays (HUDs) (Defense Industry Daily, 2010), which in turn were one of the first applications of AR, long before the term was even thought of. HUDs were developed to eliminate the need for pilots to look down on their instruments (hence to keep their "heads up"). When the technology became available, the same idea was moved

---

* http://www.thefreelibrary.com/Builder+Spectrum+Skanska+using+'virtual+reality+tours'+to+sell+homes....-a097876792

into the pilots' helmet and integrated with the weapons systems. Like in VR HMDs, the helmet contains sensors to track head position, with the effect of the pilots' gaze being used to lock on enemy targets.

The field of AR really boomed in the 1990s and the past decade. Recent advances in technology, particularly in the capabilities and features of mobile devices and the fusion of different technologies in robotics and VR, have yielded a host of different applications with very promising prospects. Those are briefly summarized below. Before looking into them, however, it must be noted that AR is now more definitively defined, not as an offshoot of VR but as a particular field that blends computing and reality in real time with the ultimate goal of enhancing user perception and performance through this blending.

AR has been employed successfully in health care in various tasks, including surgery preparation, brain surgery, and laparoscopic surgery. In this context, images obtained through CT or MRI scans are projected directly on the patient's body, providing to surgeons the three-dimensional vision and feeling (Kania, 2001).

Another popular example of AR is Google Earth.* The user views satellite pictures of Earth and by setting preferences can superimpose additional information, such as borders between countries or states, roads, names of places, extra pictures submitted by users, and calculate distances between two geographical points etc.

Similarly, taking advantage of the recent advances in mobile devices, particularly the use of cameras and GPSs, there are applications for mobile devices that can display useful information over the video view of the world provided by the phone's camera. Examples include navigational directions for GPS applications or the display of points of interest in a particular area over the view the user points the camera at (see Figure 11). Such applications are expected to become more common

---

* http://www.google.com/earth/index.html

in the future as there is growing demand from the smartphone user base to exploit the capabilities of their devices (Chen, 2009).

## 4.10  Interaction in Intelligent Environments

AmI is an emerging technological paradigm which envisages an environment populated by several interoperating computing-embedded devices of different size and capabilities, which are interweaved into "the fabric of everyday life" and are indistinguishable from it (see Chapter 49 xx).

From a technological point of view, AmI targets to distribute, embed, coordinate, and interactively deliver computing intelligence within the surrounding environment. AmI technologies integrate sensing capabilities, processing power, reasoning mechanisms, networking facilities, applications and services, digital content, and actuating capabilities distributed in the surrounding environment. AmI will have profound consequences on the type, content, and functionality of the emerging products and services as well as on the way people will interact with them, bringing about multiple new requirements for the development of interactive technologies (e.g., Butz, 2010).

While a wide variety of different technologies is involved, the goal of AmI is to either hide the presence of technology from users or smoothly integrate it within the surrounding context as enhanced environment artifacts. This way, the computing-oriented connotation of technology essentially fades out or disappears in the environment, providing seamless and unobtrusive interaction paradigms. Therefore, people and their social situation, ranging from individuals to groups, and their corresponding environments (office buildings, homes, public spaces, etc) are at the center of the design considerations.

The pervasiveness of interaction in AmI environments requires the elaboration of new interaction concepts that extend beyond the current user interface



**Figure 11**  Augmented reality.

concepts like the desktop metaphor and menu-driven interfaces (Aarts and De Ruyter, 2009).

AmI environments will integrate a wide variety of interactive devices, in many cases equipped with built-in facilities for multimodal interaction and alternative input/output (e.g., voice recognition and synthesis, pen-based pointing devices, vibration alerting, touch screens, etc.) or with accessories that facilitate alternative ways of use (e.g., hands-free kits), thus addressing a wider range of user and context requirements than the traditional desktop computer. Devices will also vary in the type and specialization of the functionality they offer, ranging from "personal gadgets" (e.g., wristwatches, bracelets, personal mobile displays and notification systems, health monitors embedded in clothing) to "general-purpose appliances" (e.g., wall-mounted displays). Regarding personal devices, an important role will be played also by smart mobile phones, which already offer sensing and location awareness facilities, as well as AR applications which go in the direction of distributed and natural interactivity.

AmI will bring about new interaction techniques as well as novel uses and multimodal combinations of existing advanced techniques, such as, for example, gaze-based interaction (Gepner et al., 2007), gestures (Ferscha et al., 2007), and natural language (Zhou et al., 2007). Progress in computer vision approaches largely contributes to the provision of natural interaction in AmI environments, making available, among other things, techniques for facial expression, gaze and gesture recognition, face and body tracking, and activity recognition.

Additionally, interaction will be embedded in everyday objects and smart artifacts. This concept refers to interfaces that use physical artifacts as objects for representation and interaction, seamlessly integrating the physical and digital worlds. Such objects serve as specialized input devices that support physical manipulation, and their shape, color, orientation, and size may play a role in the interaction.

The interaction resulting from tangible user interfaces is not mediated and it supports direct engagement of the user with the environment. Consequently, it is considered more intuitive and natural than the current keyboard- and mouse-based interaction paradigm (Aarts and De Ruyter, 2009).

Interaction in AmI environments inherently relies on multimodal input, implying that it combines various user input modes, such as speech, pen, touch, manual gestures, gaze, and head and body movements as well as more than one output modes, primarily in the form of visual and auditory feedback. In this context, adaptive multimodality is prominent to support natural input in a dynamically changing context of use, adaptively offering to users the most appropriate and effective input forms at the current interaction context. Multimodal input is acknowledged for increasing interaction accuracy by reducing uncertainty of information through redundancy (Lopez-Cozar and Callejas, 2010).

However, AmI is also anticipated to introduce increased complexity for its users. As technology "disappears" to humans both physically and mentally, devices will no longer be perceived as computers but rather as augmented elements of the physical environment (Streitz, 2007). The nature of interaction in AmI environments will change radically, evolving from HCI to human–environment interaction (Streitz, 2007) and human–computer confluence (Ferscha et al., 2007). These concepts emphasize the fusion of the technology and the environment as well as the inextricable role of interaction in all aspects of everyday life.

AmI environments will be very interaction intensive, and humans will be constantly "surrounded" by a very large number of devices of different shapes and sizes. Therefore, interaction shifts from an explicit paradigm, in which the users' attention is on computing, toward an implicit paradigm, in which interfaces themselves drive human attention when required (Schmidt, 2005). Interaction in the emerging environment will be based no longer on a series of discrete steps but on a continuous input/output exchange of information (Faconti and Massink, 2001). Continuous interaction differs from discrete interaction since it takes place over a relatively longer period of time, in which the exchange of information between the user and the system occurs at a relatively high rate in real time. A first implication is that the system must be capable of dealing in real time with the distribution of input and output in the environment. This implies an understanding of the factors which influence the distribution and allocation of input and output resources in different situations for different individuals.

Due to the intrinsic characteristics of the new technological environment, it is likely that interaction will pose different perceptual and cognitive demands on humans compared to currently available technology (Gaggioli, 2005). It is therefore important to investigate how human perceptual and cognitive functions will be engaged in the emerging forms of interaction and how this will affect an individual's perceptual and cognitive space (e.g., emotion, vigilance, information processing, and memory). The main challenge in this respect is to identify and avoid forms of interaction which may lead to negative consequences such as confusion, cognitive overload, frustration, and so on. This is particularly important given the pervasive impact of the new environment on all types of everyday activities and on the way of living.

### 4.11 Summary of Interaction Paradigms

Command-based interaction (Section 4.2), most prominently (but not exclusively) demonstrated by command line interfaces, was the first major interaction style in HCI. It is still widely used, often preferred by expert users for its high speed and flexible nature and complements very well direct manipulation interfaces. The keyboard remains the most efficient text entry medium, so it is unlikely that it will go away, and as a result, the same is true for command line interaction.

Direct manipulation interaction (Section 4.3) encompasses a wide array of styles, starting from WIMP interfaces to tangible and haptic interfaces and VR interaction styles. The characteristic of direct manipulation interfaces is that it allows users to perform actions on objects directly, either through a pointing device (like

in a WIMP interface) or through any other way, such as a gesture or directly by hand (e.g., in touch screens). It is the most widely used interaction style, specifically in the form of the GUI paradigm of the desktop metaphor that is found almost universally across desktop computers. Direct manipulation interfaces are also featured in new technologies that are growing in the market, such as tablet PCs and smartphones featuring a touch screen.

Natural language and speech dialogue (Section 4.4) remains an unrealized promise in HCI, at least regarding the vision of natural interaction for all users, who seem to prefer an all-natural-language interaction or nothing. However, it has provided useful task-specific applications that have found a place in interaction with computer systems, for example, situations where the hands are busy or for people with disabilities. It is also an important input mode in multimodal systems, which complement speech input with gestures or some other input mode to overcome strict NLP limitations. It should be expected that the speech channel as input and output will be an important part of multimodal interaction in intelligent environments.

Web interaction (Section 4.5) is unique as context of use, which is a vital parameter in HCI, while the physical interaction itself in Web use does not present a specifically different paradigm from conventional application use. The Web information space and the choices offered, as well as the technology that opens vast communication capabilities, such as wireless connectivity, make Web interaction (along with mobile interaction perhaps) the most important evolution in the way people interact with computers.

Mobile interaction (Section 4.6) brought miniaturization and mobility into the fold. Starting from noninteractive tasks such as listening to music (with the Walkman) and branching into much more interactive activities such as making phone calls and surfing the Web, mobile interaction is a major step in HCI that has opened many new possibilities, including the role of the device in ubiquitous computing. Mobile interaction is characterized (and constrained) by the physical size of the device and the context of use, which varies significantly from desktop interaction. At this time, mobile interaction is undergoing a noticeable paradigm shift from scroll and select interfaces, dependent on the mobile phone keypad interface, to the more sophisticated direct manipulation interface offered by the latest smartphones.

Virtual reality (Section 4.8) is a special case of HCI, distinguished by the goal of providing an immersive experience to the user by replacing the real word with a computer-generated environment. This is achieved primarily with the use of HMDs, which isolate the user's visual perception of the real world and provide an artificial alternative and with CAVE-type environments, which use large displays that dominate the user's vision and allow for the immersive experience. How the user interacts in the various VR environments differs significantly, dependent of the nature of the VR application. For example, a training VR application mimics the real-world controls of interaction (e.g., the plane's cockpit) but a virtual environment that has no real-world equivalent (such as a fantasy world) could use

a direct manipulation interface with a "magic wand," a tangible interface, and so on. Due to the high cost of VR, its applications are still limited to the military or large business organizations and widespread adoption of this interactive paradigm seems difficult and unlikely to become a reality in the near future.

Augmented reality (Section 4.9) blends real-world information with computer-generated visualization to facilitate user perception and user tasks. The main interactive element that distinguishes AR interaction is in the form of visual feedback provided by the computer. Unlike VR, simple AR applications (such as GPS-based navigation) are starting to penetrate the market, mainly through the exploitation of the capabilities of modern smartphones. It is expected that with the rapid growth of the market share presented by these devices more applications of this sort will be quite common. Otherwise, AR is successfully being employed in health care.

Multimodal interaction (Section 4.9) integrates natural input modes, such as speech, pen, touch, hand gestures, eye gaze, and head and body movements. It represents a paradigm shift from conventional WIMP interfaces toward providing users with greater expressive power and naturalness and achieving forms of interaction closer to natural human–human communication. Recent advances in multimodal interaction also cater for passive input that requires no explicit user command, such as vision-based tracking. This way, multimodality permeates into intelligent interaction environments where interaction is to a large extent implicit and continuous.

AmI (Section 4.10) is arguably the next major step in the overall computing paradigm and consequently the next major evolutionary step in HCI. What is fundamentally changing is that users interact no longer with a single machine but with the environment surrounding them, the computer becoming a more abstract entity, fusing intelligent artifacts through an invisible network of sensors. Interaction does not occur through a single interface and can be achieved via a multimodal paradigm that can include any of the interaction styles described in this chapter, such as gestures, speech, direct manipulation, and combinations of those. At the moment, AmI environments are largely in the research stage, since they present a host of challenges that need to be overcome.

## 5  EMERGING CHALLENGES, FUTURE TRENDS, AND CONCLUSIONS

There are several trends and challenges that determine the direction of future HCI research and development and as a consequence the future evolution of interactivity. These include trends and developments in mobile interaction, concerns about universal access, multitouch-based devices, the evolution of Web interaction, and the progressive emergence of AmI environments.

Mobile Web use has evolved into a totally different paradigm from the desktop counterpart. Similarities can be drawn with some add-ons or plug-ins of Web browsers, which offer specific Web services rather than

a free net roaming experience. Examples would include a toolbar which displays the weather or stock market prices or such. Mobile users download Web applications that provide specific services as well. For example, Apple's App Store has thousands of applications that use the Web, each with its own interface, without any particular uniform design paradigm (still in its infancy) and users consume bits of information from the Web without really noticing it. Web usage is therefore spread over different small applications, from different manufacturers and providers. The competition is no longer on the website level but on the miniapplication level. There is no homogenous paradigm of Web usage, the mobile Web is "balkanized," divided into thousands of service points with thousands of interfaces. Search is not the main issue, as it is on desktop Web surfing. According to Google user experience designer Leland Rechis, "The Pangaea of the Web is gone" (Wellman, 2007). Because of carrier portals and off-portal applications, there is no one mobile standard to develop for. In the mobile world developers have to be prepared to optimize for different devices, browsers, languages, carriers, countries, and cultures. While the international and cultural factors affect desktop Web surfing as well (Marcus and Rau, 2009) and have been examined quite well, the portal and device variety remain the most challenging issue in mobile development.

An open question regarding the future of mobile devices is if there will be a convergence of devices, that is, devices that combine the capabilities of more or an increased diversity of appliances. Most mobile phones combine the capabilities of a telephone, a camera (including video recording), an electronic organizer, and more often than not a music device. Taking into account the recent trend and market acceptance of high-end touch-based phones that offer these features, as well as Internet connectivity, in a considerably more user-friendly manner, it seems that the convergence of devices case has more merit as the safest prediction, at least on the mobile phone–size scale.

This potential of mobile devices in general needs further investigation. There are many ongoing research projects that explore the mobile device as an ubiquitous computing medium, an AR device, new interaction technique opportunities as presented by high-end phones with touch screens, accelerometers and gyroscopes, cameras, and so on.

Another important concern that influences HCI is the rapid aging of the population in the developed world. This has not only socioeconomic but also accessibility implications in (especially mobile) computer interaction. Mobile devices are particularly susceptible to accessibility issues because of their inherent small size. Furthermore, mobile interaction is challenged by situational impairments and disabilities (SIIDs), due to the plethora of the contexts of use. A major challenge therefore in future HCI research is to provide a solution to the accessibility problem regarding mobile use. Universal access is a constant challenge in HCI, but desktop computing (for which most related work has been done) presents more alternative solutions and is, arguably, easier to respond to compared to mobile interaction.

The design approach of the new smartphones has been recently transferred into the tablet PC platform. While tablet PCs have been around for some time and did feature a touch screen, these devices are not merely (or, arguably, fully) PCs with a touch screen. The Apple iPad, for example, does not offer the same capabilities as a normal netbook or laptop, such as multitasking and running any type of application. Instead it is designed to offer specific services, such as multimedia, electronic reading, and connection to the Web. A major difference is the multitouch interface, which bears significant advantages over older touch interfaces, as it offers a much richer and enjoyable user experience. This type of appliance, despite its obvious shortcomings compared with a standard (and much cheaper, for now at least) netbook, seems likely to become in the future the computing interaction of choice for a significant number of users. Its multitouch interface, enjoyable user experience, and inherent mobility are a strong combination that covers many market needs. Moreover, there is a conscious effort to make smartphones and tablet PCs and information sharing between them easy. This, it can be argued, is a first step toward a more ubiquitous computing that is not constricted into one device or location. While there are still major difficulties to overcome, this trend is anticipated to move closer to a situation where abstract cloud computing is available and compatible through a host of different device types and platforms.

AR in general is expected to be more available and accessible to the everyday user, through smartphone cameras, the aforementioned tablet PCs (also equipped with HD cameras) and other promising projects, such as SixthSense,* which is a wearable device that blends the digital world with the physical.

Interacting with the Web has changed significantly since the first days of the ARPANET. The text-based presentation of the first era was mostly confined in academia and business. The wider public became familiar with the Web through GUI-based browsers, which soon were capable of providing a richer interactive experience. Today, the Web is accessible by far more devices and in a plethora of contexts of use that is significantly altering the overall user experience. Service-specific information consumption in chunks, which characterizes mobile interaction, is also affecting desktop and laptop Web interaction. An example is RSS feeds or browser plug-in (add-ons, extensions) services. What this means in essence is that Web interaction is now expanded to include more than the usual browser-based surfing style that dominated Web interaction for more than 15 years.

Gesture-based interaction, mostly familiar to the wider audience through gaming (the Wii console is a particular example†), is also being researched in the context of AmI environments and 3D interfaces (in VR applications or information visualization and navigation of large data sets, e.g., see Underkoffler's g-Speak SOE presented at TED in 2010‡) and is gaining more attention.

---

* http://www.ted.com/talks/pranav_mistry_the_thrilling _potential_of_sixthsense_technology.html

† http://www.nintendo.com/wii/console/controllers

‡ http://www.ted.com/talks/lang/eng/john_underkoffler_drive _3d_data_with_a_gesture.html

Finally, looking at the big picture and considering the technology available, it seems that the so-called third wave of computing mentioned earlier in the chapter, where the computer will gradually disappear in the background and be embedded in the surrounding environment, is approaching. AmI enviroments may take many forms. For example, in Chapter 49xx, various applications in different domains are discussed, where each installation differs in the actual size of the environment, the hardware and software involved, the interaction styles employed, and so on. Similarly, various ongoing research projects present different approaches and manifestations of the same idea of the disappearing computer, with different emphasis on the intelligent behavior, the natural interaction involved, accessibility, aesthetics, design methodology, and so on, depending on the specific domain.

Looking at the evolutionary history of interaction, the first observation is that none of the interaction styles that appeared from the first days of computing until today have actually disappeared. The oldest and most basic interaction style, command line interface, is still widely used and the same holds true for every interaction style presented in this chapter. Following the ongoing research and development of new interaction styles, such as gesture interfaces or haptic interfaces, there is no evidence that these will render any of the current popular styles obsolete. The conclusion that can be drawn is that there is no one optimal interaction method that can be devised. On the contrary, the paradigm shift toward an AmI environment that seeks to replace current methods of interaction with computing favors a multimodal approach, where every interaction style is utilized according to its appropriateness to the context of use and the task at hand. Therefore, in the AmI context, an important challenge is to approach every interaction style and determine its optimal use and appropriate place in the services offered by the intelligent environments.

More pragmatically, a huge challenge involved in AmI environments escaping the laboratory and entering the public is the complexity inherent in such a large scale. There are issues regarding integration, interoperability, and synergy between various devices and artifacts. This also includes security and privacy concerns, increased with the presence of more than one platform and devices that must all conform to the same high standards of safety. Not to be underestimated is the cost of such an environment, including the cost of use (parallels may be drawn by the Western/Japanse contrast of mobile Web use adoption).

The technology to realize ubiquitous computing exists. Theoretically and to a certain degree practically, the wide availability of the Internet everywhere (through Wi-Fi or mobile access) and the corresponding network-capable devices can form an environment in which they can communicate and exchange information to offer ubiquitous services. The question is whether (or more to the point, when) the interaction we experience with these components individually will expand to include all in a loosely perceived abstract supersystem that intelligently observes, adapts, and responds to human needs and desires.

## REFERENCES

Aarts, E., and de Ruyter, B. (2009), "New Research Perspectives on Ambient Intelligence," *Journal of Ambient Intelligence and Smart Environments*, Vol. 1, No. 1, pp. 5–14.

Acevedo, D., Vote, E., Laidlaw, D. H., and Martha S. Joukowsky, S. (2001), "Archaeological Data Visualization in VR: Analysis of Lamp Finds at the Great Temple of Petra, a Case Study," in *Proceedings of IEEE Visualization 2001*, San Diego, CA.

Allen, J. F., Miller, B. W., Ringger, E. K., and Sikorski, T. (1995), "A Robust System for Natural Spoken Dialogue," in *Proceedings of the 1996 Annual Meeting of the Association for Computational Linguistics* (*ACL'96*), pp. 62–70.

Anderson, P. A., and Guerrero, L. K. (1998), *Handbook of Communication and Emotion*, Academic, Academic Press, San Diego, CA.

Barker, P. (2000), "Human Communication Processes," in *Iconic Communication*, M. Yazdani and P. Barker, Eds., Intellect Books, Bristol, UK, pp. 1–16.

Behr, M. E. (2002, February 11), "Interview: Jaron Lanier, 'Virtual Reality' Inventor," ExtremeTech.com, p. 1.

Billich, C. (2010), "The Insider's Guide to Mobile Web Marketing in Japan," available: mobiThinking, http://mobithinking.com/guide-mobile-Web-Japan.

Bolt, R.A. (1980), "Put-That-There: Voice and Gesture at the Graphic Interface," *Computer Graphics*, Vol. 14, No. 3, pp. 262–270.

Bunt, H., and Beun, R., Eds. (1998), *Cooperative Multimodal Communication*, Springer, New York.

Burgoon, J. K., Bonito, J. A., Bengtsson, B., Cederberg, C., Lundeberg, M., and Allspach L. (2000), "Interactivity in Human-Computer Interaction: A Study of Credibility, Understanding, and Influence," *Computers in Human Behavior*, Vol. 16, No. 6, pp. 553–574.

Butz, A. (2010), "User Interfaces and HCI for Ambient Intelligence and Smart Environments," in *Handbook of Ambient Intelligence and Smart Environments*, H. Nakashima, H. Aghajan and J. C. Augusto, Eds., Springer, Berlin, pp. 535–558.

Byrne, M. D. (2008), "Cognitive Architecture," in J. A. Jacko and A. Sears, Eds., *The Human-Computer Interaction Handbook*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 93–113.

Cai, D. A., and Rodríguez, J. I. (1996), "Adjusting to Cultural Differences: The Intercultural Adaptation Model," Intercultural Communication Studies VI:2 1996-7 D, available: http://www.trinity.edu/org/ics/ICS%20Issues/ICS%20VI/ICS-VI-2-%20Cai.pdf.

Card, S. K., Moran, T. P., and Newell, A. (1983), *The Psychology of Human-Computer Interaction*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Cassell, J. (2001), "Embodied Conversational Agents: Representation and Intelligence in User Interfaces," *AI Magazine*, Vol. 22, No. 4, pp. 67–84.

Cassidy, J. (2003), *dot.con: How America Lost Its Mind in the Internet Era*, Harper Perennial, New York.

Chapanis, A. (1988), "Interactive Human Communication," in *Computer-Supported Cooperative Work: A Book of Readings*, I. Greif, Ed., Morgan Kaufmann, San Francisco, CA, pp. 127–141.

Chen, B. X. (2009), "If You're Not Seeing Data, You're Not Seeing," *Wired*, available: http://www.wired.com/gadgetlab/2009/08/augmented-reality/.

Chou, C. (2003), "Interactivity and Interactive Functions in Web-Based Learning Systems," *Bristish Journal of Educational Technology*, Vol. 34, No. 3, pp. 265–279, available: http://www.sarasotaintranet.usf.edu/ir/Documents/DistanceLearning/chou.pdf.

Clark, H. (1996), *Using Language*, Cambridge University Press, Cambridge.

Clark, H., and Fox Tree, J. (2002), "Using 'Uh' and 'Um' in Spontaneous Speech," *Cognition*, Vol. 84, pp. 73–111.

Clark, H., and Krych, M. A. (2004), "Speaking While Monitoring Addressees for Understanding," *Journal of Memory and Language*, Vol. 50, pp. 62–81.

Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R. V., and Hart, J. C. (1992, June), "The CAVE: Audio Visual Experience Automatic Virtual Environment," *Communications of the ACM*, Vol. 35, No. 6, pp. 64–72

Defense Industry Daily (2010), "JHMCS: Fighter Pilot 'Look & Shoot' Helmets' Upgrade, Ups & Downs," available: http://www.defenseindustrydaily.com/fighter-pilot-look-shoot-helmets-upgrade-ups-downs-01271/.

Downes, E. J., and McMillan, S. J. (2000), "Defining Interactivity—A Qualitative Identification of Key Dimensions," New Media & Society, Vol. 2, No. 2, pp.157–179.

Engelbart, D. C. (1963), "A Conceptual Framework for the Augmentation of Man's Intellect," in P. Howerton and D. Weeks, Eds., *Vistas in Information Handling*, Vol. 1, Spartan Books, Washington, DC, pp. 1–29.

Engelbart D. C., and English, W. K. (1968), "A Research Center for Augmenting Human Intellect," in *AFIPS Conference Proceedings of the 1968 Fall Joint Computer Conference*, San Francisco, CA, Vol. 33, December 1968, pp. 395–410

Entner, R. (2010), "Smartphones to Overtake Feature Phones in U.S. by 2011," *nielsenwire*, The Nielsen Company:

Faconti, G., and Massink, M. (2001), "Continuous Interaction with Computers: Issues and Requirements," in *Universal Access in HCI: Towards an Information Society for All*, Vol. 3 of *Proceedings of HCI International 2001*, New Or-leans, LA, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 301–304.

Ferscha, A., Resmerita, S., and Holzmann, C. (2007), "Human Computer Confluence," in *Universal Access in Ambient Intelligence Environments, Proceedings of the 9th ERCIM Workshop on User Interfaces for All*, C. Stephanidis and M. Pieper, Eds., Königswinter, Germany, September 27–28, 2006, Springer, Berlin, pp. 14–27.

Flaherty, L. M., Pearce, K. J., and Rubin, R. B. (1998), "Internet and Face-to-Face Communication: Not Functional Alternatives," *Communication Quarterly*, Vol. 46, pp. 250–266.

Friedl, M. (2002), *Online Game Interactivity Theory with Cdrom*, Charles River Media, Newton Center, MA.

Gaggioli, A. (2005), "Optimal Experience in Ambient Intelligence," in *Ambient Intelligence*, G. Riva, F. Vatalaro, F. Davide, and M. Alcañiz, Eds., IOS Press, Amsterdam, The Netherlands, pp. 35–43.

Gentner, D., and Stevens, A., Eds. (1983), *Mental Models*, Lawrence Erlbaum Associates, Mahwah, NJ.

Gepner, D., Simonin, J., and Carbonell, N. (2007), "Gaze as a Supplementary Modality for Interacting with Ambient Intelligence Environments," in *Proceedings of 12th International Conference on Human-Computer Interaction* (*HCI International 2007*), Beijing, China, July 22–27 2007, Lecture Notes in Computer Science, Springer, Berlin.

Glass, J. (1999), "Challenges for Spoken Dialogue Systems," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding ASRU*, available: http://groups.csail.mit.edu/sls/archives/root/publications/1999/asru99-jrg.pdf.

Grudin, J. (2008), "A Moving Target: The Evolution of Human-Computer Interaction, in *The Human-Computer Interaction Handbook—Fundamentals, Evolving Technologies, and Emerging Applications*, 2nd ed., A. Sears and J. A. Jacko, Eds., Lawrence Erlbaum, Mahwah, NJ.

Ha, L., and James, E. L. (1998), "Interactivity Reexamined: A Baseline Analysis of Early Business Web Sites," *Journal of Broadcasting and Electronic Media*, Vol. 42, No. 4, pp. 457–474.

Hafner, K. (1998), *Where Wizards Stay Up Late: The Origins of the Internet*, Simon & Schuster, New York.

Heeter, C. (1989), "Implications of New Interactive Technologies for Conceptualizing Communication," in *Media Use in the Information Age: Emerging Patterns of Adoption and Consumer Use*, J. L. Salvaggio and J. Bryant, Eds., Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 217–235.

Heim, M. (1998), *Virtual Realism*, Oxford University Press, New York.

Hill, I. (1983), "Natural Language Versus Computer Language," in *Designing for Human-Computer Communication*, M. Sime and M. Coombs, Eds., Academic, New York.

Hockett, C.F. (1960), "The Origin of Speech," *Scientific American*, Vol. 203, pp. 88–96.

Hutchins, E. L., Hollan, J. D., and Norman, D. A. (1986), "Direct Manipulation Interfaces," in *User Centred System Design*, D. A Norman and S. W. Draper, Eds., Lawrence Erlbaum Associates, Hillside, NJ, pp. 87–124.

International Organization for Standardization (ISO) (2010), "Ergonomics of Human-System Interaction—Part 210: Human-Centred Design for Interactive Systems," ISO 9241-210:2010, ISO, Geneva.

Iwata, H. (2008), "Haptic Interface," in *The Human-Computer Interaction Handbook*, J. A. Jacko and A. Sears, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 229–245,

Jacob, P. S., and Rau, L. F. (1988), "Natural Language Techniques for Intelligent Information Retrieval," in *Proceedings of the 11th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (*Grenoble, France*), Y. Chiaramella, Ed., SIGIR '88, Association for Computing Machinery, New York, pp. 85–99.

Jensen, J. F. (1999), "The Concept of 'Interactivity' in Interactive Television," in *Interactive Television: TV of the Future or the Future of TV?* J. F. Jensen and C. Toscan, Eds., Aalborg University Press, Aalborg.

Jenson, S. (2002), *The Simplicity Shift: Innovative Design Tactics in a Corporate World*, Cambridge University Press.

John, B. E., and Kieras, D. E. (1996), "Using GOMS for User Interface Design and Evaluation: Which Technique?," *ACM Transactions on Computer-Human Interaction*, Vol. 3, pp. 287–319.

Jokinen, K. (2009), "Natural Language and Dialogue Interfaces," in *The Universal Access Handbook*, C. Stephanidis, Ed., CRC Press, Taylor & Francis Group, New York, pp. 495–506.

Jones, M., and Marsden, G. (2006), *Mobile Interaction Design*, Wiley, Hoboken, NJ.

Kania, K. (2001), "Virtual Reality Moves into the Medical Mainstream," *Medical Device & Diagnostic Industry Magazine*, pp. 66–79.

Kay, A. C. (1972), *A Personal Computer for Children of All Ages*, Xerox Palo Alto Research Center, Palo Alto, CA.

Kurosu, M., and Kashimura, K. (1995), "Apparent Usability vs Inherent Usability: Experimental Analysis on the Determinannts of the Apparent Usability," Conference Companion On Human Factors In Computing Systems, Denver, CO, May 7–11, pp. 292–293.

Lakoff, R. (1973), "The Logic of Politeness: Or Minding Your P's and Q's," in *Papers from the Ninth Regional Meeting of the Chicago Linguistic Society*, C. Corum, T. C. Smith-Stark, and A. Weiser A., Eds., Chicago Linguistic Society, Chicago, IL, pp. 292–305.

Liu, Y. (2003), "Developing a Scale to Measure the Interactivity of Websites," *Journal of Advertising Research*, Vol. 43, p. 207, available: http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=169835&fulltextType=RA&fileId=S0021849903030204.

Liu, Y., and Shrum, L. J. (2002), "What Is Interactivity and Is It Always Such a Good Thing? Implications of Definition, Person, and Situation for the Influence of Interactivity on Advertising Effectiveness," *Journal of Advertising*, Vol. 31, No. 4, pp. 53–64.

Lopez-Cozar, R., and Callejas, Z. (2010), "Multimodal Dialogue for Ambient Intelligence and Smart Environments," in *Handbook of Ambient Intelligence and Smart Environments*, H. Nakashima, H. Aghajan, and J. C. Augusto, Eds., Springer, Berlin, pp. 559–579.

MacKenzie, I. S., Kober, H., Smith, D., Jones, T., and Skepner, E. (2001), "LetterWise: Prefix-Based Disambiguation for Mobile Text Input," in *Proceedings of the ACM Symposium on User Interface Software and Technology—UIST 2001*, Association for Computing Machinery, New York, pp. 111–120.

Manaris, B. (1998), "Natural Language Processing: A Human-Computer Interaction Perspective," in *Advances in Computers*, Vol. 47, Marvin V. Zelkowitz, Ed., Academic, New York, pp. 1–66.

Manaris, B. (1994), "An Engineering Environment for Natural Language Interfaces to Interactive Computer Systems," *International Journal of Artificial Intelligence Tools*, Vol. 3, No. 4, pp. 557–579.

Marcus, A., and Rau, P-L. P. (2009), "International and Intercultural User Interfaces," in *The Universal Access Handbook*, C. Stephanidis, Ed., CRC Press, Taylor and Francis Group, New York.

Maybury, M. (1999), "Intelligent User Interfaces: An Introduction," in *Proceedings of the 4th international Conference on Intelligent User interfaces*, Los Angeles, CA, January 5–8, 1999, IUI '99, Association for Computing Machinery, New York, pp. 3–4.

McCarthy, J. (1988), "B. P. Bloomfield, the Question of Artificial Intelligence: Philosophical and Sociological Perspectives," *Annals of the History of Computing*, Vol. 10, No. 3, pp. 224–229.

Mine, M. (2003), "Towards Virtual Reality for the Masses: 10 Years of Research at Disney's VR Studio," in *Proceedings of the Workshop on Virtual Environments 2003*, Zurich, Switzerland, May 22–23, 2003, EGVE '03, Vol. 39, Association for Computing Machinery, New York, pp. 11–17.

Mizell, D. (2001), "Boeing's Wire Bundle Assembly Project," in *Fundamentals of Wearable Computers and Augumented Reality*, W. Barfield and T. Caudell, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 447–467.

Morville, P. (2010), "Ubiquitous Service Design, Semantic Studios," available: http://semanticstudios.com/publications/semantics/000633.php,

Mueller, M. (2002, January), "How Virtual Reality Can Help You," *Current Health*, Vol. 2, p. 16.

Nielsen Company, "Smartphones to Overtake Feature Phones in U.S. by 2011," available: http://blog.nielsen.com/nielsenwire/consumer/smartphones-to-overtake-feature-phones-in-u-s-by-2011/.

Nielsen, J. (1997), "The Difference between Web Design and GUI Design, Alertbox, May 1," available: http://www.useit.com/alertbox/9705a.html.

Nielsen-Norman Group, "User Experience—Our Definition," available: http://nngroup.com/about/userexperience.html.

Norman, D. A. (1986), "Cognitive Engineering," in *User Centered System Design*, D. A. Norman and S. W. Draper, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 31–61.

Norman, D. A. (1990), *The Psychology of Everyday Things*, Doubleday, New York.

Norman, D. A. (2002), "Emotion and Design: Atrractive Things Work Better," *Interactions Magazine*, Vol. ix, No. 4, pp. 36–42.

Norman, D. A. (2004), *Emotional Design: Why We Love (or Hate) Everyday Things*, Basic Books, New York.

Norman D. A., and Draper, S. W., Eds. (1986), *User Centred System Design*, Lawrence Erlbaum Associates, Hillside, NJ.

Norman, D. A., Miller, J., and Henderson A. (1995), "What You See, Some of What's in the Future, and How We Go About Doing It: HI at Apple Computer," Conference Companion on Human Factors in Computing Systems, Denver, CO, May 7–11, 1995, p. 155,

Ogden, W.C., and Bernick, P. (1997), "Using Natural Language Interfaces," in *Handbook of Human Computer Interaction*, 2nd ed., M. Helander, T. Landauer, and P. Prabhu, Eds., North Holland, Amsterdam.

O'Reilly, T. (2005), "What Is Web 2.0—Design Patterns and Business Models for the Next Generation of Software," available: http://oreilly.com/web2/archive/what-is-web-20.html.

Oviatt, S. (1999, November), "Ten Myths of Multimodal Interaction," *Communications of the ACM*, Vol. 42, No. 11, pp. 74–81.

Oviatt, S., and Cohen, P. (2000, March), "Perceptual User Interfaces: Multimodal Interfaces That Process What Comes Naturally," *Communications of the ACM*, Vol. 43, No. 3, pp. 45–53.

Oviatt, S., Cohen, P., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., and Ferro, D. (2000), "Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future

Research Directions," *Human-Computer Interaction*, Vol. 15, No. 4, pp. 263–322.

Pennebaker, J. W., and King, L. A. (1999), "Linguistic Styles: Language Use as an Individual Difference," *Journal of Personality and Social Psychology*, Vol. 77, pp. 1296–1312.

Pesce, M. (2000), *The Playful World: How Technology Is Transforming Our Imagination*, Ballantine Books, New York, p. 178.

Petukhova, V. V., and Bunt, H. C. (2009), "Dimensions of Communication, TiCC TR 2009–003," Tilburg Centre for Creative Computing, Tilburg University, available: http://www.tilburguniversity.nl/faculties/humanities/ticc/research/technicalreports/TR2009003.pdf.

Rafaeli, S., and F. Sudweeks (1997), "Networked Interactivity," *Journal of Computer Mediated Communication*, Vol. 2, No. 4.

Raymond, E. S. (2000), "The Cathedral and the Bazaar," available: http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/index.html.

Reeves, L. M., Lai, J., Larson, J. A., Oviatt, S., Balaji, T.S., Buisine, S., Collings, P., Cohen, P., Kraal, B., Martin, J.-C., McTear, M. Raman, TV, Stanney, K. M., Su, H., and Wang, Q. (2004, January), "Guidelines for Multimodal User Interface Design," *Communications of the ACM*, Vol. 47, No. 1, pp. 57–59.

Rice, R. E. (1984), "New Media Technology: Growth and Integration," in *The New Media: Communication, Research, and Technology*, R.E. Rice, Ed., Sage, Beverly Hills, CA, pp. 33–54.

Rothbaum, B. O., Hodges, L., Alarcon, R., Ready, D., Shahar, F., Graap, K., Pair, J., Hebert, P., Gotz, D., Wills, B., and Baltzell, D. (1999), "Virtual Reality Exposure Therapy for PTSD Vietnam Veterans: A Case Study," *Journal of Traumatic Stress*, Vol. 12, pp. 263–271

Royce, W. (1970), "Managing the Development of Large Software Systems," in *Proceedings of IEEE WESCON*, August 26, pp. 1–9, available: http://www.cs.umd.edu/class/spring2003/cmsc838p/Process/waterfall.pdf.

Schmidt, A. (2005), "Interactive Context-Aware Systems Interacting with Ambient Intelligence," in *Ambient Intelligence*, G. Riva, F. Vatalaro, F. Davide, and M. Alcañiz, Eds., IOS Press, Amsterdam, The Netherlands, pp. 159–178.

Schneiderman, B. (1983), "Direct Manipulation: A Step Beyond Programming Languages," *IEEE Computer*, Vol. 16, pp. 57–63.

Schwier, R., and Misanchuk, E. (1993), *Interactive Multimedia Instruction*, Educational Technology Publications, Englewood Cliffs, NJ.

Shannon, C. E., and Weaver, W. (1949), *A Mathematical Model of Communication*, University of Illinois Press, Urbana, IL.

Shneidermann, B. (1998), *Designing the User Interface: Strategies for Effective Human–Computer Interaction*, 3rd ed., Addison-Wesley, Reading, MA.

Slator, B., Anderson, M., and Conley, W. (1986), "Pygmalion at the Interface," *Communications of the ACM*, Vol. 29, No. 7, pp. 599–604.

Stanney, K. M., and Cohn, J.V. (2008), "Virtual Environments," in *The Human-Computer Interaction Handbook*, J. A. Jacko and A. Sears, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 621–634.

Stephanidis, C. (2001), "Adaptive Techniques for Universal Access," *User Modelling and User Adapted Interaction International Journal*, Vol. 11, No. 1/2, pp. 159–179.

Stephanidis C., Salvendy, G., Akoumianakis, D., Bevan, N., Brewer, J., Emiliani, P.L., Galetsas, A., Haataja, S., Iakovidis, I., Jacko, J., Jenkins, P., Karshmer, A., Korn, P., Marcus, A., Murphy, H., Stary, C., Vanderheiden, G., Weber, G., and Ziegler, J. (1998), "Toward an Information Society for All: An International R&D Agenda," *International Journal of Human-Computer Interaction*, Vol. 10, No. 2, pp. 107–134.

Stephanidis, C. (Ed.), Salvendy, G., Akoumianakis, D., Arnold, A., Bevan, N., Dardailler, D., Emiliani, P.L., Iakovidis, I., Jenkins, P., Karshmer, A., Korn, P., Marcus, A., Murphy, H., Oppermann, C., Stary, C., Tamura, H., Tscheligi, M., Ueda, H., Weber, G., & Ziegler, J. (1999), "Toward an Information Society for All: HCI Challenges and R&D Recommendations, International Journal of Human-Computer Interaction, Vol. 11, No. 1, pp. 1–28.

Steuer, J. (1992, Autumn), "Defining Virtual Reality: Dimensions Determining Telepresence," *Journal of Communication*, Vol. 42, No. 4, pp. 73–93.

Stone, B. (2005, April 11), "Hi-Tech's New Day," *Newsweek*, p. 62

Streitz, N. A. (2007), "From Human–Computer Interaction to Human–Environment Interaction: Ambient Intelligence and the Disappearing Computer," in *Universal Access in Ambient Intelligence Environments, Proceedings of the 9th ERCIM Workshop on User Interfaces for All*, C. Stephanidis and M. Pieper, Eds., Königswinter, Germany, September 27–28, 2006, pp 3–13.

Sturman, D. J., and Zeltzer, D. (1994), "A Survey of Glove-Based Input," *IEEE Computer Graphics and Applications*, Vol. 14, No. 1, pp. 30–39.

Sutherland, I. E. (1963), "Sketchpad: A Man-Machine Graphical Communication System," Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Sutherland, I. E. (1965), "The Ultimate Display," *Proceedings of IFIP 65*, Vol. 2, pp. 506–508.

Taylor, F. W. (1911), *Principles of Scientific Management*, Harper, New York.

Tractinsky, N. (1997), "Aesthetics and Apparent Usability: Empirically Assessing Cultural and Methodological Issues, CHI '97 Electronic Publications: Papers," available: http://www.acm.org/sigchi/chi97/proceedings/paper/nt.htm.

Traum, D. (1999), "Computational Models of Grounding in Collaborative Systems," in *Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, S. E. Brennen, A. Giboin, and D. Traum, Eds., American Association for Artificial Intelligence, Menlo Park, CA, pp. 124–131.

Tripathi, A. K. (2005), "Computers and the Embodied Nature of Communication: Merleau-Ponty's New Ontology of Embodiment," *Ubiquity*, Vol. 6, No. 44, available: http://www.acm.org/ubiquity.

Turing, A. (1950), "Computing Machinery and Intelligence," *Mind*, Vol. 59, pp. 433–460.

Turk, M., and Robertson, G. (2000), "Perceptual User Interfaces (Introduction)," *Communications ACM*, Vol. 43, No. 3, pp. 32–34.

Versweyveld, L. (2001), "Neuro-Surgeons Use 3D Virtual Reality Dextroscope Technology to Assist in Pre-Operative Planning," *Virtual Medical Worlds*, available: http://www.hoise.com/vmw/01/articles/vmw/LV-VM-06-01-9.html.

Walker, M. A., Stent, A., Mairesse, F., and Prasad, R. (2007), "Individual and Domain Adaptation in Sentence Planning for Dialogue," *Journal of Artificial Intelligence Research*, Vol. 30, pp. 413–456.

Wasow, T., Perfors, A., and Beaver, D. (2005), "The Puzzle of Ambiguity," in *Morphology and The Web of Grammar: Essays in Memory of Steven G. Lapointe*, O. Orgun and P. Sells, Eds., CSLI Publications, Stanford, CA.

Weiser, M., and Brown, J. S. (1997), "The Coming Age of Calm Technology," in *Beyond Calculation: The Next Fifty Years*, P. J. Denning and R. M. Metcalfe, Eds., Copernicus, New York, pp. 75–85

Weizenbaum, J. (1966), "ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine," *Communications of the ACM*, Vol. 9, No. 1, pp. 36–45.

Wellman, S. (2007), "Google Lays Out Its Mobile User Experience Strategy," *InformationWeek*, available: http://www.informationweek.com/blog/main/archives/2007/04/google_lays_out.html.

Wickens, C. D., and Carswell, C. M. (2006), "Information Processing," in *Handbook of Human Factors and Ergonomics*, G. Salvendy, Ed., Wiley, Hoboken, NJ, pp. 111–149.

Woods, W. A., Kaplan, R. M., and Nash-Webber, B. L. (1972), "The Lunar Sciences Natural Language Information System," Final Report, BBN Report 2378, Bolt Beranek and Newman, Cambridge, MA.

Zhou, J., Yu, C., Riekki, J., and Kärkkäinen, E. (2007), "AmE Framework: A Model for Emotion-Aware Ambient Intelligence," paper presented at the Second International Conference on Affective Computing and Intelligent Interaction (ACII2007): Doctoral Consortium, Lisbon, Portugal, September 12–14, 2007, pp. 163–69.

Zue, V.W., and Glass, J.R. (2000), Conversational Interfaces: Advances and Challenges, Proceedings of the IEEE, Vol. 88, No. 8

# PART 9

## DESIGN FOR INDIVIDUAL DIFFERENCES

# CHAPTER 51

# DESIGN FOR PEOPLE WITH FUNCTIONAL LIMITATIONS

**Gregg C. Vanderheiden and J. Bern Jordan**
**University of Wisconsin–Madison**
**Madison, Wisconsin**

## 1 INTRODUCTION

### 1.1 Not a Special Population But a Continuum — And the Future for Most of Us

Often, the topic of design for human disability and aging is thought of as a special topic, vertical market, or special application. Although there are special products or assistive technologies designed specifically for use by people with disabilities, they constitute only a small portion of the total number of products that need to be designed to accommodate persons with functional limitations. In addition to the specially designed tools, everyone, including those with disabilities, needs to have access to a wide range of technologies found in their everyday lives: at home, at school, on the job, and in the community. It is toward the more accessible design of everyday products that this chapter is directed.

Another common misconception is that the population in question is small. Although there are many different types and degrees of disabilities, some of which represent smaller numbers of people, cumulatively those with disabilities represent around 19% of the population. In addition, a majority of people who live beyond age 65 will have difficulties performing activities of daily living

(a definition which includes getting around inside the home, dressing, and eating) and instrumental activities of daily living (a definition which includes going outside the home, doing light housework, and using the telephone) due to disability. Approximately 37% of those over 65 have a severe disability; the prevalence jumps to 58% for those over 80 (Brault, 2008). In addition, many of these people experience multiple functional limitations.

Finally, it is important to note that the target audience for these design guidelines includes nearly everyone as we age and acquire disabilities (functional limitations). The only exception will be for those of us who die first.

## 1.2 Multiplier Effect

If companies were designing products for individuals, this population (people with functional limitations) would cumulatively constitute a significant portion of the market. When designing products to be used by families or within industry, the impact is multiplied. Since a family unit consists of three or four people, the percentage of families who have people with disabilities is much higher. When you turn to industry, particularly large industries, you find that the percentage of industries that employ people with disabilities is very high. Thus, if you are designing products and systems for use by larger industries, you will find that almost all of the customer base will have employees with disabilities.

## 1.3 Who Is Included in the Category "Disabled and Elderly Persons"

In considering product design, it is important to note that there is no clear line between people who are categorized as disabled and those who are not. A performance or ability distribution for a given skill or ability is generally a continuous function rather than bimodal with distinctive able and disabled groups. This distribution includes a small number of people who have exceptionally high ability, a larger number with midrange ability, and another longer tail representing those with little or no ability in a particular area. In looking at such a distribution, it is impossible simply to draw a vertical line and separate able-bodied from disabled persons. It is also important to note that each aspect of ability has a separate distribution. Thus, a person who is poor along an ability distribution in one dimension (e.g., vision) may be at the other end of the distribution (i.e., excellent) with regard to another dimension (e.g., hearing or IQ). Thus, people do not fall at the lower or upper end of the distribution overall, but generally fall into different positions, depending on the particular ability being measured.

## 1.4 The 95th Percentile Illusion

It should be clear that even if elderly and disabled persons are included in the mainstream design process, it is not possible to design all products and devices so that they are usable by all people. There will always be a "tail" of people who are unable to use a given product. To include a sizable portion of the population in the category "those who can use a product with little

or no difficulty," the 95th percentile data are often used. The problem is that there are no 95th percentile data for specific designs—there are only data with regard to individual physical or sensory characteristics. Thus, there are 95th percentile data for height, vision, hearing, and so on. As a result, it is not possible to determine when a product can be used by 95% of people. It is only possible to estimate when a product can be used by 95% of the population along any one dimension. Since people in the 5% tail for any one dimension (e.g., height) are usually not the same people as those in the 5% tail along another dimension (e.g., vision) (Kroemer, 1990), it is possible to design a product using 95th percentile data and end up with a product that can be used by far less than 95% of the population.

To illustrate this phenomenon, imagine a minipopulation of 10 people. Ten percent of them (1 of 10) have one short leg, 10% have a visual impairment, 10% have a missing arm, 10% are short, and 10% cannot hear. Let us assume that we design a product that required 90th percentile ability along each of the dimensions of height, vision, leg use, arm use, and hearing. In this instance we would end up with a product that was in fact usable by only 50% of this population. This occurs because, although only 10% of this minipopulation are limited in any single dimension, different people fall into the 10% tail for each dimension, and only 50% of the population are within the 90th percentile for all five areas.

In real life, the effect is not quite this dramatic, and its calculation is not as simple. First, the percentage of people with disabilities is less than 10% along any one dimension. Second, there is often overlap where one person would have more than one disability (e.g., elderly persons). On the other hand, there is a much wider range of different individual types of disability. In addition, the data from which the 95th percentiles are calculated often exclude persons with disabilities (Kroemer, 1990), making the percentage who could use the design(s) smaller than one would first calculate.

## 2 DISABILITY IS A CONSEQUENCE, NOT A CONDITION

> Disability is the inability to accommodate to the world as it is currently designed.
>
> *Vanderheiden, 1995, rephrasing Caplan, 1992*

The quote above is a slightly different take on Ralph Caplan's quote, "Disability is the inability to accommodate poor design," with an emphasis on the fact that design can be changed, and thus so can disability. In looking at the impact of disability and its relationship to design, it is often useful to use a model such as that shown in Figure 1, which is similar to the World Health Organization model for disability (WHO, 1980).

The model shows the relationship that both impairment and design have in creating disabilities. It also shows how circumstance can create similar reduced abilities in anyone, including those without functional
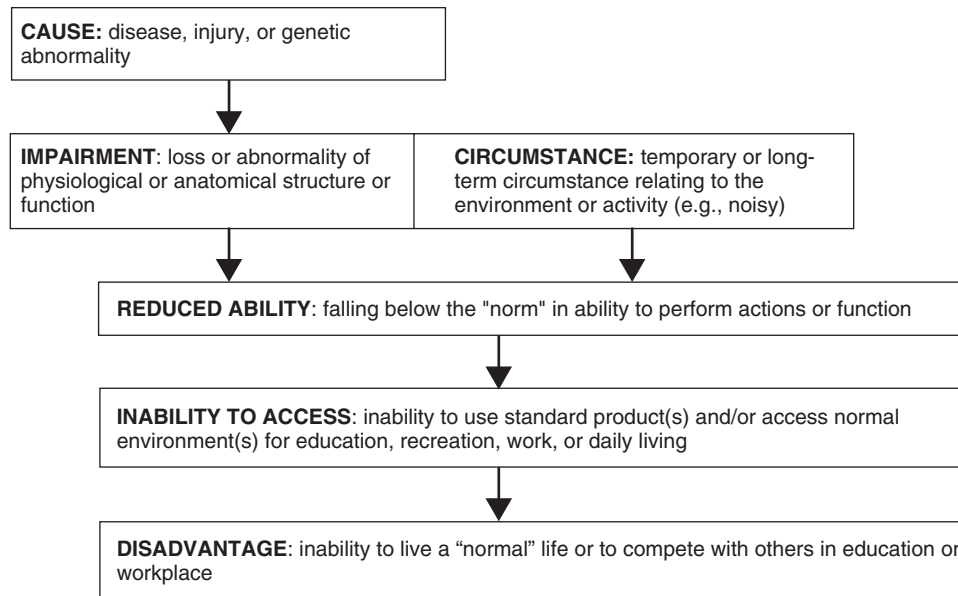
**Figure 1** Cause–effect model shows the role that both impairment and design play in disability as well as the parallel role that conditions or circumstances can play.

impairment. Combined with poor design, these circumstances can also lead to situations where people experience circumstantial disabilities or inabilities to carry out certain tasks. Thus, in addition to generally making products easier for everyone to use, better or more universal design can make a product usable even when people are under stressed conditions. Take, for example, a mother whose young son just fell and cut his head. She makes the mistake of mentioning the doctor and is now trying to use the phone while holding her screaming, kicking son in one arm to keep him from running off and hiding. Because of the screaming, she can hear very little and has some of the same functional problems as those of a person with a hearing impairment. Because her son is kicking and thrashing, she has poor motor control and has only one hand available. Because he is bleeding profusely, she is also highly distracted and is able to bring only limited cognitive skills and attention to the task at hand.

### 2.1 Three Approaches

There are basically three ways to address the problem faced by those who are unable to use the world around them:

*Change the person.* This may be accomplished through surgery, education, skill development, skill practice, or teaching strategies, tricks, or "secrets" for doing things or for doing things more easily. This would also include technologies that become "part of the person" such as eyeglasses, braces, and artificial limbs.

*Provide the person with bridging tools.* This includes devices and adapters that bridge between

the user and mainstream technologies [e.g., door knob adapters, screen readers, adaptive keyboards, telecommunication devices for the deaf (TDDs/TTYs)].

*Change the way that the world is designed.* Develop more universal and accessible designs for mainstream products.

Ergonomics is involved in all three of these areas by (1) developing new techniques, strategies, and technologies that can allow an unaided person to perform better in the workplace, home, or community; (2) developing specialized tools or assistive technologies that can adapt individual parts of the world to match the use of residual skills and abilities of a person; and, of course, (3) changing the design of the world in general so that it is more usable with a wider range of skills and abilities. The focus of this chapter is on the third approach: designing the world to be more universally usable.

### 3 UNIVERSAL DESIGN PROCESS

Universal design is the term that has been given to the practice of designing products or environments which can be used effectively and efficiently by people with a wide range of abilities operating in a wide range of situations. This includes people with no limitations as well as those operating with functional limitations relating to disabilities or simply by circumstance. For example, products developed using universal design principles would be flexible enough to be usable by people with no limitations as well as those:

- Who cannot see the product because they are blind or because their eyes are occupied temporarily (e.g., driving a car)
- Who cannot use their hands well because of aging or a physical disability or because their hands are temporarily full, cold, or gloved
- Who cannot speak or are in an environment where speech is not practical (library or noisy crowd)
- Who cannot hear the product because they are deaf or because they are in a very noisy environment (e.g., an airplane or a shopping mall at Christmas)
- Who have learning disabilities or who are able to divert only part of their attention to the task at hand
- Whose primary language is sign language or a foreign language
- Who are very young or very old

It is important to remember that universal design is a process, not a product. No matter how well something is designed to be accessible or "universally usable," there are going to be some people with severe disabilities or particular combinations of disabilities and circumstances that will not be able to use a product or service. An *ideal* design is one that is attractive, easy to learn, and effective and whose functions can be accessed efficiently and used by *everyone* across the full range of circumstances that occur for its intended use. Ideal designs however do not exist. A very good design is a commercially practical, mass market design that is usable by and attractive to the maximum possible number and diversity of users, given the best of today's collective knowledge, technologies, and materials. Universal design of mainstream products and services is the process of seeing how close to ideal designs you can get with a practical profitable commercial design.

## 3.1 Non-Disability-Related Reasons for Universal Design

### 3.1.1 Benefits for All

A general characteristic of good universal design is that it benefits many more people without disabilities than those with disabilities. This, of course, follows from the design benefiting everyone and the fact that there are more people without disabilities than with disabilities. The sidewalk curbcut is a prime example of this, as are ramps in general. Although originally designed for users of wheelchairs, they are also used by parents pushing baby carriages, people pulling baggage carriers, bicycle riders, skateboard users, kids on tricycles, and any number of other people. Even people walking can be observed to veer from their path to walk up a curbcut rather than stepping up a curb. In another example, a technique called "EZ Access" was used to allow persons who are blind to access and use touch screen–based kiosks. Once implemented, however, it was found that

it was also very useful for persons with low vision as well as those who could not read due to literacy or language problems.

### 3.1.2 New Insights

Studying the use of products by people with functional limitations can also provide insights into a design that might not otherwise be achieved. For example, it is much easier to determine which elements in a kitchen require greater strength by testing a person who is weak or who has poor grasp than it would be by employing someone with normal or extraordinary strength. Even if such a person were asked which things required more or less effort, the mere fact of having so much strength in reserve would cause the person to use it unconsciously.

### 3.1.3 Lower Cost Design

Universal design can also lead to insights that result in lower cost designs. Although universal designs are usually thought of as being more expensive, this is generally not the case. If one discounts the time it takes to reorient one's thinking and familiarize oneself with the characteristics and constraints of people with functional limitations, the resulting designs can be both easier to use and less expensive.

One example of this is the current design of elevators and their alert bells. In the past, people with disabilities had a problem getting onto elevators when they were arranged in elevator banks. Often, by the time the person using a wheelchair got to the elevator that had opened, the door had closed. New standards were proposed which would require that elevator doors stay open for a longer period of time to allow them to be boarded successfully by wheelchair users. This caused problems, because it increased the number of elevators that needed to be installed in buildings to ensure adequate service to all floors. In some thin, tall buildings, this could result in using up a substantial portion of the building for elevators.

After an injunction was sought to stop the standards, the designers and consumer advocates sat down to study the problem anew. It was determined that the problem was not the time it took to board the elevator but the time it took to get in front of the elevator. Since the elevators were computer controlled, and the computers knew where the elevators were going in advance of their arrival, it was quickly determined that lighting the alert light and sounding the bell in advance would allow persons with disabilities to position themselves in front of the elevator door and be able to board as it opened. Testing bore this out, and it was found that people in wheelchairs as well as everyone else could actually begin the boarding process much more quickly and in much less time than the elevators were then staying open. Following the modification in timing of the alert light and bell, designers were able to decrease the time that the doors stayed open, allowing builders either to use fewer elevators or to provide better service to the floors.

According to data from the U.S. Current Population Survey, the employment–population ratio in 2009 was approximately 19% for people with a disability compared with 65% for people without disabilities.

In 2009, nearly 900,000 people with disabilities were looking for and unable to find work (Bureau of Labor Statistics, 2010). If their annual salaries only average $25,000, that amounts to $22 billion in lost productivity as well as lost tax revenues. This is in addition to the large costs in the form of transfer payments and program costs for those who cannot live independently. Total U.S. federal, state, and local public expenditures for disability-related programs was $294 billion for 1997 and estimated at $426 billion for 2002 (Braddock, 2002). What portion of this could be saved if the design of the environment allowed people to live more independently or to stay on their jobs longer?

## 4 DEMOGRAPHICS

As shown in Figures 2.4, the prevalence of the various types of functional limitation (visual, hearing, physical, cognitive) varies significantly as a function of age. In children we see a much higher percentage of mental retardation and language and learning disabilities than of other disabilities (Figure 2). As people age, sensory and physical disabilities become more prevalent (Figure 3). Not evident from these charts is the fact that in older persons we see a much higher incidence of multiple disabilities, including combinations such as hearing and visual impairments, which interfere with many of the adaptive strategies developed for those who have hearing or visual impairments alone. Finally, we can see that the percentage of people who have functional limitations within the population increases sharply as a function of age. In fact, a wide majority of those over the age of 75 (Figure 4) will have functional limitations, and almost half will have severe functional limitations of one type or another. Thus, over our lifetimes (if we live long enough), most of us will not only benefit from but will require more universal design.

### 4.1 Characteristics of Users with Functional Limitations

In considering design for people with functional limitations, it is important to examine their abilities both without and with tools and strategies that they normally employ. For example, it is important to look at an amputee's abilities both with and without different types of artificial limbs. These present very different mechanical and manipulative characteristics. Many touch buttons, for example, cannot be activated using different
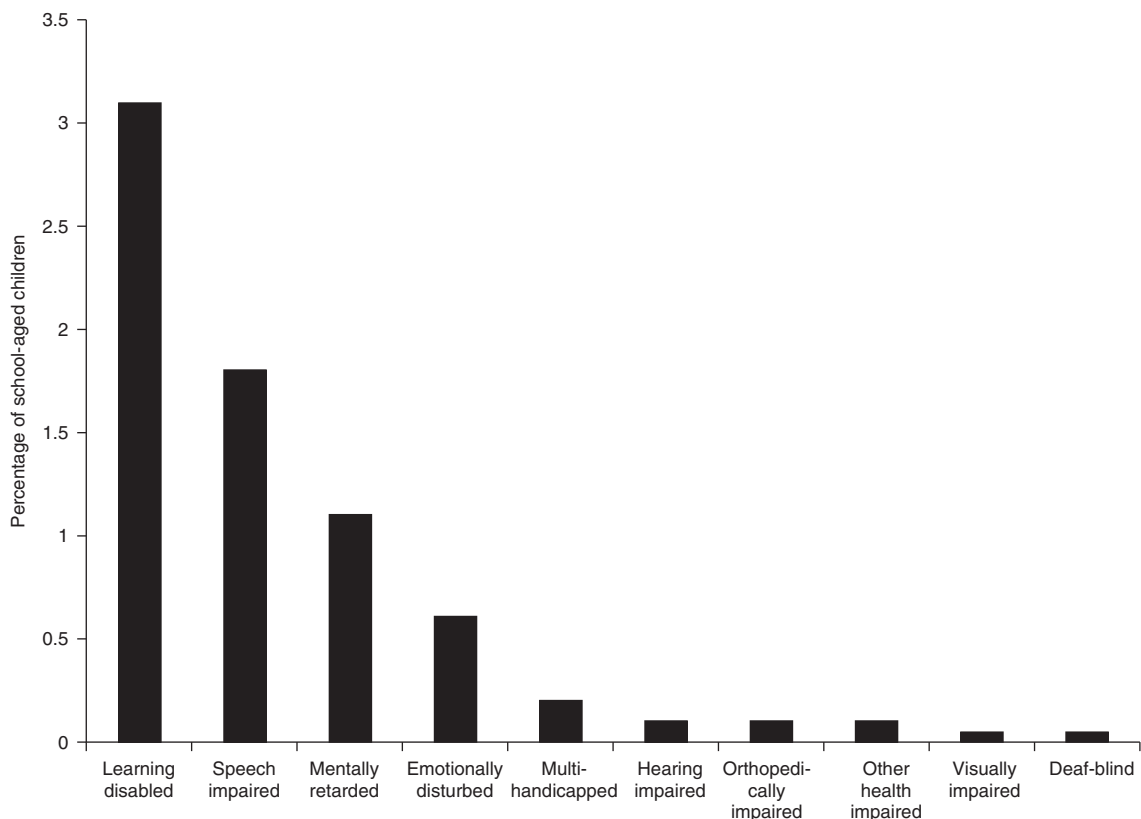


**Figure 2** Prevalence of impairment (primary diagnosis for school-aged children (3–21 years) in the United States. Each child is counted in only one category. (Data from Kraus and Stoddard, 1989, based on reports of the Office of Special Education and Rehabilitation Services, 1988. OSEP state reported date, 1986–1987 school year.)
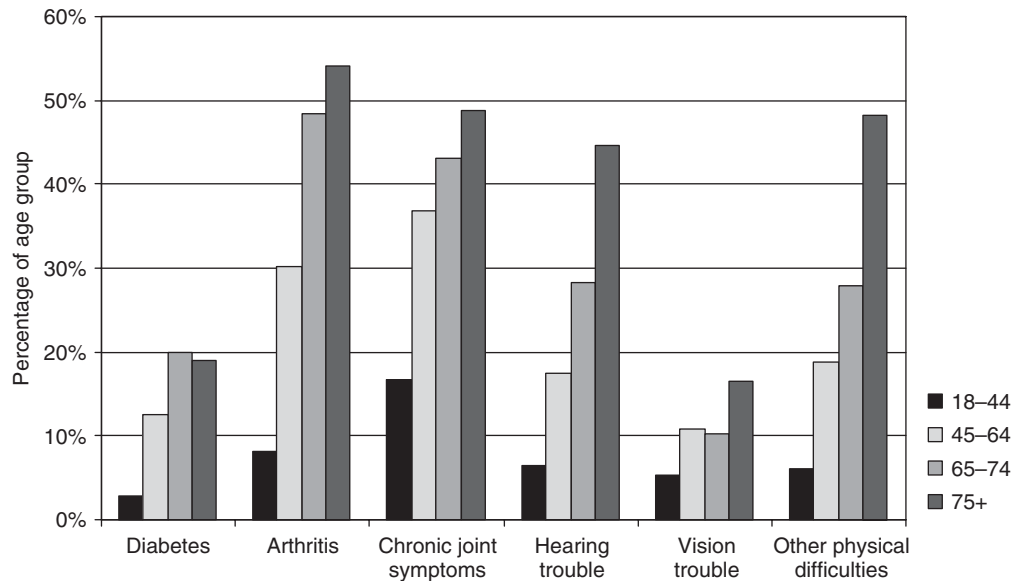
**Figure 3** U.S. prevalence of selected impairments within age groups. Data categories are not exclusive. (Data from Sondick et al., 2010, based on the 2009 National Health Interview Surveys, Centers for Disease Control and Prevention.)

types of artificial arms. Acoustic wave touch screens may be accessible using soft plastic cosmetic arms, but not hooks.

It is also important to consider people without their assistive devices, since many people do not have them, either because they cannot afford them or because they prefer to avoid the stigma (e.g., not wanting to use hearing aids or even very strong glasses). This, of course, adds to the variability and complicates any attempts at comprehensive surveying of needs, abilities, or characteristics. Although a comprehensive survey of the types of assistive technologies used by persons with different disabilities cannot be presented here, a partial listing is provided in Table 1. For a more comprehensive review, readers are referred to Galvin and Scherer (1996) or Cook and Polgar (2007).

## 5 RESEARCH IN ERGONOMICS AND PEOPLE WITH FUNCTIONAL LIMITATIONS

Most of the research on people with functional limitations is taken not from research on people with disabilities but rather from experiments done with "normal" persons operating under stress or adverse conditions (e.g., blinded by smoke, encumbered by a spacesuit). These studies represent much more controlled conditions than those represented by the great diversity of types, combinations, and degrees of disability but do yield interesting information that can be used by people with disabilities. As noted above, the results of work with persons with disabilities can also be applied to these other environments/locations where people have reduced abilities due to circumstance.

There are major problems in carrying out research in that the variation and range of ability or constraint is so great. Visual impairments, for example, can take a very wide range of forms, and each of these can vary in degree from very mild to severe reduction or total loss. As a result, it is not possible to make blanket statements about these populations. Instead, the research generally tries to characterize the diversity, to quantify numbers of people within particular ranges, and/or to chart the functional characteristics for major groups. For example, people who are experiencing hearing loss due to aging tend to lose hearing at certain frequencies more than others. (These results are reflected in the design guidelines that follow.) People with photosensitive epilepsy tend to be much more susceptible to certain frequencies than to others. The fact that there are no set patterns and that one can find people with just about any type, degree, and combination of disabilities makes developing design guidelines difficult. However, design principles do exist, as well as strategies that can significantly increase the accessibility and usability of products by a much wider range of people.

Note that this chapter refers to persons rather than populations. *Population* tends to imply somewhat homogeneous groups (although there may be variance within the group). When talking about people with functional limitations, we are talking about something that is a continuum that flows across many dimensions simultaneously. A classic example is people who are older who may have reductions in visual, hearing, physical, and/or cognitive abilities simultaneously. These abilities will also take different tracks and combinations in different people and will be progressive over time, making design of environments and products challenging. Clearly, designs must be flexible to accommodate
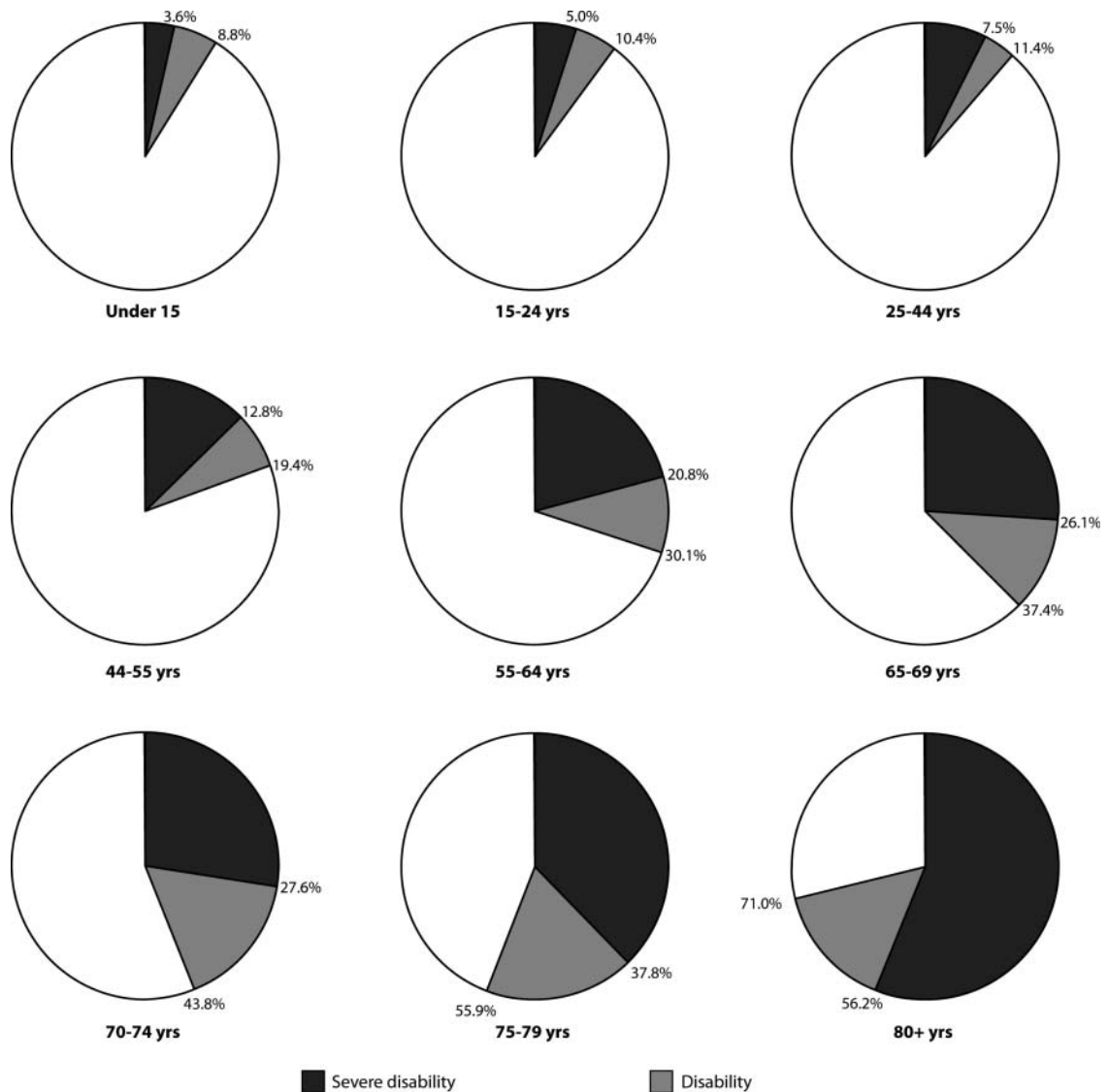
**Figure 4**  Disability as a function of age. Pie charts show the percentage of people who report having a disability within each age group. (Data from Brault, 2008, based on the 2004 Survey of Income and Program Participation, U.S. Census Bureau.)

different people, but in these cases they must be flexible to accommodate the same person over time or sometimes during different periods of the same day.

## 6   REGULATIONS AND GUIDELINES

There has been a steady advancement in both the capabilities of technology and the extensiveness of its use. Growing along with the advancement of technology is recognition of the importance of accessible technology, both in the United States and abroad.

In the United States, the U.S. Access Board (http://www.access-board.gov/) develops accessibility guidelines and regulations. Some of the notable accessibility legislation that applies to information and communication technology includes the American with Disabilities Act Accessibility Guidelines (civil rights legislation), Section 508 of the Rehabilitation Act of 1973 (which relates to federal procurement of accessible products and services and is being updated at the time of this publishing), and Section 255 of the Telecommunications Act of 1996 (which relates to the manufacture and sale

of telecommunication devices and is being updated and harmonized with the Section 508 regulations).

Many standards are developed at the international level and are often used in part or adopted by countries as legislation or guidance. The International Organization for Standardization (ISO) has a number of standards that relate to accessibility, the most noteworthy of which are ISO/International Electrochemical Commission (IEC) Guide 71:2001 ("Guidelines to address the needs of older persons and people with disabilities when developing standards"), ISO 9241-20:2008 ["Accessibility guidelines for information/communication technology (ICT) equipment and services"], and ISO 9241-171:2008 ("Guidance on software accessibility"). The Web Accessibility Initiative (WAI) of the World Wide Web Consortium (W3C) has published extensive guidelines on Web accessibility: Web Content Accessibility Guidelines (WCAG) 2.0, 2008. The Education and Outreach Working Group of the W3C WAI also maintains a list of policies relating to Web accessibility available at http://www.w3.org/WAI/Policy/.

Readers who are interested in more information about standards, regulations, and guidelines are referred to Hodgkinson (2009), ISO/IEC (2009b), or Vanderheiden (2009b).

## 7 OVERVIEW BY MAJOR DISABILITY GROUPS

Although there is a tremendous variety of specific causes, as well as combinations and severity of disabilities, we can most easily relate their basic impact to the use of consumer products by looking at five major categories of impairment: (1) visual impairments, (2) hearing impairments, (3) physical impairments, (4) cognitive/language impairments, and (5) seizure disorders. In addition, we discuss some of the common situations of multiple impairments.

### 7.1 Visual Impairments

Visual impairment represents a continuum from people with very poor vision, to people who can see light but no shapes, to people who have no perception of light at all. However, for general discussion it is useful to think of this population as representing two broad groups: those with low vision and those who are legally blind. In 2000, there were an estimated 3.3 million people with visual impairments (including blindness) (Congdon et al., 2004). In the elderly population the percentage of persons with visual impairments is very high.

As established by the American Medical Association in 1934, a person is termed *legally blind* when their visual acuity (sharpness of vision) is 20/200 or worse *after correction* or when their field of vision is less than $20°$ in the best eye after correction (Hoover and Bledsoe, 1981). There are approximately 937,000 people in the United States who are legally blind (Congdon et al., 2004). *Low vision* includes problems (after correction) such as dimness of vision, haziness, film over the eye, foggy vision, extreme near- or

farsightedness, distortion of vision, spots before the eyes, color distortions, visual field defects, tunnel vision, no peripheral vision, abnormal sensitivity to light or glare, and night blindness. There are approximately 2.4 million people in the United States with severe visual impairments who are not legally blind (Congdon et al., 2004). Many diseases causing severe visual impairments are common in those who are aging (glaucoma, cataracts, macular degeneration, and diabetic retinopathy). With current demographic trends toward a larger proportion of elderly, the incidence of visual impairments will certainly increase.

### 7.1.1 Functional Limitations Caused by Visual Impairments

Those who are legally blind may retain some perception of shape and contrast or of light versus dark (the ability to locate a light source) or they may be totally blind (having no awareness of environmental light). Those with visual impairments have the most difficulty with visual displays and other visual output (e.g., hazard warnings). In addition, there are problems in utilizing controls where labeling or actual operation is dependent on vision (e.g., where eye–hand coordination is required, as with a computer mouse or touch screen). Written operating instructions and other documentation may be unusable, and there can be difficulties in manipulation (e.g., insertion/placement, assembly).

Because many people with visual impairments still have some visual capability, many of them can read with the assistance of magnifiers, bright lighting, and glare reducers. Many such people with low vision are helped immensely by use of larger lettering, sans serif typefaces, and high-contrast coloring. Those with color blindness may have difficulty differentiating between certain color pairs. This generally does not pose much of a problem except in those instances when information is color coded or where color pairs are chosen that result in poor figure–ground contrast. Key coping strategies for people with more severe visual impairments include the use of braille and large raised lettering. Note, however, that braille is preferred by less than 10% of those who are legally blind (American Foundation for the Blind, 1996), normally those blind from early in life. Raised lettering must be large and is therefore better for indicating simple labels than for extensive text.

### 7.2 Hearing Impairments

Hearing impairment is one of the most prevalent chronic disabilities in the United States. From the 2009 National Health Interview Survey (NHIS), it is estimated that 34 million adults in the United States (15%) report at least some difficulty hearing (Sondick et al., 2010). From the most recent detailed NHIS about hearing impairments in 1990–1991, approximately 0.5% of the U.S population had severe to profound impairments, where people can at best understand words shouted into their better ear (Ries, 1994). *Hearing impairment* means any degree and type of auditory disorder; *deafness* means an extreme inability to discriminate conversational speech through the ear. Deaf people, then, are those who cannot use their hearing for communication. People with a lesser

**Table 1  Partial List of Assistive Technologies and Strategies**

| Type of Disability of Functional Limitation | Assistive Technologies and Strategies Used |
|---|---|
| Hearing impairment | <ul><li>Hearing aids</li><li>Cochlear implants</li><li>Amplifiers</li><li>Assistive listening devices (remote microphone that transmits to a receiver worn by the user)</li><li>Headphones</li><li>Inductive loops (that couple to the hearing aid)</li><li>Direct connection (wire from audio device to the hearing aid)</li><li>*(See also deafness strategies and technologies)*</li></ul> |
| Deafness | <ul><li>Telecommunication device for the deaf or text telephone (TDD/TT)</li><li>Text messaging (SMS, instant messaging, and real-time text)</li><li>Relay service (a special operator or interpreter with a TDD/TT)</li><li>Video relay service</li><li>Closed captions</li><li>Sign language</li><li>Sign language interpreters</li><li>Lip reading</li></ul> |
| Low vision | <ul><li>Lights</li><li>Magnifiers</li><li>Telescopes</li><li>Closed-circuit television</li><li>High-contrast display mode</li></ul> |
| Blindness | <ul><li>Braille (used by approximately 10% of those who are legally blind)</li><li>Dynamic braille displays (device with a series of braille cells with pins that move up and down to form the braille characters)</li><li>Tactile symbols and shapes</li><li>Raised line drawings</li><li>Long cane</li><li>Tape recorders</li><li>Synthetic speech</li><li>Portable note-takers with synthetic speech or braille</li><li>Talking clocks, watches, calculators</li><li>Satellite positioning systems and electronic map databases</li><li>Talking signs (infrared broadcasters in the environment that are picked up by small hand-held units)</li><li>Descriptive television (audio description track)</li><li>Voice output screen readers (on computer systems)</li></ul> |
| Physical impairment | <ul><li>Reachers</li><li>Artificial arms, legs, and hands or hooks</li><li>Canes and crutches</li><li>Walkers</li><li>Wheelchairs</li><li>Splints and braces</li><li>Mouthsticks and headsticks</li><li>Communication, writing, and control aids using a wide variety of input techniques, including sip and puff, Morse code, eyegaze, joystick, single-switch scanning, multiswitch encoding, etc.</li><li>Keyguards</li><li>Hand and arm rests</li><li>Universal remote consoles/controllers</li></ul> |

**Table 1** (*Continued*)

| Type of Disability of Functional Limitation | Assistive Technologies and Strategies Used |
| --- | --- |
| Speech impairment | • Voice amplifiers<br>• Voice synthesizers<br>• Artificial larynxes |
| Cognitive impairment | • Memory aids<br>• Cuing systems<br>• Calculators<br>• Text-to-speech aids<br>• Network-based assistance on demand |

degree of hearing impairment are called *hard of hearing*. Usually, a person is considered deaf when sound must reach at least 90 dB (5–10 times louder than normal speech) to be heard, and even amplified speech cannot be understood.

Hearing impairments can be found in all age groups, but loss of hearing acuity is part of the natural aging process. Of those aged 55–64, 15.4% have hearing impairments, and 29.1% over age 65 have hearing impairments (Ries, 1994). The number of people with hearing impairments will increase with the increasing age of the population and the increase in the severity of noise exposure.

Hearing impairment may be sensorineural or conductive. *Sensorineural hearing loss* involves damage to the auditory pathways within the central nervous system, beginning with the cochlea and auditory nerve and including the brain stem and cerebral cortex (this prevents or disrupts interpretation of the auditory signal). *Conductive hearing loss* is damage to the outer or middle ear, which interferes with sound waves reaching the cochlea. Causes of both types of hearing loss include heredity, infections, tumors, accidents, and aging (presbycusis, or "old hearing") (Schein, 1981).

### 7.2.1 Functional Limitations Caused by Hearing Impairments

The primary difficulty for people with hearing impairment in using standard products is receiving auditory information. This problem can be compensated for by presenting auditory information redundantly in visual and/or tactile form. If this is not feasible, an alternative solution to this problem would be to provide a mechanism, such as a jack, which would allow the user to connect alternative output devices. Increasing the volume range and lowering the frequency of products with high-pitched auditory output would be helpful to some less severely impaired persons. (Progressive hearing loss usually occurs in higher frequencies first.)

Basic voice input is now becoming more widespread as a feature of mobile and smartphones and may be extended to more commercial products in the future as the technology advances. This, too, will present a problem for many deaf persons. Whereas many have some residual speech, which they work to maintain, those who are deaf from birth or a very early age often are also nonspeaking or have speech that cannot be

recognized using current voice input technology. Thus, alternatives to voice input will be necessary for these people to access voice input products.

Familiar coping strategies for hearing-impaired people include the use of hearing aids, sign language, lip reading, TDDs (telecommunication devices for the deaf), and text communication (e.g., instant messaging, text messaging, and real-time text). Some hearing aids are equipped with a T-coil as well, which provides direct inductive coupling with a second coil (such as in a telephone receiver) to reduce ambient noise. Some other commercial products could make use of this capability. ASL (American Sign Language) is commonly used by people who are deaf. It should be noted, however, that this is a completely different language from English. Thus, deaf people who primarily use ASL may understand English only as a second language and may therefore not be as proficient with English as are native speakers. TDDs used to be the major mechanism for communication over the phone. These have largely been replaced by the use of SMS text messaging on mobile phones and instant messaging on computers. As we move to telephony via Voice over Internet Protocol (VoIP), real-time text can be built directly into any phone with a display—opening up text communication on almost any device used for voice telecommunication.

## 7.3 Physical Impairments

### 7.3.1 Functional Limitations Caused by Physical Impairments

Problems faced by people with physical impairments include poor muscle control, weakness and fatigue, difficulty walking, talking, seeing, speaking, sensing, or grasping (due to pain or weakness), difficulty reaching things, and difficulty doing complex or compound manipulations (push and turn). People with spinal cord injuries may be unable to use their limbs and may use mouth- or headsticks for most manipulations. Some people may not be able to perform simultaneous actions. Twisting motions may be difficult or impossible for people with many types of physical disabilities (including cerebral palsy, spinal cord injury, arthritis, multiple sclerosis, and muscular dystrophy).

Some people with severe physical disabilities may not be able to operate even well-designed products

directly. They usually must rely on assistive devices that take advantage of their specific abilities and standard products that are compatible with or can be used with their assistive devices. Commonly used assistive devices include mobility aids (e.g., crutches, wheelchairs), manipulation aids (e.g., prosthetics, orthotics, reachers), communication aids (e.g., single switch-based artificial voice), and computer–device interface aids (e.g., eyegaze-operated keyboard).

## 7.3.2 Nature and Causes of Physical Impairments

Neuromuscular impairments include:

- Paralysis (total lack of muscular control in part or most of the body)
- Weakness (paresis; lack of muscle strength, nerve enervation, or pain)
- Interference with control, via spasticity (where muscles are tense and contracted), ataxia (problems in accuracy of motor programming and coordination), and athetosis (extra, involuntary, uncontrolled, and purposeless motion)

*Skeletal impairments* include joint movement limitations (either mechanical or due to pain), small limbs, missing limbs, or abnormal trunk size. Some major causes of these impairments are described next.

**Arthritis**   Arthritis is defined as pain in joints, usually reducing range of motion and causing weakness. Rheumatoid arthritis is a chronic syndrome. Osteoarthritis is a degenerative joint disease. It was estimated that, in 1990, 38 million people in the United States had some form of arthritis or rheumatic condition. That number is expected to increase to 59 million by the year 2020 (Lawrence et al., 1998).

**Cerebral Palsy (CP)**   Cerebral palsy is defined as damage to the motor areas of the brain prior to brain maturity (most cases of CP occur before, during, or shortly following birth). Annually the incidence of CP is between 2.0 and 2.5 of every 1000 live births (Odding et al., 2006). CP is a type of injury, not a disease (although it can be caused by a disease), and does not get worse over time; it is also not "curable." Some causes of CP are problems with fetal development, lack of oxygen, and injuries related to preterm birth (Pellegrino, 2002). The most common types are (1) spastic, where a person moves stiffly and with difficulty; (2) ataxic, characterized by a disturbed sense of balance and depth perception; and (3) athetoid, characterized by involuntary, uncontrolled motion. Most cases are combinations of the three types.

**Spinal Cord Injury**   Spinal cord injury can result in paralysis or paresis (weakening). The extent of paralysis or paresis and the parts of the body affected are determined by how high or low on the spine the damage occurs and the type of damage to the cord. Quadriplegia involves all four limbs and is caused by injury to the cervical (upper) region of the spine; paraplegia involves only the lower extremities and occurs where injury was below the level of the first thoracic vertebra (mid-lower back). There are 229,000–306,000 people with spinal cord injuries in the United States, with 12,000–15,000 new cases projected annually (Foundation for Spinal Cord Injury Prevention, Care and Cure, 2009; Wyndaele and Wyndaele, 2006). Car accidents are the most frequent cause (42%), followed by falls and jumps (27%), gunshot wounds or other violent acts (15%), and sports and recreational injuries (7.6%) (Foundation for Spinal Cord Injury Prevention, Care and Cure, 2009).

**Traumatic Brian Injury (TBI)**   The term *head injury* is used to describe a wide array of injuries, including concussion, brain stem injury, closed head injury, cerebral hemorrhage, depressed skull fracture, foreign object (e.g., bullet), anoxia, and postoperative infections. Like spinal cord injuries, head injury and stroke often result in paralysis and paresis, but there can be a variety of other effects as well. Annually, about 1.7 million Americans sustain a traumatic brain injury. However, many of these are not disabled permanently or severely; nearly 80% are treated and released from an emergency department (Faul et al., 2010)

**Stroke (Cerebral Vascular Accident; CVA)**   The three main causes of stroke are thrombosis (blood clot in a blood vessel blocks blood flow past that point), hemorrhage (resulting in bleeding into the brain tissue; associated with high blood pressure or rupture of an aneurism), and embolism (a large clot breaks off and blocks an artery). Worldwide, it is the second most common cause of death and a major cause of disability (Donnan et al., 2008). The response of brain tissue to injury is similar whether the injury results from direct trauma (as above) or from stroke. In either case, function in the area of the brain affected either stops altogether or is impaired (Anderson, 1981).

**Loss of Limbs or Digits (Amputation or Congenital)**   This may be due to trauma (e.g., explosions, mangling in a machine, severance, burns) or surgery (e.g., due to cancer, peripheral arterial disease, diabetes). Usually, prosthetics are worn, although these do not result in full return of function. There are approximately 1.6 million persons living with the absence of a limb in the United States. Annually, approximately 185,000 Americans undergo an amputation of a limb (Ziegler-Graham et al., 2008).

**Parkinson's Disease**   This is a progressive disease of older adults characterized by muscle rigidity, slowness of movements, and a unique type of tremor. There is no actual paralysis. The usual age of onset is over 50 with increasing prevalence with age. Approximately 1% of the population over 60 and 4% of those over 80 have Parkinson's disease (de Lau and Breteler, 2006).

**Multiple Sclerosis (MS)**   Multiple sclerosis is defined as a progressive disease of the central nervous system characterized by the destruction of the insulating material covering nerve fibers. The problems these

patients experience include poor muscle control; weakness and fatigue; difficulty walking, talking, seeing, sensing, or grasping objects; and intolerance of heat. Onset is between the ages of 10 and 40. This is one of the most common neurological diseases, affecting 250,000–350,000 people in the United States alone (Anderson et al., 1992).

***ALS (Lou Gehrig's Disease)***    ALS (amyotrophic lateral sclerosis) is a fatal degenerative disease of the central nervous system characterized by slowly progressive paralysis of the voluntary muscles. The major symptom is progressive muscle weakness involving the limbs, trunk, breathing muscles, throat, and tongue, leading to partial paralysis and severe speech difficulties. This is an uncommon disease (with an estimated annual incidence of three to five cases per 100,000 people in the United States). It occurs mostly to those between ages 45 and 74 with men affected more often than women (Cronin et al., 2007).

***Muscular Dystrophy (MD)***    Muscular dystrophy is a group of hereditary diseases causing progressive muscular weakness; loss of muscular control; contractions; and difficulty in walking, breathing, reaching, and use of hands involving strength.

## 7.4   Cognitive/Language Impairments

### 7.4.1   Functional Limitations Caused by Cognitive/Language Impairments

The type of cognitive impairment can vary widely, from severe retardation to inability to remember, to the absence or impairment of specific cognitive functions (most particularly language). Therefore, the types of functional limitations that can result also vary widely.

Cognitive impairments are varied but may be categorized as memory, perception, problem solving, and conceptualizing disabilities. Memory problems include difficulty getting information from short-term, long-term, and remote memory. This includes difficulty recognizing and retrieving information. Perception problems include difficulty taking in, attending to, and discriminating sensory information. Difficulties in problem solving include recognizing the problem; identifying, choosing, and implementing solutions; and evaluation of outcome. Conceptual difficulties can include problems in sequencing, generalizing previously learned information, categorizing, cause and effect, abstract concepts, comprehension, and skill development. Language impairments can cause difficulty in comprehension and/or expression of written and/or spoken language.

There are very few assistive devices for people with cognitive impairments. Simple cuing aids or memory aids are sometimes used. As a rule, these people benefit from use of simple displays; low language loading; use of patterns; simple, obvious sequences; and cued sequences.

### 7.4.2   Types and Causes of Cognitive/Language Impairments

***Intellectual Disability***    A person is considered to have an intellectual disability if he or she has an IQ below 70 (average IQ is 100) and if they have difficulty functioning independently. The prevalence of intellectual disabilities is estimated at 0.87% in the United States (Larson et al., 2001), meaning that in 2010 there were approximately 2.6 million Americans with intellectual disabilities For most, the cause is unknown, although infections, Down syndrome, premature birth, birth trauma, or lack of oxygen may all cause intellectual disabilities. Those with mild intellectual disability have an IQ between 55 and 69 and achieve the fourth- to seventh-grade levels in education. They usually function well in the community and hold semiskilled and unskilled jobs. People with moderate intellectual disability have an IQ between 40 and 54 and are trainable in educational skills and independence. They can learn to recognize symbols and simple words, achieving approximately a second-grade level. They often live in group homes and work in sheltered workshops.

***Language and Learning Disabilities***    Aphasia, an impairment in the ability to interpret or formulate language symbols as a result of brain damage, is frequently caused by left cerebral vascular accident (stroke) or head injury. Specific learning disabilities are chronic conditions of presumed neurological origin that interfere selectively with the development, integration, and/or demonstration of verbal and/or nonverbal abilities. Aside from their specific learning disability, many people with learning disabilities are highly intelligent. Approximately 8% of children age 6–11 years have learning disabilities (Pastor and Reuben, 2002).

***Age-Related Disease***    Alzheimer's disease is a degenerative disease that leads to progressive intellectual decline, confusion, and disorientation. Dementia is a brain disease that results in the progressive loss of mental functions, often beginning with memory, learning, attention, and judgment deficits. The underlying cause is obstruction of blood flow to the brain. Some kinds of dementia are curable, whereas others are not.

## 7.5   Seizure Disorders

A number of injuries or conditions can result in seizure disorders. Epilepsy is a chronic neurological disorder. It is reported that approximately 150,000 children and adolescents will be medically evaluated because of seizures each year. Of these, approximately 30,000 will be diagnosed with epilepsy (Hauser, 1994). A seizure consists of an explosive discharge of nervous tissue, which often starts in one area of the brain and spreads through the circuits of the brain like an electrical storm. The seizure discharge activates the circuits in which it is involved, and the function of these circuits will determine the clinical pattern of the seizure. Except at those times when this electrical storm is sweeping through it, the brain is working perfectly well in a person with epilepsy. Seizures can vary from momentary loss of attention to grand mal seizures, which result in the severe loss of motor control and awareness. Seizures can be triggered in people with photosensitive epilepsy by rapidly flashing lights, particularly in the range 10–25 Hz (Harding and Jeavons, 1994).

### 7.6 Multiple Impairments

It is common to find that whatever caused a single type of impairment also caused others. This is particularly true where disease or trauma is severe or in the case of impairments caused by aging. Deaf blindness is one commonly identified combination. Most of these people are neither profoundly deaf nor legally blind but are both visual and hearing impaired to the extent that strategies for deafness or blindness alone will not work. People with developmental disabilities may have a combination of mental and physical impairments that result in substantial functional limitations in three or more areas of major life activity. Diabetes, which can cause blindness, also often causes loss of sensation in the fingers. This makes braille or raised lettering impossible to read. Cerebral palsy is often accompanied by visual impairments, by hearing and language disorders, or by cognitive impairments.

### 8 USER NEEDS–BASED APPROACH

It can be very difficult to keep track of the myriad of disabilities when practicing universal design. Using the framework of user needs allows a different view of the design problem. Instead of trying to keep track of all the different disabilities and things they need, the practitioner can consider three general needs that all users have (users need to be able to perceive, operate, and understand) with one additional need that applies only to some people with disabilities (compatibility with their assistive technology).

**Perceive**   People must be able to perceive all of the controls, operation feedback, and information that is displayed either passively or dynamically that is necessary to use the product. People must be able to find and refind all of the controls and perceive their status (e.g., the state of an on/off control or position and setting of a dial). Displayed information includes dynamic information on a screen, labels, instructions, printed output, and manuals and may be provided visually or through audio (usually speech).

**Operate**   People must be able to safely invoke and carry out all of the functions of a product within the time allowed and with equivalent privacy and security to other users. While the time allowed may be time limits that are imposed by a system, they are also time imposed by efficiency, competition, or productivity requirements. Everybody needs to be able to operate a device safely, and some disabilities may pose an additional risk to injury if people have difficulty seeing, moving, or changing their behavior to avoid potential seizures or physical injury.

**Understand**   People must be able to understand both how to use a product and all the output and displayed information. If there are helpful features or an access mode that a person needs to activate to use the product, they need to be able to discover the features and activate them.

**Assistive Technology (AT) Compatibility**   Unless direct, cross-disability access is built into a product, a person must be able to use their assistive technology with the product. With personal-use products, people may still prefer to be able to use their assistive technology rather than the built-in access.

An overview of all user needs organized in this fashion is available in Vanderheiden (2009a) or in ISO/IEC (2009a). For this chapter, we have chosen to organize the design guidelines by component to aid its use as a handbook for active designers.

### 9 DESIGN GUIDELINES

To facilitate use by product design teams, this section is organized by the functionality of products and user needs rather than by disability area. Functional categories are as follows:

- *Output/displays:* includes all means of presenting information to the user
- *Input/controls:* includes keyboards and all other means of communicating to the product
- *Manipulations:* includes all actions that must be performed directly by a person in concert with the product or for routine maintenance (e.g., inserting disk, loading tape, changing ink cartridge)
- *Documentation:* primarily operating instructions
- *Safety:* includes alarms and protection from harm

Each guideline is phrased as an objective, followed by a statement of the problem(s) faced by people with disabilities. The problem statement is accompanied by more specific examples. Next, design options are presented to provide some suggestions as to how the objective could be achieved. The guidelines are stated as generically as possible. Therefore, all, some, or none of the design options and ideas presented may apply in the case of any specific product. The recommended approach is to implement those options that together go the farthest toward achieving the objective of the guideline for your product. It is understood that this is not an ideal world, so it may currently be too expensive to implement all those ideas that would best achieve the objective. It is also anticipated that there will be other ways of meeting accessibility objectives than those discussed here, and such discoveries are encouraged.

### 9.1 Output/Displays

Maximize the number of people who can/will . . .

O-1 hear auditory output clearly enough (Perceive).

O-2 not miss important information if they cannot hear (Perceive).

O-3 have a line of sight to visual output and reach printed output (Perceive).

O-4 see visual output clearly enough (Perceive).

O-5 not miss important information if they cannot see (Perceive).

O-6 understand the output (visual, auditory, other) (Understand).

O-7 view the output display without triggering a seizure (Operate).

### 9.1.1 O-1 Maximize the Number of People Who Can Hear Auditory Output Clearly Enough (Perceive)

**Problem** Information presented auditorily (e.g., synthesized speech, cuing and warning beeps, buzzers, tones, machine noises) may not be heard effectively.

*Example*: People who have mildly to moderately impaired hearing may not be able to discern sounds that are too low in volume. People who have mild hearing impairments may be unable to turn the volume up sufficiently in some environments (e.g., libraries, where others would be disturbed, or in noisy environments, where even the highest volume is insufficient). People with moderate hearing impairments are often unable to hear sounds in higher frequencies (above 2000 Hz) (Hunt, 1970). People with hearing aids may have difficulty separating background noise from the desired auditory information. People with cognitive impairments may easily be distracted by too much background noise. Auditory information that is short or not repeated or repeatable (e.g., a short beep or voice message) may be missed or not understood.

*Note*: Severely hearing impaired (and deaf) people cannot use audio output at all. See Section O-2 for guidelines to address this problem.

### Design Options and Ideas to Consider

- Providing a volume adjustment, preferably using a visual volume indicator; sound should be intelligible (undistorted) throughout the volume range
- Providing automatically adjusted volume that is relative to the environmental noise level
- Making audio output (or volume range if adjustable) as loud as practical
- Using sounds that have strong middle- to low-frequency components (500–3000 Hz)
- Providing alerts and other auditory warnings that include at least two strong middle- to low-frequency components, with recommended ranges of 300–750 Hz for one component and 500–3000 Hz for the other (Berkowitz and Casali, 1990)
- Providing adjustable pitch for tones and sounds and selectable voices for speech
- Providing a headphone jack to enable a person with impaired hearing to listen at high volume without disturbing others, to enable such a person to isolate themselves effectively from background noise, and to facilitate use of neck loops and special amplifiers (Figures 5 and 6)
- Providing a separate volume control for the headphone jack so that people without hearing impairments can listen as well (at standard listening levels)
- When a headphone jack is not possible:
  - Placing the sound source on the front of the device and away from loud mechanisms would facilitate hearing.

- Locating the speaker on the front of the device would also facilitate use of a small microphone and amplifier to pick up and present the information (via speaker, neck loop, or vibrator).
- Facilitating direct use of the telecoil in hearing aids by incorporating a built-in inductive loop in a product (e.g., in telephone receiver's earpiece)
- Reducing the amount of unmeaningful sound or background noise produced by the product
- Having a warning beep precede the message to allow people to attend
- Using a male (lower) voice for speech synthesis
- Providing control of speech rate or speed
- Repeating the message or providing a mechanism for users to have the message repeated
- Providing methods for pausing, rewinding, and repeating speech
- Presenting auditory information continuously or periodically until the desired message is confirmed or acted upon
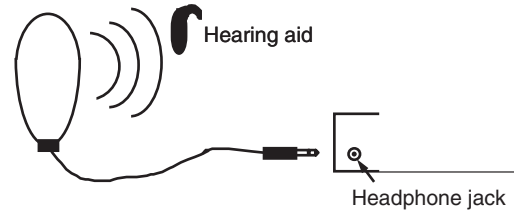


**Figure 5** A neck ring or ear loop can be plugged into a headphone jack on an audio source and provide direct inductive coupling between the audio source and a special induction coil on a person's hearing aid. This cuts out background noise that would be picked up by the hearing aid's microphone and provides clearer reception of the audio signal.
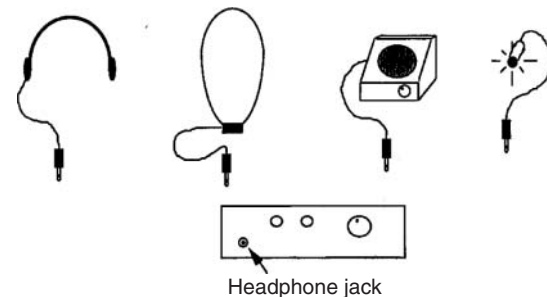


**Figure 6** A headphone jack permits the connection of headphones, neck/ear loops, amplifiers, or sound indication lights.

### 9.1.2 O-2 Maximize the Number of People Who Will Not Miss Important Information If They Cannot Hear (Perceive)

**Problem** Audio output (e.g., synthesized speech, cuing and warning beeps, buzzers, tones) may not be heard at all or may be insufficient for communicating information effectively.

*Example*: People who are severely hearing impaired or deaf may not hear audio output, even at high volume and low frequencies (Figure 7). People with language or cognitive impairments may not be able to respond to information given only in auditory form. (This may also be true if the language used is not a person's primary language.) People who are deaf-blind may not hear audio output. People with unimpaired hearing must sometimes use products in environments where the sound must be turned off (e.g., libraries) or where the environment is too noisy to hear any sound output reliably.

#### Design Options and Ideas to Consider

- Providing all important auditory information in visual form as well (or having it available; includes any speech output as well as auditory cues and warnings)
- Providing a tactile indication of auditory information (e.g., vibrating alarms)
- Facilitating the connection or use of tactile aids
- Providing an optional remote audio/visual or tactile indicator

### 9.1.3 O-3 Maximize the Number of People Who Will Have Line of Sight to Visual Output and Reach Printed Output (Perceive)

**Problem** Visual displays or printouts may be unreadable due to their placement.

*Example*: People who are in a wheelchair or who are extremely short may be unable to read displayed information due to the physical placement or angle of the display screen. People in wheelchairs, with missing or paralyzed arms, or with the ability to move limited by cerebral palsy or disease (e.g., severe arthritis, MS, ALS,

muscular dystrophy) may be unable to reach printed output (e.g., receipts produced by an automatic teller machine), due to printer placement.

#### Design Options and Ideas to Consider

- Locating display screens so they are readable from varying heights, including a wheelchair (see Section I-1 for specific anthropomorphic data and Section O-4 regarding image height)
- Locating printed output within easy reach of those who are in wheelchairs
- Facilitating manipulation of printouts by reaching and grasping aids
- Providing displays that can be adjusted in angle or height
- Providing redundant audio output in addition to visual display if the visual display cannot be made physically accessible to a person in a wheelchair (see Section O-5)

### 9.1.4 O-4 Maximize the Number of People Who Can See Visual Output Clearly (Perceive)

**Problem** Visual output (e.g., information presented on screens, paper printouts, cuing and warning lights or dials) may not be seen effectively.

*Example*: People who are visually impaired may not be able to see output that is too small. Those who are visually impaired may have difficulty discerning complex typefaces or graphics. People who are color blind may not be able to differentiate between certain color pairs. People with poor vision have more difficulty seeing letters or pictures against a background of similar hue or intensity (low contrast). People with visual impairments may be much more sensitive to glare (Figure 8). Those who have visual impairments may not be able to see detail in low lighting. Some people with severe lack of head control (e.g., cerebral palsy) may not be able to maintain continuous eye contact with a display and therefore may miss portions of dynamic (i.e., moving, changing) displays.

See Section O-5 for guidelines for people who cannot use visual output at all and Section O-6 for problems in understanding displayed output.
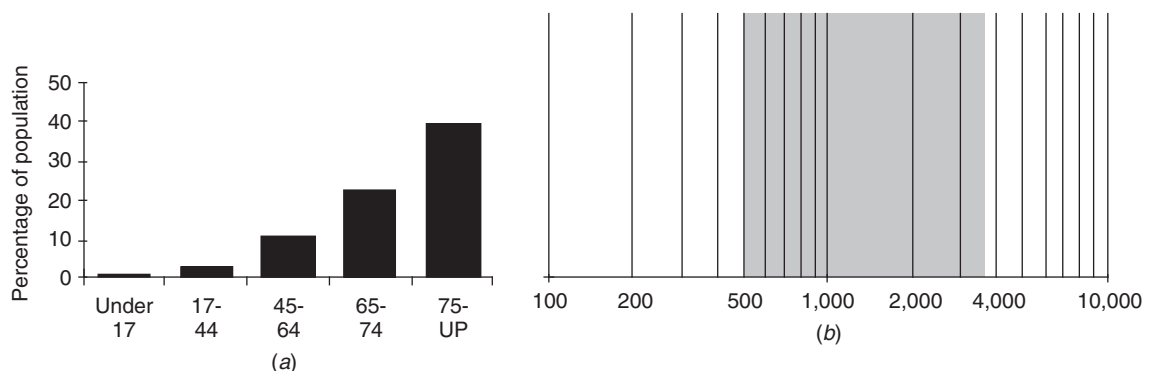


**Figure 7** (*a*) Hearing loss as a function of age; (*b*) recommended frequency for alerting devices. [(*a*) From Schow et al., 1978; (*b*) based on Hunt, 1970, and Berkowitz and Casali, 1990.]
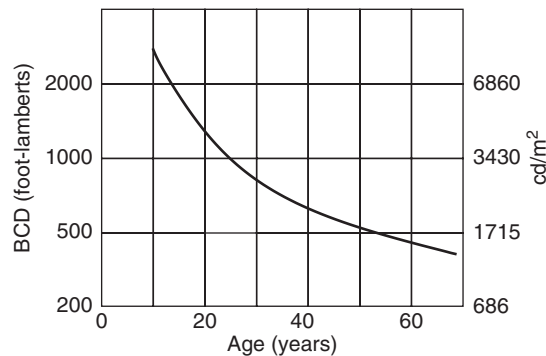
**Figure 8** Ability to tolerate glare decreases sharply as a function of age. Data are based on a 1° glare source size and a background luminance of 1.6 foot-lamberts. (From Bennet, 1977.)

### Design Options and Ideas to Consider

- Making letters and symbols on visual output as large as possible or practical
- Using upper- and lowercase type (title or sentence case) to maximize readability
- Making sure that letter spacing, the space between lines (leading), and the distance between messages are sufficient that the letters and messages stand out distinctly from each other
- Providing zoom or adjustable display image size
- Providing a video jack for attaching larger image displays or utilizing special assistive devices (e.g., electronic magnifiers)
- Using high contrast between text, graphics, user interface elements, and background
- Keeping letters and symbols on visual output as simple as possible; using sans serif typefaces for non-body-text lettering (e.g., labels, dials, displays) (see Section D-1)
- Using only black and white or using colors that vary in intensity or luminosity so that the color itself carries no information
- Providing adjustable color selection (hue and/or intensity)
- Replacing or supplementing color coding with different shape or relative position coding
- Providing contrast and/or brightness adjustment
- Minimizing glare (e.g., by employing filtering devices on display screens and/or avoiding shiny surfaces and finishes)
- Providing the best possible lighting for displays or areas containing instrumentation (good even illumination without hot spots and brighter than background illumination)
- Providing adjustable speed for dynamic displays (so they can be slowed down for those who lack motor control of their head and who must read with repeated glances)

- Avoiding use of light blue color to convey important information (harder for aging eyes to perceive as they yellow)
- Increasing contrast on liquid crystal displays (LCDs) by allowing user to adjust viewing angle
- Providing a pause control for text or other information that moves or scrolls

### 9.1.5  O-5 Maximize the Number of People Who Will Not Miss Important Information If They Cannot See (Perceive)

**Problem**  Visual output (e.g., information presented on screens, paper printouts, cuing and warning lights, and dials) may not be seen at all by some users.

*Example*: People who are severely visually impaired or blind may not be able to see visual output, even when magnified and clarified (as recommended in Section O-4). People who cannot read may be unable to use visually presented text. People who are deaf and blind may only be able to perceive tactile output. People who do not have any visual impairment may miss warnings, cues, or other information if it is presented only in visual form while their attention is diverted.

### Design Options and Ideas to Consider

- Providing all important visual information (redundantly) in audio and/or tactile form.
- Accompanying visual cues and warnings by a sound, one component of which is of middle to low frequency (500–3000 Hz) (see Section O-1)
- Making information that is displayed visually (both text and graphics) also available electronically at an external connection point (standard or special port) to facilitate the use of special assistive devices (e.g., voice synthesizers, braille printers), preferably in an industry or company standard format (see Figure 16)

### 9.1.6  O-6 Maximize the Number of People Who Can Understand the Output (Visual, Auditory, Other) (Understand)

**Problem**  Visual and/or auditory output may be confusing or difficult to understand.

*Example*: Some people with specific learning disabilities or with reduced or impaired cognitive abilities are easily confused by complex screen layouts (e.g., multiple "windows" of information), have difficulty understanding complex or sophisticated verbal (printed or spoken) output, or have a short attention span and are easily distracted when reviewing a screen display. For many people who are deaf as well as many other U.S. citizens, English is a second language and not well understood. People who are using a device in an alternate mode (e.g., speech output mode on a kiosk) may not understand instructions that are meant for people using the primary mode.

### Design Options and Ideas to Consider

- Using simple screen layouts
- Providing the user with the option to look at one thing at a time

- Shortening menus
- Hiding (or layering) seldom used commands or information
- Keeping language as simple as possible
- Accompanying words with pictures or icons (Note, however, that the use of graphics may present more difficulty for people who are blind. See Section O-5.)
- Using Arabic rather than Roman numerals (i.e., 1, 2, 3 instead of I, II, III)
- Using attention-attracting (e.g., underlining, boldfacing) and grouping techniques (e.g., putting a box around things or color blocking) to emphasize important information
- Highlighting key information
- Putting most important information at the beginning of written text (but not spoken announcements, where it might be missed)
- Providing an attention-getting sound or words before audio presentation
- Keeping auditory presentations short
- If providing menu choices, always state the choice first and the action second (e.g., use, "For deposits, press 1." Do not use, "Press 1 for deposits.")
- Providing auto-repeat or a means to repeat auditory messages
- Presenting information in as many (redundant) forms as possible/practical (i.e., visual, audio, and tactile) or providing as many display options as possible
- Providing digital readouts for product-generated numbers where the numeric or precise value is important
- Providing dials or bar graphs where qualitative information is more important (e.g., half full, full) (See Sections I-4 and I-6)

### 9.1.7  O-7 Maximize the Number of People Who Can View the Output Display without Triggering a Seizure (Operate)

**Problem**  People with seizure sensitivities (e.g., epilepsy) may be affected by some dynamic animations and screen cursor or display update frequencies, increasing the chance of a seizure while working on or near a display screen.

### *Design Options and Ideas to Consider*

- Avoiding screen refresh or update flicker or flashing frequencies which are most likely to trigger seizure activity (Figure 9 provides a general overview of the frequencies most likely to trigger a seizure.)
- Avoiding flashing where there are more than three flashes within any 1-s period where the combined area of the flashing would occupy more than 25% of the central vision (central $10°$) (Harding and Jeavons, 1994)
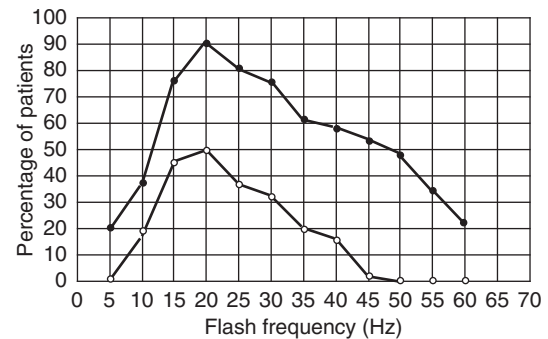


**Figure 9**  Percentage of photosensitive patients in whom a photoconvulsive response was elicited by a 2-s train of flashes with the eyes open and closed. As can be seen, the greatest sensitivity is at 20 Hz, with a steep drop-off at higher and lower frequencies. (From Jeavons and Harding, 1975.)

- Following the guidelines established for the Web Content Accessibility Guidelines 2.0 (http://www.w3.org/TR/WCAG20/#seizure), which allow exceptions to the general guideline of three flashes in 1 s for small or weak flashing, as well as other guidance on this

For the general flash threshold, a flash is defined as a pair of opposing changes in luminance (i.e., an increase in luminance followed by a decrease, or a decrease followed by an increase) of 20 candelas per rectangle meter ($cd/m^2$) or more and where the screen luminance of the darker image is below 160 $cd/m^2$. For the red flash threshold, a flash is defined as any pair of opposing transitions to or from a saturated red at any luminance level.

*Note*: Video waveform luminance is not a direct measure of display screen brightness. Not all display devices have the same gamma characteristic, but a display with a gamma value of 2.2 may be assumed for the purpose of determining electrical measurements made to check compliance with these guidelines. For the purpose of measurements made to check compliance with these guidelines, pictures are assumed to be displayed in accordance with the home viewing environment described in the International Telecommunication Union recommendation ITU-R BT.500, in which peak white corresponds to a screen illumination of 200 $cd/m^2$. Specifications are based on "ITC Guidance Note for Licensees on Flashing Images and Regular Patterns in Television" (revised and reissued in July 2001).

### 9.2  Input/Controls

Maximize the number of people who can . . .

I-1 reach the controls (Operate).

I-2 find the individual controls/keys if they cannot see them (Perceive).

I-3 read the labels on the controls/keys (Perceive).

I-4 determine the status or setting of the controls if they cannot see them (Perceive).

I-5 physically operate controls and other input mechanisms (Operate).

I-6 understand how to operate controls and other input mechanisms (Understand).

I-7 connect special alternative input devices (AT Compatibility).

### 9.2.1  I-1 Maximize the Number of People Who Can Reach the Controls (Operate)

**Problem**  Controls, keyboards, and so on, may be unreachable or unusable.

*Example*: People who use a wheelchair, who are very weak, or who are extremely short may be unable to reach some controls, keypads, and so on, well enough to use them. People with poor motor control may be able to reach the controls but may find the knobs, buttons, and so on, too small or close together to operate accurately. People with severe weakness may be able to reach the controls but may find the act of reaching or holding position in order to manipulate the controls too tiring.

#### Design Options and Ideas to Consider

- Locating controls, keyboards, and so on, so they are within easy reach of those who are in wheelchairs or have limited reach
- Locating controls so that the user can reach and use them with the least change in body position
- Locating controls that must be used constantly in the closest positions possible and where there is wrist or arm support
- Using keys with smaller radiuses of curvature (with sharper edges)
- Providing a (redundant) speech recognition input option
- Offering remote controls (wired, wireless, or bus operated)

### 9.2.2  I-2 Maximize the Number of People Who Can Find the Individual Controls/Keys If They Cannot See Them (Perceive)

**Problem**  People with visual impairments may be unable to find controls.

*Example*: People who are severely visually impaired may be unable to locate controls tactilely because they are on a flat membrane or glass panel (e.g., calculators, microwave ovens) or because they are placed too close together or in a complicated arrangement. People who have diabetes may have both visual impairments and failing sensation in fingertips, making it difficult to locate controls that have only subtle tactile cues.

#### Design Options and Ideas to Consider

- Varying the size of controls (also texture or shape), with the most important controls being larger to facilitate their location and identification
- Using keys, buttons, and controls that are raised or project from the background

- Providing controls whose shapes are associated with their functions
- Providing sufficient space between controls for easy tactile location and identification as well as easier labeling (large print or braille)
- Using keys with small radius of curvature on edges (sharper edges)
- Locating controls adjacent to what they control
- Using standard control arrangements (e.g., telephone number arrangement, direction pad)
- Making layout of controls logical and easy to understand, to facilitate tactile identification (e.g., stove burner controls in corresponding locations to actual burners)
- Providing a raised lip or ridge around flat (membrane or glass) panel buttons
- Providing a redundant speech recognition input option

See Figures 10 and 11.

### 9.2.3  I-3 Maximize the Number of People Who Can Read the Labels on the Controls/Keys (Perceive)

**Problem**  Labels on controls, keys, and so on, are difficult or impossible to see, due to their size, color, or location.

*Example*: People with low vision may have difficulty identifying controls or keys on a keyboard because the label lettering is too small and/or because the contrast between letters/graphics and background is poor. People with color blindness may have difficulty distinguishing controls that are color coded or use certain pairs of colors for labels and background. People with physical impairments may have difficulty reading labels on the sides or backs of objects. People who are blind may not be able to see printed labels at all.

#### Design Options and Ideas to Consider

- Making lettering used for labels as large as possible or practical
- Making sure that letter spacing, the space between lines (leading), and the distance between labels are sufficient so that the letters and labels stand out distinctly from each other
- Placing important labels or instructions on the front or an easily accessible side of large or stationary devices, where they can be read from wheelchairs
- Using sans serif fonts for non-body-text lettering (e.g., labels, dials)
- Using title case or all lowercase letters rather than all caps
- Using high contrast between letters/graphics and background
- Providing sufficient illumination of controls and instructions
- Providing backlit controls
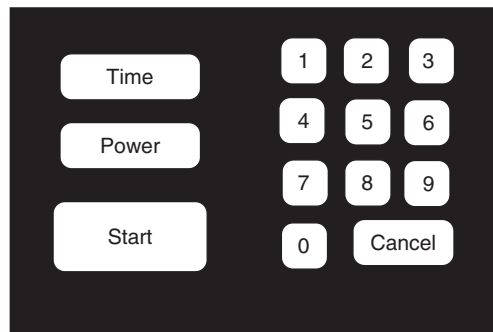
Keypad on which bottom-edge views below are based.

A flat membrane or glass keypad provides no tactile indication as to where the keys are, even if one memorize the arrangement.

Providing a slight raised lip around the keys allows their location to be discerned easily by touch. The ridge around the key also helps prevent slipping off of the key when using a mouthstick, reacher, etc., to press the keys.

Raised bumps are tactilely discernable, but it is harder to press the key without slipping off, particularly if you are using a mouthstick, reacher, or other manipulative aid unless there is a flat on top of the key,

Using indentations or hollows on the touchpad provides most of the advantage of ridges but is easier to clean. Hollows can be the same size as the key or of a consistent small circular size centered on the keys. Shallow edges, such as those on the left button, are harder to sense with fingers than the sharper curve of the middle button.

Raised keys with indents provide better feedback than just indents (as in example above), especially if the keys have different shapes or textures that correspond to their functions.

**Figure 10** The shape of a key or button can have a significant effect on people's ability to accurately locate and operate it.

- Supplementing color coding with use of different button/key shape or letter/graphic labels
- Providing tactile labels
- Avoiding use of blues, greens, and violets to encode information (since the yellowing of the cornea with age can cause confusion with some shades of these colors)
- Using easily interchangeable keycaps to allow replacement with special or optional keycaps
- Arranging controls in groupings that facilitate tactile identification (e.g., using small groups of keys that are separated from the other keys, or placing frequently used keys near tactile landmarks such as along the edges of a keyboard)
- Using established or standard layouts for keypads and keyboards (e.g., typewriter, adding machine, phone)
- Using voice output to speak the names of keys or buttons as they are pressed (This capability would need to be turned on and off as needed.)

- If a flat membrane panel cannot be avoided, providing a stick-on tactile overlay that provides tactile demarcation of the key locations and functions

See Sections O-4 and O-6 for related guidelines for output/displays.

### 9.2.4 I-4 Maximize the Number of People Who Can Determine the Status or Setting of the Controls If They Cannot See Them (Perceive)

**Problem** Determination of control status or setting may depend solely on vision.

*Example*: People with visual impairments may be unable to see a control setting or on/off indicator (e.g., where a dial is set, whether a button is pushed in, whether a light is on, flashing or off, or what a numeric setting on a visual display reads).

### Design Options and Ideas to Consider

- Providing multisensory indication of the separate divisions, positions, and levels of the controls
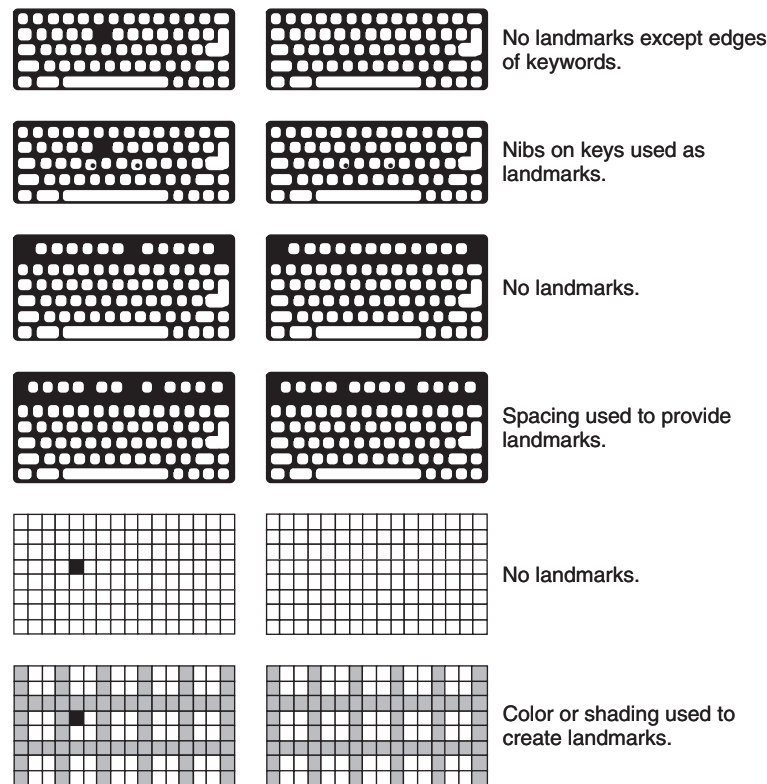
**Figure 11**   Quick self-demonstration of the impact of landmarks on key finding by people who cannot see labels on a key due to blindness or very low vision. *Instructions*: For each keyboard, visually locate the key on the right-hand keyboard that corresponds to the key marked on the left. Note the increase in speed and accuracy when landmarks (nibs or breaks in the key patterns) are provided.

(e.g., use of detents or clicks to indicate center position or increments, raised lines)
- Using absolute reference controls (e.g., pointers) rather than relative controls (e.g., pushbuttons to increase/decrease, or round, unmarked knobs)
- Using moving pointers with stationary scales
- Providing multisensory indications of control status (e.g., in addition to a status light indicating "on," provide an intermittent audible tone and/or tactilely discernable vibration)
- Using direct keypad input
- Providing speech output to read or confirm the setting

See Sections O-3, O-4, and O-5 for design options covering visual displays. See Figures 12 and 13.

### 9.2.5   I-5 Maximize the Number of People Who Can Physically Operate Controls and Other Input Mechanisms (Operate)

**Problem**   Controls (or other input mechanisms) may be difficult or impossible for those with physical disabilities to operate effectively.

*Example*: People with severe weakness may be unable to operate controls at all or may have great difficulty performing constant, uninterrupted input. People with only one arm or without arms (but utilizing assistive devices such as headsticks or mouthsticks) may not be able to activate multiple controls or keys at the same time. People with artificial hands or reaching aids may have difficulty grasping small knobs or operating knobs or switches which require much force. People with poor coordination or impaired muscular control have slower or irregular reaction times, making time-dependent input unreliable. People lacking fine movement control may be unable to operate controls requiring accuracy (e.g., a mouse or joystick) or twisting or complex motions. People with limited movement control (including tremor, poor coordination, or those using headsticks or mouthsticks) can inadvertently bump extra controls on their way to a nearby desired control.

### Design Options and Ideas to Consider

- Minimizing the need for strength by minimizing force required as much as possible or by providing adjustable force on mechanical controls
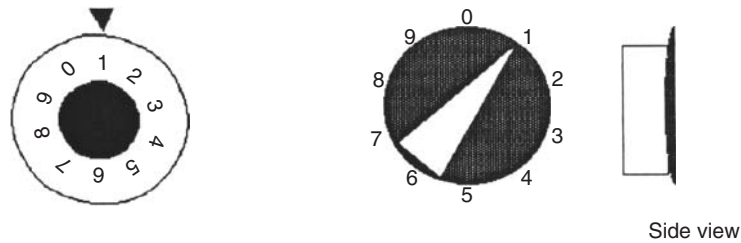
- No nonvisual indication of setting. If vision blurred, one cannot tell setting.

- Difficult to put large print or braille labels on knob.

- Harder to grasp and requires twisting motion.

- Highly visible raised pointer.

- Instant tactile indication of orientation allows setting to be read even if user is blind.

- Easy to put larger-print or braille labels on back panel.

- Use of detents (large and small) can facilitate internumeral settings.

- Black base disk provides high contrast and helps in control location/orientation on panel.

- Design is also easy to grasp and can be turned by pushing the point around — no twisting if the knob turns freely enough.

FOR EXAMPLE: What are the settings of the knobs below?



**Figure 12**  The design of a knob can greatly affect its usability by people with low vision or blindness.



POOR:
Round smooth knob,
no tactile orientation cue.

BETTER:
Has tactile orientation cue but
user has to feel around to find it.

BETTER:
Orientation cue is less
ambiguous. However, the user
must still feel the ends to be
sure which is the pointer end.

BEST:
Has tactile orientation cue
which is unambiguous and
can be felt immediately upon
grasping knob.
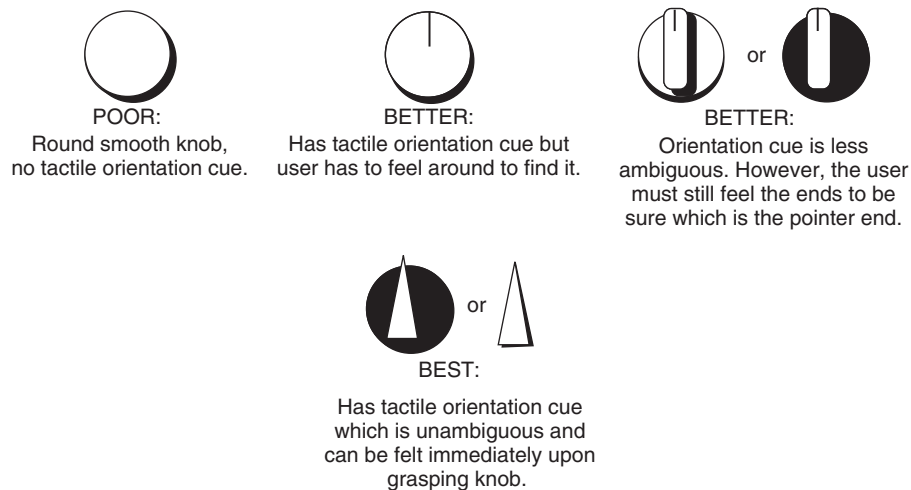
**Figure 13**  Knob and pointer design can have substantial effect on usability by people who are blind.

- If stiff resistance is provided to prevent accidental activation, it could drop off after activation. Other non-strength-related safety interlocks could also be considered.
- Spacing the controls to provide a guard space between controls, thus also leaving room for adaptations such as attaching levers to hard-to-turn knobs or room to replace knobs with larger, easier-to-turn knobs or cranks
- Minimizing or providing alternatives to performing constant, uninterrupted actions (e.g., button locks or push on–push off buttons to eliminate the need to press and hold some buttons continuously)
- Where simultaneous actions are required (e.g., pressing shift or control key while typing another key), providing an alternative method to achieve a result that does not require simultaneous actions (e.g., sequential option as in StickyKeys)
- Providing for operation with the left or right hand
- Using concave and/or nonslip buttons (see Figure 10 for a discussion), which are easier to use with mouthsticks or headsticks; on flat membrane keypads, providing a ridge around buttons
- If product requires a quick response (i.e., a reaction time of less than 5 s or release of a key or button in less than 1.5 s), allowing the user to adjust the time interval or to have a non-time-dependent alternative input method
- If product requires fine motor control, providing an alternative mechanism for achieving the same objectives that does not require fine motor control (e.g., on a mouse-based computer, provide a way to achieve mouse actions from the keyboard)
- Avoiding controls that require twisting or complex motions (e.g., push and turn) (*Note*: There are rotating knobs that can be turned by brushing the control with a hand and do not require twisting of the wrist.)
- Spacing, positioning, and sizing controls to allow manipulation by people with poor motor control or arthritis
- Where many keys must be located in close proximity, providing an option that delays the acceptance of input for a preset, adjustable amount of time (i.e., the key must be held down for the preset amount of time before it is accepted), thus helping some users who would otherwise bump and activate keys on the way to pressing their desired key (*Note*: This option must be difficult to invoke accidentally and be provided on request only, as it can have the effect of making the keyboard appear to be "broken" to naive users.)
- Making keyboards adjustable from horizontal ($0°–15°$ is standard) (Grandjean, 1987; Mueller, 1990)
- Providing an optional keyguard or keyguard mounting for keyboards
- Providing optional redundant voice control
- Providing textured controls and avoiding slippery surfaces/controls
- Providing means to stabilize body part used to operate the device (e.g., a wrist rest)

See Figure 14.
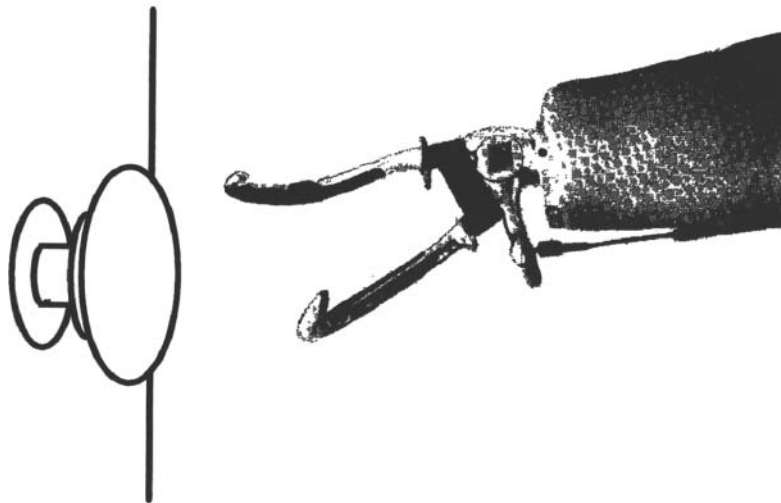


**Figure 14** People with arthritis, artificial hands, hooks, disabilities that restrict wrist rotation, or disabilities that cause weakness have difficulty with knobs or controls that require twisting. Such controls are also difficult to use for people with loss of upper body strength, range of motion, and flexibility, as is common with elderly persons. These really should be avoided in bathrooms where soap and water create a slippery environment. (Lever handles, now required in many building codes, facilitate access.)

### 9.2.6 I-6 Maximize the Number of People Who Can Understand How to Operate Controls and Other Input Mechanisms (Understand)

**Problem** The layout, labeling, or method of operating controls and other input mechanisms can be confusing or unclear.

*Example*: People with reduced or impaired cognitive function may be confused by complex, cluttered control layouts, with many and/or many types of controls; may have difficulty making selections from large sets; may have trouble remembering sequences (see also Section M-5); may be confused by dual-purpose controls; or may not relate appropriately to control settings indicated solely by notches/dots or numbers. People with reduced or impaired cognitive function, language impairments, and illiteracy or for whom English is a second language may have difficulty relying solely on textual labels, especially where abbreviations are used. They sometimes have difficulty making associations between label and control or may have trouble with timed responses involving text.

### Design Options and Ideas to Consider

- *Reducing the number of controls*
  - Limiting the number of choices where practical
  - Using layering of controls where only the most frequent or necessary controls or commands are visible unless you open a door or ask for additional levels of commands (e.g., hiding less frequently used controls, or at least grouping the most frequently used controls together and placing them prominently)
  - Where possible, making products automatic or self-adjusting, thus removing the need for controls (e.g., TV fine tuning and horizontal hold)
- *Simplifying the controls*
  - Minimizing dual-purpose controls
  - Using direct selection techniques where practical (selection techniques where the person need only make a single, simple, non-time-dependent movement to select)
  - Using visual/graphic indications for settings along with, or instead of, numbers or notches/dots (i.e., substitute concrete indications for abstract indications)
  - Reducing or eliminating lag/response times
  - Minimizing control or mode ambiguity
  - Providing a busy indicator or, preferably, a progress indicator when a product is busy and cannot take further input or when there is a delay before the requested action is taken
  - Integrating, grouping, and otherwise arranging controls to indicate function or sequence of operation

- *Making labels easy to understand*
  - Placing the label on or, less preferably, immediately adjacent to the control (does not apply to scales, which should not be on the controls but on the background)
  - Placing a line around the button and label (or from button to label) to show association (should be kept away from any lettering, especially if it is raised to avoid tactile confusion with the lettering)
  - Using simple concise language
  - Using redundant labeling (e.g., color code plus label)
  - Avoiding abbreviations in labeling (e.g., PrtScr, FF, C)
  - Leaving space around keys (makes it easier to match labels to keys and easier to add special labels)
  - Using multisensory presentation of feedback information
  - Providing labels on interinterval marks
- *Reducing, eliminating, or providing cues for sequences*
  - Allowing use of programmable function keys or using a "default" mode
  - Using preprogrammed buttons for common sequences
  - Allowing entry of a short code or scanning of a barcode to program a longer sequence
  - Simplifying required sequences, limiting the number of steps
  - Arranging controls to indicate sequence of operation
  - Adding memory cues or simple operating instructions on the device where possible
  - Cueing required sequences of action
  - Providing an easy exit that returns the user to the original starting point from any point in the program/sequence (exit should be prominent and clear)
  - Using wizards or step-by-step dialogues for a task sequence
- *Building on users' experiences (make the similarity obvious)*
  - Laying out controls to follow function
  - Making operation of controls follow movement stereotypes
  - Using common layouts or patterns for controls
  - Using common color-coding conventions in addition to textual or graphic labeling
  - Standardizing by using the same shape/color/icon/label for the same function or action (within and across products and manufacturers)

### 9.2.7 I-7 Maximize the Number of People Who Can Connect Special Alternative Input Devices (AT Compatibility)

**Problem** Standard controls (or other input mechanisms) cannot be made accessible for all of those with severe impairments.

*Example*: People with paralysis of their arms, severe weakness, tremor, or other severe physical impairments may not be able to use controls or input mechanisms that require the use of hands. People who are blind cannot use input devices that require constant eye–hand coordination and visual feedback (e.g., a standard computer mouse, trackball, or touch screen without special accommodation).

### *Design Options and Ideas to Consider*

- Providing a standard infrared remote control (e.g., audio/visual equipment)
- Providing alternative means for eye–hand coordination input devices (e.g., mice, trackballs,

relative joysticks) or allowing special devices to be substituted by the user, which will achieve as many of the functions as possible

- Using standard human interface device universal serial bus (USB) drivers and input so users can connect their own USB devices
- Providing tactile or auditory cues to allow direct use of touch pads or techniques to allow touch screens to function alternatively as auditory or tactile touch pads
- Providing a standard connection point (connector, infrared link, or wireless) for special alternative input devices (e.g., eyegaze keyboards, communication aids)
- Providing a network connection and accessible Web-based interface for control

See Figures 15 and 16.
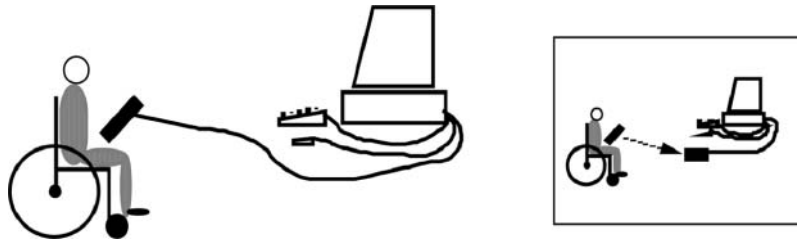


**Figure 15** By using standard USB human interface driver protocols, it is possible for users who cannot use the standard keyboard and mouse to create ''authentic'' keystrokes and mouse movements with their own input devices. This would allow these people to access the computer and all of its software.



People who are blind or unable to read the displayed information could use an assistive device and have information presented in auditory or tactile (braille) form and to provide input to the terminal.

People who are unable to operate the standard controls could use an assistive device to control the terminal using an input system they can control (eyegaze, sip and puff, single-switch scanning, etc.).

- **Public information terminal**
- **Restaurant and hotel guide at airport**
- **Automated teller machine**
- **Electronic building directory**
- **Point-of-sale terminal**
- **Information or sales kiosk at airport or mall or other public information/transaction terminal**

**Figure 16** A wireless bidirectional link could provide a low-cost environment and vandal-resistant mechanism for connecting assistive devices to information, control, and transaction terminals.

## 9.3   Manipulations

This includes all actions that must be directly performed by a person in concert with the device or for routine maintenance (e.g., inserting disk, loading tape, changing ink cartridge)

Maximize the number of people who can ...

M-1 physically insert and/or remove objects as required to operate a device (Operate).

M-2 physically handle and/or open the product (Operate).

M-3 remove, replace, or reposition often-used detachable parts (Operate).

M-4 understand how to carry out the manipulation necessary to use the product (Understand).

### 9.3.1   M-1 Maximize the Number of People Who Can Physically Insert and/or Remove Objects as Required to Operate a Device (Operate)

**Problem**   Insertion and/or removal of objects required to operate some devices (e.g., removable media, USB flash drives, Blu-ray discs, DVDs, credit cards, keys, coins, currency) may be physically impossible. In addition, damage to the object or device can occur from unsuccessful attempts.

*Example*: People using mouthsticks or other assistive devices may have difficulty grasping an object and manipulating it as required to insert or retrieve it from the device. People with poor motor control may be unable to place a semifragile object accurately into the device and retrieve without damage (e.g., bending of a credit card). People with severe weakness may have difficulty reaching the slot or positioning the object for insertion or removal. People who are blind may be unable to determine proper orientation or alignment for insertion (e.g., hotel key card may be held upside down, backward, or at the wrong angle).

### Design Options and Ideas to Consider

- *Facilitating orientation and insertion*
  - Ensuring that objects can be inserted (and removed) with minimal user reach and dexterity
  - Providing a simple funneling system or other self-guidance/orienting mechanism that will position the object properly for insertion
  - Allowing receptacles to be repositioned or reangled to be more reachable
  - Whenever possible, allowing the object to be inserted in several ways (e.g., a six-sided wrench can be positioned on a mating bolt six different ways; two-sided keys can be inserted upside down)
  - Providing visual contrast between the insertion point and the rest of the device (making a more obvious "target")
  - Clearly marking the proper orientation both visually and tactilely
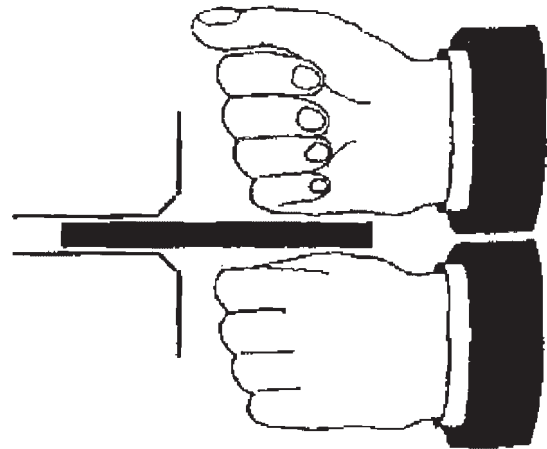


**Figure 17**   Mechanisms that eject items at least 1 in. and preferably 2 in. facilitate grasping the item with tools, reachers, teeth, or fists for those who cannot use their hands or fingers effectively.

- *Facilitating removal*
  - Providing ample ejection distance to facilitate easy gripping and removal (ejection distance as large as possible while retaining a stable ejection; see Figure 17)
  - Using pushbutton ejection or automatic (motorized) ejection mechanism
- *Facilitating handling*
  - Making objects to be inserted rugged and able to take rough handling
  - Using objects with high friction surfaces for ease in grasping

### 9.3.2   M-2 Maximize the Number of People Who Can Physically Handle and/or Open the Product (Operate)

**Problem**   Handles, doorknobs, drawers, trays, and so on, may be impossible for some people to grasp or open.

*Example*: People using mouthsticks or other assistive devices may be unable to grasp handles, doorknobs, and so on, in order to open or operate the product and may find it impossible to open doors or drawers without handles (e.g., those using recessed "lips" or those utilizing only side pressure to open). People with limited arm and hand movement (e.g., due to arthritis or cerebral palsy) may have problems grasping handles that are in-line (straight). People with only one hand or with poor coordination may have difficulty opening products that require two simultaneous actions (e.g., stabilizing while opening or operating two latches that spring closed).

### Design Options and Ideas to Consider

- Using doors with open handles, levers, or doors that are pushed, then spring open
- Avoiding use of knobs or lips to open products

- Avoiding dual latches that must be operated simultaneously
- Using latches that are operable with a closed fist
- Using bearings for drawers or heavy objects that must be moved
- Providing electric pushbutton or remote control power openers
- Shaping product and door handles, and so on, to minimize the need for bending the wrist or body
- Using components that can be operated with either hand

See Section I-5 for additional suggestions.

### 9.3.3   M-3 Maximize the Number of People Who Can Remove, Replace, or Reposition Often-Used Detachable Parts (Operate)

***Problem***   Covers, lids, and other detachable parts may be difficult to remove, replace, or reposition.

*Example*: People with poor motor control may be unable to replace a cover or lid once it has been detached because it was dropped to the floor or into an inaccessible part of the product. People with weakness may have difficulty repositioning a keyboard, monitor, or TV set if the resistance to movement is high.

***Design Options and Ideas to Consider***

- Employing devices with covers or lids that could be hinged, have sliding covers, or be operated electronically
- Tethering covers and lids with a cord or wire
- Making device components repositionable with a minimum of force
- Eliminating or limiting tasks needed for consumer assembly, installation, or maintenance of product

### 9.3.4   M-4 Maximize the Number of People Who Can Understand How to Carry Out the Manipulations Necessary to Use the Product (Understand)

***Problem***   Some people may have difficulty remembering how to operate the product, performing tasks in the correct order or within the required time, making choices, doing required measurements, or problem solving.

*Example*: Some people (particularly those with learning disabilities or cognitive impairments) have difficulty remembering passwords or codes required to operate a device (e.g., PIN for automated teller machine). They may also be unable to remember which control to push to start or stop the device or have difficulty with serial order recall (the ability to remember items or tasks in sequence) and thus cannot follow complex or numerous steps. Others may have a slower or delayed reaction time, due to their inability to remember things quickly or to make responses that are dependent on timed input. Some get confused when there is a time lag for a response after they issue a command or when they expect an immediate result and have trouble in choosing from available selection options (e.g., selecting paper size on a printer, choosing settings on a stereo set). Some cannot understand the concept of measuring or quantifying. Some have significant difficulty finding out what and where the problem is when a device is not functioning properly and may have difficulty identifying solutions to problems they have identified.

***Design Options and Ideas to Consider***   Many of the problems in this category are similar to the problems outlined in Section I-6 and many of the same design ideas would apply, including the following:

- Keeping things as simple as possible
- Providing cues or prompts for sequences of actions required
- Writing the instructions directly on the device
- Having programmable keys for commonly used sequences
- Providing an easy way out of any situation
- Eliminating any timed responses (or make the times adjustable)
- Providing feedback to the user when the device is busy or "thinking"
- Hiding seldom-used controls which are not used primarily, to limit available choices

Other design suggestions include:

- Incorporating premeasuring methods whenever a quantifiable amount is required
- Providing prompts to inform users about the source(s) of problems and lead them to action to be taken to solve the problems (e.g., lights and color-coded pictorials used in copying machines)
- Eliminating or simplifying consumer assembly, installation, and maintenance of the product
- Providing a "standard" key or default mode to operate standardized functions (e.g., a key on a copier to give standard-sized copies)
- Providing an automatic mode so that the machine will make self-adjustments

## 9.4   Documentation

Maximize the number of people who can . . .

D-1 access the documentation.
D-2 understand the documentation.

### 9.4.1   D-1 Maximize the Number of People Who Can Access the Documentation (Perceive)

***Problem***   Printed documentation (e.g., operating or installation instructions) may not be readable.

*Example*: People with low vision may not be able to read documentation due to small size or poor format.

Poor choice of colors may make diagrams ambiguous for people with color blindness. People who are blind cannot use printed documentation, especially graphics. People with severe physical impairments may find it difficult or impossible to handle printed documentation.

### Design Options and Ideas to Consider

- Providing documentation in alternate formats: electronic, large-print, audio tape/disc, and/or braille
- Using large fonts
- Using sans serif fonts
- Making sure that letter spacing, the space between lines (leading), and the distance between topics are sufficient that the letters and topics stand out from each other distinctly
- Supplementing any information that is presented via color coding so it can be interpreted in some other way which does not rely on color (e.g., bar charts may use various black-and-white patterns under the colors or patterns in the colors)
- Providing a text description of all graphics (this is especially important for use in electronic, audio, and large-print forms)
- Providing basic instructions directly on the device as well as in the documentation
- Making printed documentation "scanner/OCR friendly"

### 9.4.2 D-2 Maximize the Number of People Who Can Understand the Documentation (Understand)

**Problem** Printed documentation (e.g., operating or installation instructions) may not be understandable.

*Example*: People with cognitive impairments may have difficulty following multistep instructions. People with language difficulties or for whom English is a second language (including people with deafness) may have difficulty understanding complex text. People with learning difficulties may have difficulty distinguishing directional terms.

### Design Options and Ideas to Consider

- Providing clear, concise descriptions of the product and its initial setup
- Providing descriptions that do not require pictures (words and numbers used redundantly with pictures and tables), at least for all the basic operations
- Formatting with plenty of white space used to create small text groupings and bullet points
- Highlighting key information by using large, bold letters and putting it near the front of the text
- Frontloading key information by putting it near the front of the text
- Providing step-by-step instructions which are numbered or bulleted or have check boxes

- Using affirmative instead of negative or passive statements
- Keeping sentence structure simple (i.e., one clause)
- Supplying a glossary
- Avoiding directional terms (e.g., left, right, up, down) where possible
- Providing a Quick Start or basic "bare bones" form or section to the documentation that gets you up and running with just the basic features

See also Sections O-6, I-6, and M-4.

## 9.5 Safety

Maximize the number of people who can . . .

S-1 perceive hazard warnings (Perceive).

S-2 use the product without injury due to unperceived hazards or the user's lack of motor control (Operate).

### 9.5.1 S-1 Maximize the Number of People Who Can Perceive Hazard Warnings (Perceive)

**Problem** Hazard warnings (alarms) are missed, due to single-sensory presentation or lack of understandability.

*Example*: People who are deaf may not hear auditory alarms. People with hearing impairments may not hear auditory alarms that have only a narrow frequency spectrum. People with visual impairments may not see visual warnings. People with cognitive impairments may not understand the nature of a warning quickly enough.

### Design Options and Ideas to Consider

- Using a broad-frequency spectrum with at least two frequency components between 500 and 3000 Hz for alarm signals
- Using redundant visual and auditory format for alarms (e.g., flashing lights plus alarm siren)
- Reducing glare on any surfaces containing warning messages
- Using common color-coding conventions and/or symbols along with simple warning messages
- Providing an optional, portable, vibrating module for use by persons who are deaf

### 9.5.2 S-2 Maximize the Number of People Who Can Use the Product without Injury Due to Unperceived Hazards or the User's Lack of Motor Control (Operate)

**Problem** Users are injured because they are unaware of an "obvious" hazard or because they lack sufficient motor control to avoid hazards.

*Example*: People with visual impairments may not see a hazard that is obvious to those with average sight. People with lack of strength or muscle control may inadvertently topple a device while in use so that it injures them. People with a lack of muscle control may inadvertently put their limbs or fingers in places not

**Table 2  Principles of Universal Design**

**Principle 1:** Equitable Use

The design is useful and marketable to people with diverse abilities.

    Guidelines:

    1a.   Provide the same means of use for all users: identical whenever possible; equivalent when not.

    1b.   Avoid segregating or stigmatizing any users.

    1c.   Make provisions for privacy, security, and safety equally available to all users.

    1d.   Make the design appealing to all users.

**Principle 2:** Flexibility in Use

The design accommodates a wide range of individual preferences and abilities.

    Guidelines:

    2a.   Provide choice in methods of use.

    2b.   Accommodate right- or left-handed access and use.

    2c.   Facilitate the user's accuracy and precision.

    2d.   Provide adaptability to the user's pace.

**Principle 3:** Simple and Intuitive to Use

Use of the design is easy to understand, regardless of the user's experience, knowledge, language skills, or current concentration level.

    Guidelines:

    3a.   Eliminate unnecessary complexity.

    3b.   Be consistent with user expectations and intuition.

    3c.   Accommodate a wide range of literacy and language skills.

    3d.   Arrange information consistent with its importance.

    3e.   Provide effective prompting and feedback during and after task completion.

**Principle 4:** Perceptible Information

The design communicates necessary information effectively to the user, regardless of ambient conditions or the user's sensory abilities.

    Guidelines:

    4a.   Use different modes (pictorial, verbal, tactile) for redundant presentation of essential information.

    4b.   Provide adequate contrast between essential information and its surroundings.

    4c.   Maximize ''legibility'' of essential information.

    4d.   Differentiate elements in ways that can be described (i.e., make it easy to give instructions or directions).

    4e.   Provide compatibility with a variety of techniques or devices used by people with sensory limitations.

**Principle 5:** Tolerance for Error

The design minimizes hazards and the adverse consequences of accidental or unintended actions.

    Guidelines:

    5a.   Arrange elements to minimize hazards and errors: most used element most accessible; hazardous elements eliminated, isolated, or shielded.

    5b.   Provide warnings of hazards or errors.

    5c.   Provide fail-safe features.

    5d.   Discourage unconscious action in tasks that require vigilance.

**Principle 6:** Low Physical Effort

The design can be used efficiently and comfortably and with a minimum of fatigue.

    Guidelines:

    6a.   Allow user to maintain a neutral body position.

    6b.   Use reasonable operating forces.

    6c.   Minimize repetitive actions.

    6d.   Minimize sustained physical effort.

**Principle 7:** Size and Space for Approach and Use

Appropriate size and space are provided for approach, reach, manipulation, and use regardless of user's body size, posture, or mobility.

    Guidelines:

    7a.   Provide a clear line of sight to important elements for any seated or standing user.

    7b.   Make reach to all components comfortable for any seated or standing user.

    7c.   Accommodate variations in hand and grip size.

    7d.   Provide adequate space for the use of assistive devices or personal assistance.

*Source*: B. R. Connell, M. Jones, R. Mace, J. Mueller, A. Mullick, E. Ostroff, J. Sanford, E. Steinfeld, M. Story, and G. Vanderheiden, © 1997 NC State University, The Center for Universal Design.

intended for contact or other hazardous places (e.g., the cassette tape drive of a stereo or VCR contains sharp edges that can cut fingers jammed inside with force). People with cognitive impairments may be unable to remember to shut off devices when not in use.

### Design Options and Ideas to Consider

- Avoiding pinch points on moving parts
- Eliminating or audibly warning of hazards that rely on the user's visual ability to avoid
- Making all surfaces, corners, protrusions, and device entrances free of sharp edges or extreme heat
- Deburring any internal parts accessible by a body part, even if contact with a body part is not normally expected (e.g., inside an open cassette tape door on a stereo set)
- Providing automatic shutoff of devices that would present a hazard if left on (e.g., irons)
- Ensuring that devices have stable, nonslip bases, or the ability to be attached to a stable surface
- Providing guards that are difficult to defeat where components present a danger of injury

### 9.6 Universal Design Tools

A group of architects, product designers, and human factors engineers have gotten together to develop a common set of universal design principles and guidelines (see Table 2). Members of the team also have developed tools that work with the principles including a Guide to Evaluating the Universal Design Performance of Products. The most current version of the principles and guidelines can be found at the author's website, listed at the end of the chapter.

The Trace Center at the University of Wisconsin–Madison has also developed an online design tool. The tool facilitates the design of more accessible mainstream products by highlighting aspects that contribute to enhanced and expanded usability. It also provides strategies, techniques, and examples for various product types.

### 10 CONCLUSION

Universal design should not really exist as a separate topic. In fact, it is just an extension of good human factors design today. The fact that it is currently a separate topic is probably an artifact of both the heavy military influence in the early ergonomic design process and the focus on serving the largest and most homogeneous segment of the population. However, legislation and commercial interests, a shifting and aging population, and the high costs of health care are combining to provide increased emphasis on this area. In the computer area, Apple, IBM, Microsoft, and other computer companies are all expanding the human interface and general design of their products to allow them to accommodate people with a much wider range of skills and abilities. Similarly, homebuilders, household product manufacturers,

and others are extending and modifying their lines to serve people with more diverse abilities. There is an acute shortage, however, of people with background and experience in what might be universal design. It is to be hoped that over time the term and the field of universal design will fade as it becomes part and parcel of the standard design process.

### 11 FOR FURTHER INFORMATION

Further information on topics covered in this chapter as well as updated versions of design guidelines, resource materials, and references can be found at http://trace.wisc.edu/.

### Acknowledgments

### REFERENCES

American Foundation for the Blind (1996), "Demographics Update: Estimated Number of Adult Braille Readers in the United States," *Journal of Visual Impairment & Blindness*, Vol. 90, p. 287.

Anderson, D. W., Ellenberg, J. H., Leventhal, C. M., Reingold, S. C., Rodriguez, M., and Silberberg, D. H. (1992), "Revised Estimate of the Prevalence of Multiple Sclerosis in the United States," *Annals of Neurology*, Vol. 31, pp. 333–336.

Anderson, T. P. (1981), "Stroke and Cerebral Trauma: Medical Aspects," in *Handbook of Severe Disability*, W. C. Stolov and M. R. Clowers, Eds., U.S. Department of Education, Rehabilitation Services Administration, Washington, DC.

Bennett, C. A. (1977), "The Demographic Variables of Discomfort Glare." *Lighting Design and Application*, Vol. 7, pp. 22–24.

Berkowitz, J. P., and Casali, S. P. (1990), "Influence of Age on the Ability to Hear Telephone Ringers of Different Spectral Content," *Proceedings of the Human Factors Society 34th Annual Meeting*, Vol. 1, pp. 132–136.

Braddock, D. L. (2002), "Public Financial Support for Disability at the Dawn of the 21st Century," *American Journal on Mental Retardation*, Vol. 107, No. 6, pp. 478–489.

Brault, M. W. (2008, December), "Americans with Disabilities: 2005, Household Economic Studies," P70-117, U.S. Census Bureau, Washington DC, available: http://www.census.gov/prod/2008pubs/p70-117.pdf.

Bureau of Labor Statistics (2010, August 25), "Persons with a Disability: Labor Force Characteristics Summary," press release, available: http://www.bls.gov/news.release/disabl.nr0.htm.

Caplan, R. (1992, August), "Disabled by Design," *Interior Design*, Vol. 63, pp. 88–91.

Congdon, N., et al. (2004), "Causes and Prevalence of Visual Impairment among Adults in the United States," *Archives of Ophthalmology*, Vol. 122, pp. 477–485.

Cook, A. M., and Polgar, J. M. (2007), *Cook & Hussey's Assistive Technologies: Principles and Practice*, 3rd ed., Mosby-Year Book, St. Louis, MO.

Cronin, S., Hardiman, O., and Traynor, B. J. (2007), "Ethnic Variation in the Incidence of ALS: A Systematic Review," *Neurology*, Vol. 68, pp. 1002–1007.

de Lau, L. M., and Breteler, M. M., (2006), "Epidemiology of Parkinson's disease," *The Lancet Neurology*, Vol. 5, pp. 525–535.

Donnan, G. A., Fisher, M., Macleod, M., and Davis, S. M. (2008), "Stroke," *The Lancet*, Vol. 371, No. 9624, pp. 1612–1623.

Faul, M., Xu, L., Wald, M. M., and Coronado, V. G. (2010), *Traumatic Brain Injury in the United States: Emergency Department Visits, Hospitalizations and Deaths 2002–2006*, Centers for Disease Control and Prevention, National Center for Injury Prevention and Control, Atlanta, GA.

Foundation for Spinal Cord Injury Prevention, Care & Cure (2009), "Spinal Cord Injury Statistics: Spinal Cord Injury Facts & Figures at a Glance 2009," available: http://www.fscip.org/facts.htm, retrieved September 9, 2010.

Galvin, J., and Scherer, M. (1996), *Evaluating, Selecting, and Using Appropriate Technology*, Aspen, Gaithersburg, MD.

Grandjean, E., Ed. (1987), *Ergonomics of Computerized Offices*, Taylor & Francis, Bristol, PA.

Harding, G. F. A., and Jeavons, P. M. (1994), *Photosensitive Epilepsy*, MacKeith Press, London.

Hauser, W. A. (1994), "The Prevalence and Incidence of Convulsive Disorders in Children," *Epilepsia*, Vol. 35, No. S2, pp. S1–S6.

Hodgkinson, R. (2009), "10th Report on International ICT Accessibility Standards: Proposed, Being Developed and Recently Published," available: http://tiresias.org/research/standards/report_10.htm.

Hoover, R. E., and Bledsoe, C. W. (1981), "Blindness and Visual Impairments," in *Handbook of Severe Disability*, W. C. Stolov and M. R. Clowers, Eds., U.S. Department of Education, Rehabilitation Services Administration, Washington, DC.

Hunt, R. M. (1970), "Determination of an Effective Tone Ringer Signal," paper presented at the 38th Convention of the Audio Engineering Society, New York.

International Organization for Standardization/International Electrochemical Commission (ISO/IEC) (2009a), "ISO/IEC TR 29138-1:2009: Information Technology—Accessibility Considerations for People with Disabilities—Part 1: User Needs Summary," ISO, Geneva.

International Organization for Standardization/International Electrochemical Commission (ISO/IEC) (2009b), "ISO/IEC TR 29138-2:2009: Information Technology—Accessibility Considerations for People with Disabilities—Part 2: Standards Inventory," ISO, Geneva.

Jeavons, P. M., and Harding, G. R. A. (1975), *Photosensitivity Epilepsy*, Heinemann, London.

Kraus, J. F., and MacArthur, D. L. (1996), "Epidemiologic Aspects of Brain Injury," *Neurologic Clinics*, Vol. 14, No. 2, pp. 435–450.

Kraus, L., and Stoddard, S. (1989), *Chartbook on Disability in the United States*, prepared for U.S. Department of Education, National Institute on Disability and Rehabilitation Research, Washington, DC.

Kraus, L., Stoddard, S., and Gilmartin, D. (1996), *Chartbook on Disability in the United States*, prepared for U.S. Department of Education, National Institute on Disability and Rehabilitation Research, Washington, DC.

Kroemer, K. H. E. (1990), *Engineering Physiology Bases of Human Factors/Ergonomics*, Van Nostrand Reinhold, New York.

LaPlante, M. P. (1988), *Data on Disability from the National Health Interview Survey, 1983-85*, U.S. Department of Education, National Institute on Disability and Rehabilitation Research, Washington, DC.

Larson, S. A., Lakin, K. C., Anderson, L., Kwak, N., Lee, J. H., and Anderson, D. (2001), "Prevalence of Mental Retardation and Developmental Disabilities: Estimates from the 1994/1995 National Health Interview Survey Disability Supplements," *American Journal on Mental Retardation*, Vol. 106, pp. 231–252.

Lawrence, R. C., et al. (1998), "Estimates of the Prevalence of Arthritis and Selected Musculoskeletal Disorders in the United States," *Arthritis and Rheumatism*, Vol. 41, pp. 778–799.

Mueller, J. (1990), *The Workplace Workbook: An Illustrated Guide to Job Accommodation and Assistive Technology*, RESNA Press, Washington, DC.

Odding, E., Roebroeck, M. E., and Stam, H. J. (2006), "The Epidemiology of Cerebral Palsy: Incidence, Impairments and Risk Factors," *Disability & Rehabilitation*, Vol. 28, pp. 183–191.

Pastor, P. N., and Reuben, C. A. (2002), "Attention Deficit Disorder and Learning Disability: United States, 1997–98," *Vital Health Statistics*, Series 10, No. 206, National Center for Health Statistics, Hyattsville, MD.

Pellegrino, L. (2002), "Cerebral Palsy," in *Children with Disabilities*, 5th ed., M. L. Batshaw, Ed., Paul H. Brooks Publishing, Baltimore, MD, pp. 443–466.

Ries, P. W. (1994), "Prevalence and Characteristics of Persons with Hearing Trouble: United States, 1990-91," *Vital Health Statistics*, Series 10, No. 188, National Center for Health Statistics, Hyattsville, MD.

Schein, J. D. (1981), "Hearing Impairments and Deafness," in *Handbook of Severe Disability*, W. C. Stolov and M. R. Clowers, Eds., U.S. Department of Education, Rehabilitation Services Administration, Washington, DC.

Schow, R. L., et al. (1978), *Communication Disorders of the Aged: A Guide for Health Professionals*, University Park Press, Baltimore, MD.

Sondick, E. J., Maddans, J. H., and Gentleman, J. F. (2010), "Summary Health Statistics for US Adults: National Health Interview Survey, 2009," *Vital Health Statistics*, Series 10, No. 249, National Center for Health Statistics, Hyattsville, MD.

Vanderheiden, G. C. (2009a), "Accessible and Usable Design of Information and Communication Technologies," in *The Universal Access Handbook*, C. Stephanidis, Ed., CRC Press, Boca Raton, FL, pp. 3-1–3-26.

Vanderheiden, G. C. (2009b), "Standards and Guidelines," in *The Universal Access Handbook*, C. Stephanidis, Ed., CRC Press, Boca Raton, FL, pp. 54-1–54-21.

World Health Organization (WHO) (1980), *International Classification of Impairments, Disabilities, and Handicaps: A Manual of Classification Relating to the Consequences of Disease*, WHO, Geneva.

Wyndaele, M., and Wyndaele, J. (2006), "Incidence, Prevalence and Epidemiology of Spinal Cord Injury: What Learns a Worldwide Literature Survey?" *Spinal Cord*, Vol. 44, pp. 523–529.

Ziegler-Graham, K., MacKenzie, E. J., Ephraim, P. L., Travison, T. G., and Brookmeyer, R. (2008), "Estimating the Prevalence of Limb Loss in the United States: 2005 to 2050," *Archives of Physical Medicine and Rehabilitation*, Vol. 89, pp. 422–429.

## UNIVERSAL DESIGN RESOURCES

*The Accessible Housing Design File*, 1991, Barrier Free Environments, Wiley, New York.

*Beautiful Barrier-Free: A Visual Guide to Accessibility*, 1992, C. Liebrock and S. Behar, Wiley, New York.

*Building for a Lifetime: The Design and Construction of Fully Accessible Homes*, 1994, M. Wylde, A. Baron-Robbins, and S. Clark, Taunton, Newton, CT.

*Building Sight: A Handbook of Building and Interior Design Solutions to Include the Needs of Visually Impaired People*, 1995, P. Barker, J. Barrick, and R. Wilson, American Foundation for the Blind Press, New York.

*A Consumer's Guide to Home Adaptation*, 1992, Adaptive Environments Center, Inc., Boston, MA, available: http://www.adaptenv.org/index.php?option=Resource &articleid=219&topicid=25.

*Culture and Environment*, 1984, I. Altman and M. M. Chemers, Cambridge University Press, Cambridge.

*Design for Dignity: Studies in Accessibility*, 1993, W. Lebovitch, Wiley, New York.

"Designing Consumer Product Displays for the Disabled," in *Proceedings of the Human Factors Society 34th Annual Meeting*, 1990, J. T. Ward, Vol. 1, Human Factors Society, Santa Monica, CA pp. 448–451.

*Designing for Diversity: Gender, Race, and Ethnicity in the Architectural Profession*, 2001, K. H. Anthony, University of Illinois Press, Champaign, IL.

*Design Like You Give a Damn: Architectural Responses to Humanitarian Crises*, 2006, K. Stohr and C. Sinclair, Eds., Metropolis Books, New York.

*Design Meets Disability*, 2009, G. Pullin, MIT Press, Cambridge, MA.

*Design for the Other 90%*,, 2007, C. E. Smith., Editions Assouline, Paris.

*Emotional Design: Why We Love (or Hate) Everyday Things*, 2004, D. A. Norman, Basic Books, New York.

*Ergonomics at Work*, 1987, D. J. Osborne, Wiley, New York.

*The Eyes of the Skin: Architecture and the Senses*, 2nd ed., 2005, J. Pallasmaa, Wiley Academy, Chichester, England.

*The Hidden Dimension*, 1990, E. T. Hall, Anchor Books, New York.

*House Form and Culture*, 1969, A. Rapoport, Prentice Hall, Englewood Cliffs, NJ.

*Human Factors in Engineering and Design*, 6th ed., 1987, M. S. Sanders and E. J. McCormick, McGraw-Hill, New York.

*Human Factors Research Needs for an Aging Population*, 1990, S. J. Czaja, Ed., Panel on Human Factors Research Issues for an Aging Population, Committee on Human Factors, Commission on Behavioral and Social Sciences and Education, National Research Council, National Academies Press, Washington, DC, available: http://books.nap.edu/catalog/1518.html.

*Inclusive Design: Designing and Developing Accessible Environments*, 2001, R. Imrie and P. Hall, Francis & Taylor, London.

*Inclusive Housing: A Pattern Book*, 2010, E. Steinfeld and J. White, W.W. Norton & Co., New York.

*Just Below the Line: Disability, Housing, and Equity in the South*, 2010, K. H. Smith, J. Webb, and B. T. Williams, University of Arkansas Press, Fayetteville.

*Making Life More Livable: Simple Adaptations for Living at Home after Vision Loss*, 2001, I. R. Dickman, American Foundation for the Blind Press, New York.

*Play for All Guidelines: Planning, Design and Management of Outdoor Play Settings for All Children*, 1992, R. Moore, S. Goltsman, and D. Oacofano, Eds., MIG Communications, Berkeley, CA.

*Practicing Universal Design: An Interpretation of the ADA*, 1994, W. L. Wilkoff and L. W. Abed, Wiley, New York.

*The State of the Science: Emerging Research and Development in Universal Design*, in press, J. Maisel, Ed., Bentham Sciences, available: http://www.bentham. org/ebooks/9781608050635/index.htm.

*Symbol Sourcebook: An Authoritative Guide to International Graphic Symbols*, 1972, H. Dreyfuse, McGraw-Hill, New York.

*Tactile Graphics*, 1992, P. K. Edman, American Foundation for the Blind Press, New York.

*Touch, Representation, and Blindness*, 2000, M. A. Heller, Oxford University Press, New York.

*Transgenerational Design*, 1994, J. J. Pickle, Van Nostrand Reinhold, New York.

*Universal Design: 17 Ways of Thinking and Teaching*, 2002, J. Christophersen, Ed., Norwegian State Housing Bank (Husbanken), Oslo, Norway.

*Universal Design File*, 1998, M. F. Story, J. L. Mueller, and R. Mace, Center for Universal Design, North Carolina State University, Raleigh, NC, available: http://www.design.ncsu.edu/cud/.

*Universal Design Handbook*, 2nd ed., 2010 (2001, 1st ed.), W. Preiser, and E. Ostroff, Eds., McGraw-Hill Professional, New York.

*Universal Design Newsletter*, quarterly, Universal Designers and Consultants, Takoma Park, MD, available: http://www.universaldesign.com/newsletters.php.

*Universal Design New York*, 2001, G. S. Danford and B. Tauke, Eds., City of New York Office of the Mayor, New York, available: http://udeworld.com/dissemination/ publications.html.

*Universal Design New York 2*, 2003, D. Levine, Ed., City of New York Office of the Mayor, New York, available: http://udeworld.com/dissemination/publications.html.

*Universal Kitchen and Bathroom Planning: Design that Adapts to People*, 1998, M. J. Peterson, McGraw-Hill, New York.

## ORGANIZATIONS

**AccesslT**
National Center on Accessible Information
    Technology in Education
Box 357920
University of Washington
Seattle, WA 98195-7920
866-968-2223 (V)
866-866-0162 (TTY)
http://www.washington.edu/accessit/index.php/

AccesslT, at the University of Washington, promotes the use of electronic and information technology (E&IT) for students and employees with disabilities in educational institutions at all academic levels. The website contains a searchable, growing database of questions and answers regarding accessible E&IT. Funding for AccesslT is provided by the National Institute on Disability and Rehabilitation Research and the National Science Foundation.

**Center for Assistive Technology and Environmental Access (CATEA)**
Georgia Institute of Technology
School of Architecture
490 Tenth Street, NW
Atlanta, GA 30332-0156
404-894-4960 (V/TTY)
http://www.catea.gatech.edu/

CATEA is a multidisciplinary engineering and design research center dedicated to enhancing the health, activity, and participation of people with functional limitations through the application of assistive and universally designed technologies in real-world environments, products, and devices. CATEA's work is organized under four laboratories: the Rehabilitation Engineering and Applied Research Laboratory (REAR Lab; http://rearlab.gatech.edu/), the Accessible Workplace Laboratory (http://www.catea.gatech.edu/work.php), the Enabling Environments Laboratory (EE lab; http://www.catea.gatech.edu/eelab.php), and the Accessible Education and Information Laboratory (http://www.catea.gatech.edu/aei.php).

**Center for Inclusive Design and Environmental Access (IDeA)**
378 Hayes Hall, School of Architecture and Planning
3435 Main Street
University at Buffalo
Buffalo, NY 14214-3087
716-829-5902
http://www.ap.buffalo.edu/idea/

The IDeA Center is dedicated to making environments and products more usable, safer, and healthier in response to the needs of an increasingly diverse population. The IDeA Center's activities are based on the philosophy of inclusive design, often called "universal design" or "design for all." It is a way of thinking that can be applied in any design activity, business practice, program, or service involving interaction of people with the physical, social, or virtual worlds. Maintains the Universal Design E-World site, which is a participatory environment with Web-based tools to support the community of practice in universal design.

**Center for Universal Design**
College of Design
North Carolina State University
Campus Box 8613
Raleigh, NC 27695-8613
919-515-8359
http://www.ncsu.edu/www/ncsu/design/sod5/cud/

The Center for Universal Design (CUD) is a national information, technical assistance, and research center that evaluates, develops, and promotes accessible and universal design in housing, commercial and public facilities, outdoor environments, and products. Its mission is to improve environments and products through design innovation, research, education, and design assistance.

**Inclusive Technologies**
Temper Complex
37 Miriam Drive
Matawan, NJ 07747
908-907-2387 (V/TTY)
http://www.inclusive.com/

Inclusive Technologies provides a full range of consulting services to companies, public agencies, consumers, researchers, purchasers, and policymakers on how mainstream products can better meet the needs of all users, including users with disabilities and elders.

**Institute for Human Centered Design**
200 Portland Street
Boston, MA 02114
617-695-1225 (V/TTY)
http://humancentereddesign.org/

The Institute for Human Centered Design (IHCD), founded in 1978 as Adaptive Environments, is an international nongovernmental educational organization committed to advancing the role of design in expanding opportunity and enhancing experience for people of all ages and abilities through excellence in design. IHCD's work balances expertise in legally required accessibility with promotion of best practices in human-centered or universal design.

IHCD has been the lead organization in the international universal design movement, having hosted or cohosted five international conferences as well as international student design competitions, smaller regional meetings, and publication of Web and print materials.

**J.L. Mueller, Inc.**
4717 Walney Knoll Court
Chantilly, VA 22021
703-222-5808
http://www.jlmueller.com/index.html/

J.L. Mueller, Inc. was founded in 1982 by Jim Mueller, an industrial designer who has worked in the field of design for people with disabilities since 1974. He is recognized as one of the most experienced designers in working with people with disabilities to increase independence at home, at school, and at work through design. The website includes information about the principles of universal design, applicable legislation, and workplace accommodations.

### National Center for Accessible Media (NCAM)
http://ncam.wgbh.org/

The Carl and Ruth Shapiro Family National Center for Accessible Media (NCAM) at Boston public broadcaster WGBH is a research and development facility dedicated to addressing barriers to media and emerging technologies for people with disabilities in their homes, schools, workplaces, and communities. NCAM is part of the Media Access Group at WGBH which includes two production units, The Caption Center (est. 1972) and Descriptive Video Service® (DVS®) (est. 1990).

### R.L. Mace Universal Design Institute (UDI)
410 Yorktown Drive, Suite 203
Chapel Hill, NC 27516
919-960-6734 (office)
http://udinstitute.org/index.php/

The Ronald L. Mace Universal Design Institute is a nonprofit organization based in North Carolina dedicated to promoting the concept and practice of accessible and universal design. The institute's work manifests the belief that all new environments and products, to the greatest extent possible, should and can be usable by everyone regardless of age, ability, or circumstance.

The institute advances the concept of universal design in all design disciplines, including housing, public use buildings, outdoor and urban environments, and related products.

### Technology Access Program
Gallaudet University
800 Florida Avenue, NE
Kendall Hall
Washington, DC 20002
202-651-5257 (V/TTY)
http://tap.gallaudet.edu/

The Technology Access Program (TAP) conducts research related to communication technologies and services, with the goal of producing knowledge useful to industry, government, and deaf and hard-of-hearing consumers in the quest for equality in communications.

### Trace Research and Development Center
University of Wisconsin—Madison
2107 Engineering Centers Building
1550 Engineering Drive
Madison, WI 53706
608-262-6966 (V)
608-263-5408 (TTY)
http://trace.wisc.edu/

The Trace R&D Center at the University of Wisconsin—Madison focuses on the design of mainstream information and communication technology for use by all people. Trace is also the home of the Rehabilitation Engineering Research Center (RERC) on Information Technology Access and (in partnership with Gallaudet University) the RERC on Telecommunication Access, both funded by the National Institute on Disability and Rehabilitation Research.

### Universal Design Education Online
http://www.udeducation.org/

This site supports educators and students in their teaching and study of universal design. It provides a place where educators can interact with each other and where a growing community of learners exchange information for the benefit of all. The site is designed for use by faculty members, students (of any age and stage), and user/experts. It supports professional design education as well as professional development/continuing education, featuring a variety of materials for a range of disciplines, levels, and interests

### Web Accessibility Initiative
http://www.w3.org/WAI/

The Web Accessibility Initiative (WAI) works with organizations around the world to develop strategies, guidelines, and resources to help make the Web accessible to people with disabilities. It is one of the four domains of the World Wide Web Consortium (W3C); the WAI develops its work through W3C's consensus-based process, involving different stakeholders in Web accessibility. These include industry, disability organizations, government, accessibility research organizations, and more.

### WebAIM (Web Accessibility in Mind)
http://webaim.org/

WebAIM's mission is to expand the potential of the Web for people with disabilities by providing the knowledge, technical skills, tools, organizational leadership strategies, and vision that empower organizations to make their own content accessible to people with disabilities. They are known for their online instructional media and software tools.

# CHAPTER 52

# DESIGN FOR AGING

**Walter R. Boot**
**Florida State University**
**Tallahassee, Florida**


**Timothy A. Nichols**
**Microsoft Game Studios**
**Redmond, Washington**


**Wendy A. Rogers and Arthur D. Fisk**
**Georgia Institute of Technology**
**Atlanta, Georgia**

## 1 INTRODUCTION

Age is a critical variable relevant to design considerations in human factors research and practice. This conclusion is founded on three primary facts:

1. The number of older adults in developed countries today is higher than ever and is increasing.
2. There are critical age-related differences between younger and older adults which necessitate specific design considerations.

3. The proportion of older adults within the global workforce and of all users of systems and products is increasing steadily (Figure 1).

### 1.1 Increasing Population over Age 65

The world's older adult (65+ years) population is increasing by approximately 800,000 each month (Kinsella and Velkoff, 2001). U.S. Census data show clearly the increasing trend for adults over age 65 and over age 85, as the average lifespan is increasing (U.S. Bureau of the Census, 1996; Figure 2).
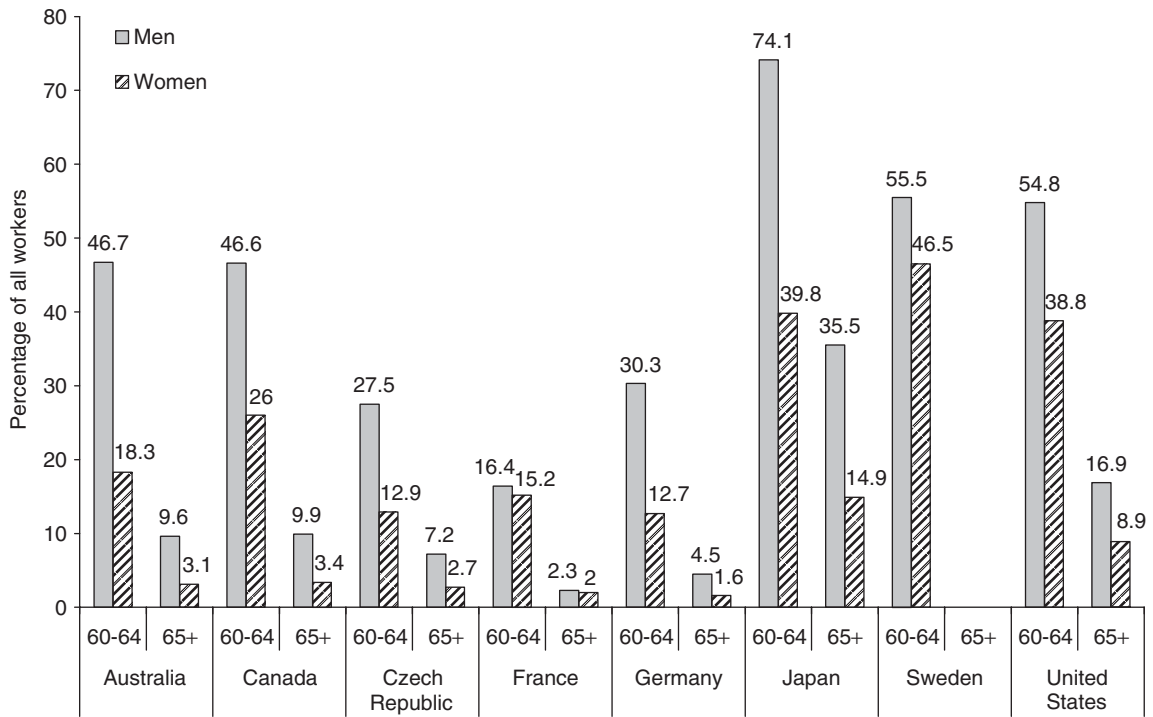
**Figure 1**    Percentage of men and women between 60 and 64 and 65+ in selected countries who are in the workforce.
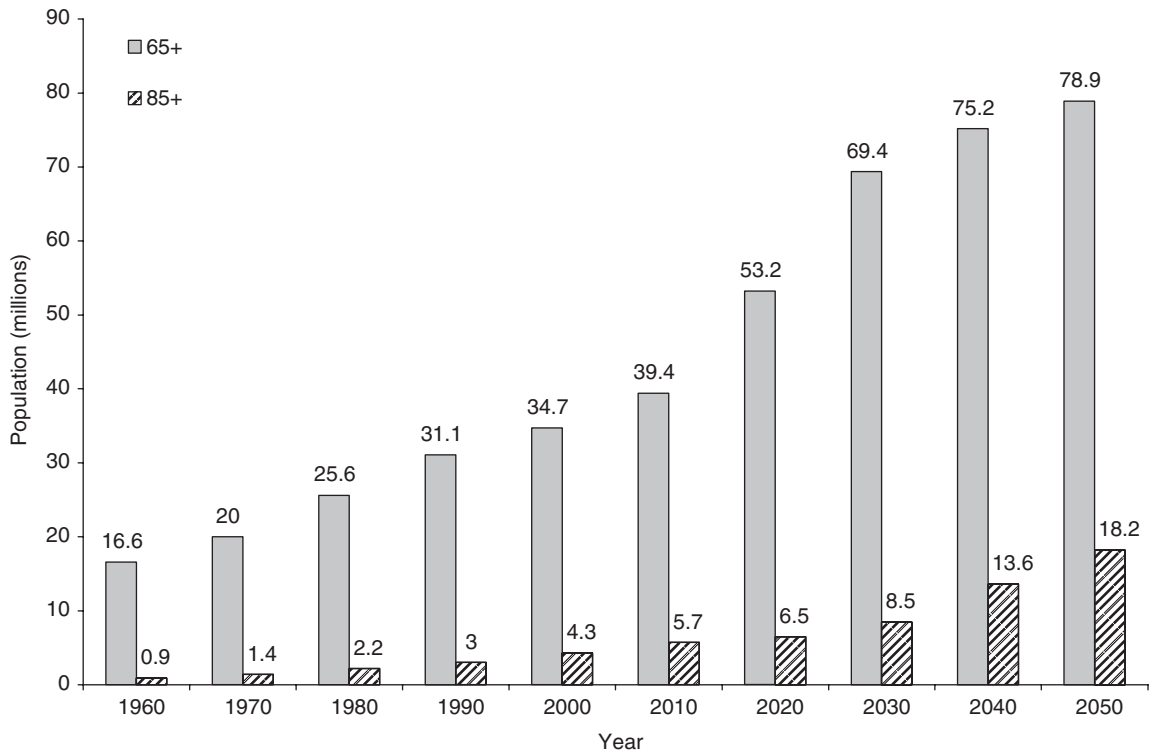


**Figure 2**    Population estimates for the age groups 65+ and 85+ through the year 2060.

## 1.2 Design-Critical Age-Related Differences

Why is it important to consider older adult users? The statistics of older adult population rates alone are not cause for the field of human factors to take notice. The meaningful issue is whether older adults require significantly different design considerations than younger adults. This issue is addressed in the present chapter by specifying the cognitive, perceptual, motor, and motivation differences between the two age groups and also considering whether these differences translate into functional differences. For example, age-related response time differences of 1.5 s may not be functionally meaningful when searching for information on the Internet, but this age-related difference could be critical in the driving environment or when responding to a medical emergency in the home where delayed responses can lead to serious consequences. Primary goals of this chapter are to describe functionally meaningful changes that occur with age and their effects on performance and to recommend design considerations and design principles.

## 1.3 Older Adults' Use of Technology

Given the need for age-specific design considerations, a critical third issue for human factors researchers and practitioners is to understand the extent to which older adults interact with systems and products. Certainly, if older adults comprise only a very small proportion of users of systems and products, this would temper the degree to which the field must be concerned with the age-related changes and characteristics of older adults. However, this is not the case. In a survey of product use (Hancock et al., 2001), adults over age 65 reportedly used a variety of household products as frequently as adults under age 65, and they used health care products more frequently. Furthermore, as discussed in Section 6, older adults express a desire to learn to use new technologies (e.g., Rogers et al., 1998) and many are using the computer and the Internet for a variety of tasks (e.g., Olson et al., 2011). In summary, the capabilities and limitations of the growing older adult population must be understood and accounted for in the design process and in human factors research to ensure that this segment of the population can interact with products and systems in a safe, efficient, and effective manner.

## 1.4 Definitions

### 1.4.1 Age

The focus of the chapter is on the relevance of "age" to design. However, age is no more than a chronological marker for the myriad changes that people undergo as they age—hence the term age-*related* changes. The changes discussed here are not caused by a person's age but rather are correlated with age, and there is substantial variability in the degree to which individual older adults show some changes and the degree of change for a particular person. Furthermore, given the continuous nature of this variable, is it proper or even useful to think of age in discrete chunks (e.g., younger vs. older)? In fact, there are several reasons to do so (see Nichols et al., 2003, for a review). For research purposes, we recommend defining "older" adults as 65–85 years,

"middle-aged" adults as 40–55, and "younger" adults as 18–30. Subgroups such as young–old (e.g., 56–64) and old–old (85+), are often used in aging research, and the younger group may at times be extended (e.g., 18–39). These age range decisions are often, in part, based on the user population to which the research is intended to generalize as well as the age-related changes that are being investigated. For example, younger and older groups in a study of air traffic controllers would be based on different age ranges than a study of normal driving. The standard maximum age for an air traffic controller is 56, whereas the maximum age of drivers extends into the 80s and 90s. Thus, what is older in one context differs from what is older in the other context. We recommend that researchers and practitioners follow the above age range guidelines for the following reasons:

- *Variability in Performance*. Age-related variance will be controlled to some degree (as opposed to a study that uses a sample ranging from 40 to 75). As will be highlighted, considerable age-related changes occur across the lifespan, and those changes may have a significant impact on behavior and task performance.

- *Precision and Consistency*. By following appropriate age classification guidelines and reporting participants' age, the field of human factors will better provide precise and useful information on the performance of adults of different ages and design will better fit the user.

- *Parsimony*. For many studies, it may be simpler to think of a single variable than the host of variables affected by age. Although age is merely a marker predicting performance, age provides a broad and useful designate for many of the age-related changes discussed in this chapter.

### 1.4.2 Products and Systems

The design issues discussed in this chapter are intended to be relevant to a wide range of products and systems (i.e., any system or product older adults might use). This is necessarily general, as older adults interact with a wide range of products and systems in their daily activities. Although the guidelines are intended to be general, their applicability will differ as a function of the demands of the task at hand (e.g., a technology that requires fine motor control will be more influenced by age-related changes in movement control). Guidelines specific to particular devices and contexts are available in Fisk et al. (2009).

## 1.5 How to Use the Recommendations

The recommendations presented in this chapter provide a means of reducing the potential solution space for resolving design-related questions. However, these recommendations do not provide a complete solution. Variability at all levels of behavior in older adults is a hallmark characteristic of aging, making iterative design and user testing with older adults crucial. The design issues discussed here should be beneficial in understanding and predicting older adults' performance, but

the chapter is not intended to replace necessary human factors methods such as iterative design and user testing. For the implications and design recommendations, our goal is to provide general design principles relevant to design for older users. However, many of the age-related changes discussed in this chapter occur gradually and improvements in design and training targeted at older adults are likely to benefit younger and middle-aged adults as well (Fisk, 1999).

## 1.6 Overview of the Chapter

The remainder of this chapter is organized as follows: First we discuss aspects of behavior from the perspective of age-related differences. Within each section we present general implications for design. We then present a "case study" of a hand-held gaming system designed for older adults. This case study illustrates the implications of the age-related changes for system design.

### 1.6.1 Age-Related Changes

To guide the discussion of age-related changes, consider a general model of information processing as presented in Figure 3. The three general categories of activities are perceptual encoding, central processing, and responding; each is influenced to some degree by the normal aging process. As such, the following categories of age-related changes are reviewed briefly and then discussed in terms of the relevance of such changes to design: (1) perception, (2) cognition, and (3) movement control. In addition, given their general overarching role we also discuss briefly (4) beliefs, attitudes, and motivation.

### 1.6.2 Design Implications and Suggestions

Following each section on age-related changes, the implications of these data for design will be described. We also present an in-depth hypothetical case study of a hand-held gaming system for older adults. While the discussion accompanying each age-related change will involve specific consequences for design, the goal of the latter case study is to provide robust design recommendations based on consideration of multiple age-related changes.

In providing design suggestions, there is a tension between providing overly general suggestions that are not specific enough to be useful versus overly specific suggestions that do not generalize across products or systems. Our approach is to focus on implications and recommendations for those systems and products particularly relevant to a given age-related change. Thus, when we present design implications at the end of each section, these will be reasonably broad (i.e., the design implications for the selective attention section will be relevant to the design of automobiles but not to the design of furniture).

### 1.6.3 Consistent Themes and Guidelines

There are consistent themes that are relevant to design considerations for an older population (Table 1). These themes are evident in many of the design suggestions and implications in this chapter.
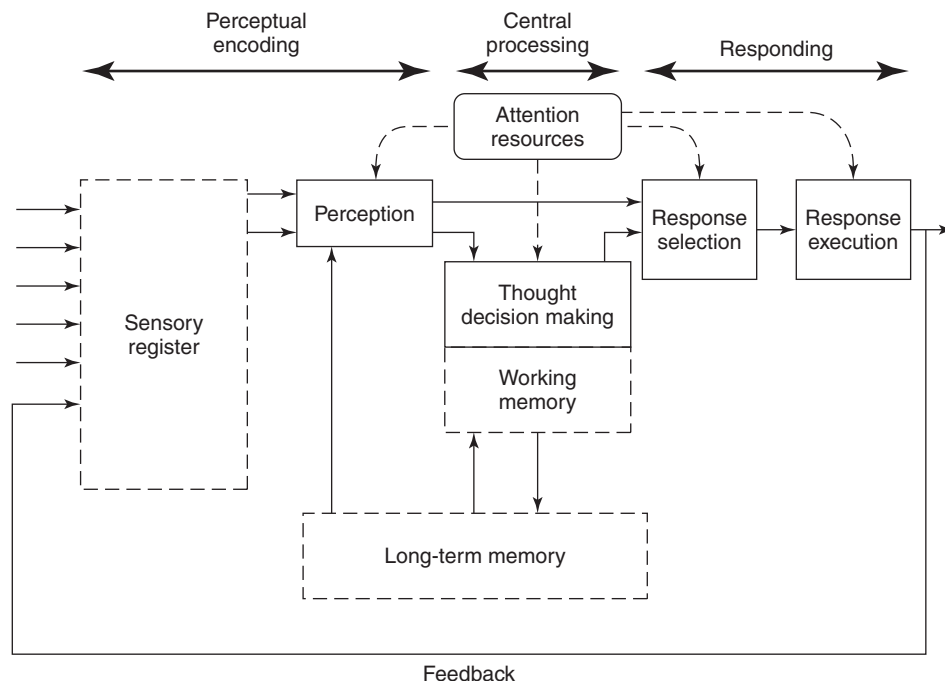


**Figure 3**   A schematic representation of information processing. (Adapted from Wickens et al., 1998.)

**Table 1 Six Major Themes of Designing for Aging**

| Design Theme | Definition |
|---|---|
| Provide environmental support through context, cues, and organization | Performance can be supported by placing information in the task environment, which reduces cognitive demands on the user. |
| Improve the physical stimulus | Age-related perceptual declines can be offset to some degree by improving the physical stimulus and increasing older adults' ability to perceive and recognize the stimulus. |
| Display information consistently | Some level of consistency is required for learning to occur, and with greater degree of consistency, learning is more efficient. |
| Provide training | Through appropriate training, older adults' performance can be brought closer in line with that of younger adults. |
| Capitalize on crystallized knowledge | Fact knowledge is generally unaffected by aging, and design can take advantage of this knowledge. |
| Recognize knowledge, interest, and motivation | Older adults' motivation and desire to interact with technology are often underestimated. |

**Table 2 Age-Related Perceptual Changes in Vision, Audition, and Haptics**

| | |
|---|---|
| ***Visual Changes*** | |
| Visual acuity | The ability to resolve detail decreases. |
| Visual accommodation | The ability to focus on close objects decreases. |
| Color vision | The ability to discriminate and perceive shorter wavelength light decreases. |
| Contrast detection | The ability to detect contrast decreases. |
| Dark adaptation | The ability to adapt quickly to darker conditions decreases. |
| Glare | The susceptibility to glare increases. |
| Illumination | More illumination is required to see adequately. |
| Motion perception | Motion is not as readily detected and motion estimation is reduced. |
| Useful field of view | The useful visual field is reduced. |
| ***Auditory Changes*** | |
| Auditory acuity | The ability to detect sound decreases, particularly at higher frequencies and particularly for males. |
| Auditory localization | The ability to localize sound decreases, particularly at higher frequencies and when directly in front of or behind the user. |
| Audition in noise | The ability to perceive speech and complex sounds decreases. |
| ***Haptic Changes*** | |
| Haptic control | Older adults have more variability in maintaining constant force when grasping an object. |
| Proprioceptive perception | Threshold to differentiate being touched at a single point versus two points is higher for older adults. |
| Temperature perception | Thresholds increase with age. |
| Vibration perception | Thresholds increase with age. |

## 2 PERCEPTION

Most products and systems are designed to provide information via the visual and auditory modalities, and these two sensory systems have been well studied in the aging population (see Table 2 for a summary of the visual and auditory changes; see Fisk et al., 2009, for more in-depth discussion). These systems show significant and substantial declines in older adults. For information to affect behavior, the information must first enter the sensory system and then be encoded. If, due to sensory system degradation, information is sensed or perceived incompletely, it may be processed incorrectly.

Haptic perception (i.e., the sense of touch) also provides an important channel for information. Consider the vehicle shaking when the car is misaligned or the feedback from a computer keyboard that informs the typist if the key has been pressed sufficiently. Haptic cues are also being designed into systems to provide information such as rumble strips on the roadway or the vibration of a mobile phone. As such, we need to understand the degree to which older adults show changes in these perceptions (see Table 2).

### 2.1 Vision

In general, the ability to resolve an image accurately is dependent on the available luminance and the contrast of the scene. The most common age-related causes of visual impairment are age-related macular degeneration (ARMD), cataracts, and glaucoma (Desai et al., 2001)

(Figure 4). Due to changes in the structure of the aging eye and visual processing system, older adults are less able to resolve details and are less sensitive to critical environmental characteristics such as luminance, contrast, color, and motion.

### 2.1.1 Acuity

Although visual impairment can affect people across the lifespan, age is the best predictor of visual decline. Visual acuity, the best known measure of visual ability, is typically measured relative to what a "normal" person
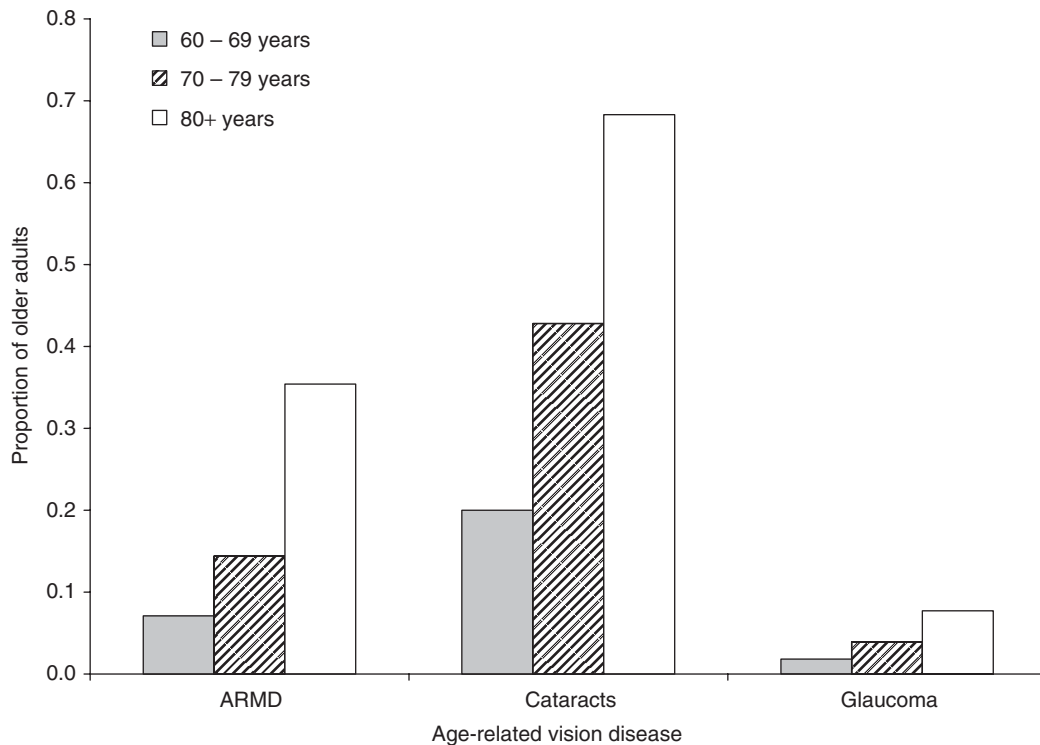
**Figure 4** Prevalence of ARMD, cataracts, and glaucoma across three groups of older adults. (Adapted from Desai et al., 2001.)

can see at 20 ft away (thus, the measure 20/20) (Snellen, 1862, cited in Bennett, 1965). Acuity is affected by various eye pathologies (many of which are more commonly observed with age) as well as deterioration of the brain's visual pathways. In ARMD, the most common cause of vision loss in older adults, people experience a reduced ability to resolve fine detail (Bellman and Holz, 2001). People with ARMD suffer from reduced central vision due to degeneration of the *macula*, an area in the center of the retina. Cataracts also affect acuity, resulting in a clouding of the lens, but these are often treatable, whereas vision loss due to ARMD cannot currently be restored. Glaucoma results from damage to the optic nerve due to an increase in pressure in the eye, as fluid flow through the front of the eye is hindered. The buildup of fluid at the front of the eye results in increased pressure at the back of the eye, causing irreparable degeneration of the optic nerve fibers (Goldstein, 1999). As a result, peripheral vision deteriorates, and if not treated, central vision will deteriorate as well.

Older adults are able to compensate for loss of acuity. For example, despite poorer visual acuity, older adults were shown to perceive blurred text signs better than younger adults (Kline et al., 1999). In this study, the acuity of both younger and older adults was reduced artificially and the size at which blurred text signs could be read was measured. The primary findings were that both age groups were better able to read familiar text

signs than novel text signs and, when both age groups' acuity was comparably impaired (e.g., at 20/40), older adults could read text at a smaller size than younger adults. Presumably the nature of optical blur is such that low-vision persons are less affected (Legge et al., 1987a). Thus, older adults' ability to deal with blurred information may be related to their ability to adapt to changes in visual acuity as well as to rely on *top-down processing* (i.e., interpreting information based on well-learned, crystallized knowledge).

***Design Implications and Suggestions*** The loss in acuity has profound effects on the way in which information should be displayed for older adults. Increasing the size, brightness, and contrast of an item will improve older adults' perception of information. For example, text should typically be displayed in a 12-point font size or greater. It is especially important for novel information to be made perceptually salient for older adults. Top-down processing can result in the correct identification of a perceptually indistinct stimulus (Kline et al., 1999), but when processing is primarily bottom-up (as with novel stimuli), clarity of the stimulus is crucial. For example, a driver who has reduced near vision may not be able to perfectly discriminate the points on the odometer, but the consistent spacing and common look of the odometer should provide sufficient information about the vehicle's speed. Thus, it is important to design for such

consistent aspects of a system, because users who rely on these consistent aspects to aid their top-down processing interpretation of displays can be supported (this is a form of environmental support).

If a display cannot be changed to accommodate low-acuity people, it is important to maximize the effect of top-down processing. Contextual cues, another form of environmental support, can be provided to increase the likelihood that a stimulus will be recognized. For example, a driver who has trouble reading traffic control signs from afar may rely on the color and shape of the signs, which allows the driver to identify the type of sign. However, acuity is not the only age-related decrement in visual perception that must be considered.

### 2.1.2   Accommodation

Older adults have difficulty with visual accommodation (termed *presbyopia*), which involves adjusting the curvature of the lens to focus on objects of different depths. In fact, reductions in accommodative ability are primarily responsible for losses in acuity in near vision, typically starting at age 40. By age 65, lens accommodation is so reduced that only objects at a certain distance can be focused on the retina, meaning that information not displayed at this distance cannot be clearly perceived by the person (Schneider and Pichora-Fuller, 2000).

***Design Implications and Suggestions***   The need to focus at different distances should be minimized as much as possible. In a system with multiple displays, all of the displays should be placed as close to the optimal reading distance from the user as possible. This will reduce the necessity for head movement to bring information into perfect clarity as the multiple displays are scanned.

### 2.1.3   Color Vision

Older adults are less able to discriminate shorter wave-length light, such as blues and greens, due to a yellowing of the lens (Said and Weale, 1959). Furthermore, the ability to discriminate color declines with age (Kraft and Werner, 1999), and age-related differences in color discrimination increase at lower light levels and lower color saturations (Pinckers, 1980; Knoblauch et al., 1987).

***Design Implications and Suggestions***   Color coding is still a feasible option in information visualization (e.g., representing multidimensional data at a single time), but color codes should avoid shorter wavelength light as a general rule or use only a single blue or green, thus eliminating the need to compare within this range of color. Color coding should not be used when many distinct levels are required, and colors should be well saturated.

### 2.1.4   Contrast Detection

The importance of high contrast in detecting stimuli and in tasks such as reading is important for adults of all ages, but particularly for older adults. People with poorer acuity are more affected by reductions in contrast whether they are younger adults (Legge et al., 1987b) or older adults (Fozard, 1990). However, even if matched for visual acuity, older adults have reduced contrast sensitivity (Mitzner and Rogers, 2003). Reduced contrast sensitivity is due in part to the scattering of light as it enters the eye, such that light from the image is scattered across the retina, creating a more uniform dispersal of light on the retina (Schneider and Pichora-Fuller, 2000).

***Design Implications and Suggestions***   Optimal contrast in which text should be presented is a bright white on black, or vice versa. If colors are used to present information, colors close together on the spectrum should not be used together (e.g., a red icon on an orange background). Furthermore, the deleterious effects of reduced contrast can be alleviated through the use of context, essentially allowing top-down processes to aid the older user in identifying the stimulus correctly (Mitzner and Rogers, 2003).

### 2.1.5   Illumination, Glare, and Dark/Light Adaptation

In the aging eye, changes in the cornea scatter light before it reaches the retina, the lens absorbs more light, and pupil size is reduced, allowing less light to reach the retina (Schneider and Pichora-Fuller, 2000). Even though less light reaches the retina, glare is problematic for older adults. Glare occurs when a person is exposed to levels of light higher than the eye is currently adapted. Thus, glare is a notable problem when driving at night. Similarly, bright sunlight can cause glare from either direct line of sight or reflected surfaces. Due to the scattering of light in the older eye, glare is more of a problem for older adults, as the excess light is more distributed across the eye, essentially reducing perception for a greater degree of the visual field. The aging eye also is slower to adapt to light and dark. With reductions in the amount of light that reaches the retina, the retina is therefore slower to adapt to the changing light conditions. Furthermore, the chemical processes that cause dark adaptation are slowed with age (Jackson et al., 1999).

***Design Implications and Suggestions***   These age-related changes in the sensation of light cannot be solved via an external perceptual aid, as they are specifically due to deterioration of the cornea. Older adults will benefit from increased illumination in all environments, including driving and activities at home and work. Unfortunately, Charness and Dijkstra (1999) demonstrated that the homes of older adults are not lit optimally, particularly at night (although older adults appeared to compensate for their visual impairments by using significantly more light in their homes than younger adults).

Appropriate lighting is critical in optimizing the perception of information. If possible, increase the level of illumination to at least 100 cd/m$^2$, as measured by the reflection from reading surfaces (Charness and Dijkstra, 1999). Lighting levels should be even when possible, in roadways and in office and home layouts. To reduce glare, light sources should be diffused and positioned to create ambient light as opposed to direct light. Mirrors and shiny surfaces should be avoided, as the undiffused reflections can cause glare. Multiple light sources serve

to reduce harsh shadows and to even out the light in the environment.

These age-related changes in vision are critically relevant in the driving domain, which is one of the most perceptually and cognitively demanding tasks as well as one of the most widely and frequently performed. When driving at night, illumination levels are already extremely low. Glare can occur as the result of passing headlights and streetlights, causing older adults to be blinded temporarily by the dispersed light across their retinas. When driving from daylight into a tunnel, older adults' visual perception will suffer relative to younger adults. In the design of roads and tunnels, lighting should be made as constant as possible to reduce the negative effects of low illumination, glare, and slower adaptation to dark and light.

### 2.1.6  Perception of Motion

Older adults are less able to detect motion relative to younger adults. To investigate sensitivity to motion, researchers have created sets of elements, a subset of which move synchronously (Trick and Silverman, 1991; Tran et al., 1998). Older adults appear to have a higher element threshold than younger adults, indicating that they required more movement of elements to detect motion in the array.

Age-related differences in perception of motion have been investigated in a driving scenario (Atchley and Andersen, 1998; Andersen et al., 2000). A three-dimensional display was used to simulate a vehicle windshield and display an environment at one of two velocities followed by a constant deceleration. The deceleration occurred such that the car would either stop short of an obstacle, stop directly at the obstacle, or crash into the obstacle. The display was stopped prior to stopping or crashing, and the participants were required to indicate what the outcome would be. Relative to younger adults, older adults were less sensitive to detecting collisions, indicating that a crash was inevitable when no collision occurred.

***Design Implications and Suggestions***  Motion detection findings suggest that in combination with age-related declines in movement speed and visual perception, older adults' declining perception of potential collision information could be a factor in their ability to safely avoid vehicular collisions. When motion is a critical cue, it must be accentuated for older adults. However, this design implication must be incorporated carefully because, all other things being equal, motion can be more of a distraction for older adults than for younger adults.

### 2.2  Audition

Auditory information is presented in a wide variety of environments. Museums and exhibition halls may play descriptive recordings at displays, training materials may include a video with a model explaining how to perform some task, computers in the office or home emit alerts and other auditory signals, and many systems rely on auditory stimuli to communicate system status information. Safe and efficient system interaction can depend on the user's

ability to hear normally, but audition is another perceptual domain wherein older adults show declines.

### 2.2.1  Auditory Acuity

In addition to an overall decrement in auditory acuity, age-related losses in hearing occur differentially across frequency ranges, with greater loss occurring for higher frequencies (greater than 8000 Hz) (Schneider and Pichora-Fuller, 2000; Fozard and Gordon-Salant, 2001). Furthermore, men have worse high-frequency perception than women (Moscicki et al., 1985).

***Design Implications and Suggestions***  A high-frequency stimulus is sometimes used as an alert or indicator in computer applications. Older adults, particularly males, may not perceive these stimuli. In fact, the age-related changes in auditory acuity suggest that high-frequency sounds should be avoided in any system or product that older adults might use. A high-frequency alert or indicator may be differentiated from background noise better than lower frequencies, but if it is not perceived, it is useless. If auditory stimuli are designed to attract attention when the user's vision is elsewhere, auditory alerts should not exceed 4000 Hz (Fisk et al., 2009). For non-alert-related stimuli, it is important to provide the user with control over the intensity of the stimulus. Because of individual differences in overall thresholds, volume control should be provided so that people can calibrate for themselves. However, often it is not sufficient to provide overall volume control but rather the ability to modulate various frequencies.

### 2.2.2  Localization

Data suggest that older adults are less adept at localizing sounds in space, specifically being prone to front/back localization errors (Abel et al., 2000). When high-frequency deficits occur, localization is more difficult in the elevation dimension (up vs. down) than in the azimuth (right vs. left) (Noble et al., 1994). Furthermore, higher frequency stimuli are harder to localize for all ages because high-frequency stimuli reach both ears at the same time (Lorenzi et al., 1999).

***Design Implications and Suggestions***  The reduced ability to localize high-frequency sounds is another reason to avoid high-frequency auditory stimuli. When an auditory stimulus is intended to direct the older user's attention to the source of the stimulus, the stimulus should be presented at between 5000 and 8000 Hz. Furthermore, auditory stimuli designed to orient attention should not be presented directly behind or in front of the user. This is especially relevant in workstations or other scenarios where the user is likely to remain in the same space. Sounds that must be localized should be presented for durations long enough for people to turn their heads and localize the sound, thus eliminating the error-prone front/back scenario.

### 2.2.3  Degraded Stimulus Environment

Noises are not often pure auditory stimuli. Many auditory signals as well as speech occur within a noisy environment—for example, at a workstation with the

hum of computer fans and conversing co-workers in the background. Research has shown that older adults have greater difficulty than younger adults perceiving speech in such degraded auditory conditions. There is some debate concerning whether the locus of the difference is primarily cognitive or perceptual (Schneider et al., 2000). Regardless of the absolute locus of the effect, noise degrades auditory perception more for older than for younger adults.

***Design Implications and Suggestions***   This difficulty in perception under noisy conditions demonstrates the importance of using cues in the visual modality instead of the auditory when presenting information in a potentially noisy environment. However, auditory cues can be used to augment visual cues via redundant or dual coding. Dual coding is beneficial even in quiet environments, as users' visual attention may be directed elsewhere when information needs to be communicated to them. Younger adults may perceive an auditory stimulus easily, but older adults will have more difficulty. Speech perception specifically can be hindered in high-noise environments, particularly if the people have poor hearing (Schneider et al., 2000). However, given that the locus of the problem is perceptual (as opposed to cognitive), age-related differences are likely to be evident for auditory stimuli other than speech.

For optimal perception, the signal should be presented independent of any noise. For example, in training materials, there should be no sound except for the relevant instructional materials (e.g., no background music). If the auditory signal can be amplified independent of background noise, users should be offered this capability (e.g., headphones at a museum display). If this is not possible (such as in the automated speech in an elevator or subway car), text should be presented to provide redundant information. Compressed speech is more difficult for older adults to perceive [although Sharit et al. (2003) indicate that 10% compression has little effect on young, middle-aged, or older adults]. It is recommended that speech rates not exceed 140 words per minute (Fisk et al., 2009). In public presentation of information, where ambient speech and other noise may be present, provide wireless headphones to amplify the signal, if feasible. Sound-absorbing materials on floors, walls, and ceilings may be used.

## 2.3   Haptics

Haptics can be defined as the sense of touch. Haptic sensitivity is assessed in a number of ways including thresholds (Grünwald, 2008). In many instances, older adults have higher thresholds for detecting an increase in temperature or an increase in vibration. Likewise, the ability to detect being touched by a single point versus two points shows a decline for older adults.

Grasping an object and maintaining a constant force require haptic control which also shows age-related deficits. Older adults have even more difficulty maintaining force control when simultaneously engaged in a cognitively demanding task (Voelcker-Rehage et al., 2006).

***Design Implications and Suggestions***   Together, these age-related changes in haptics degrades the quality of haptic information processing, with implications for successful interaction with technology. For example, if vibration is used as a cue, care should be taken in selecting vibration frequency. Sensitivity to low-frequency (25-Hz) vibration is relatively unimpaired with age through the decade of the 60s, but sensitivity to higher frequency vibration (60 Hz and above) shows linear decline with age from the teenage years.

With respect to sensititivey of touch, there is general more age-related loss in sensitivity for lower limbs compared to upper limbs. Therefore, upper body sites (e.g., hands) should be preferred to lower body ones (e.g., feet) for conveying haptic information.

## 3   COGNITION

### 3.1   Attention

Attention is not a unitary scientific construct. Indeed, there are well-known varieties of attention (James, 1890/1950; Parasuraman and Davies, 1984). Closely linked to visual perception is the construct of useful field of view which incorporates both visual processing speed and attentional capacity.

Two general categories of attention are selective attention and attentional capacity. Research on selective attention has focused on the ability to focus on and process a restricted set of goal-relevant information while ignoring available information that is not relevant to the goal (Johnston and Dark, 1986). Attentional capacity research has investigated the amount of "mental work" that humans can perform at a given time, often employing dual-task methods, where the trade-off in performance between the two tasks can provide a measure of attentional capacity given to each task. Selective attention and attentional capacity are affected deleteriously by age (see Rogers and Fisk, 2001, for a review), but certain interventions and design provisions have been shown to reduce these decrements to some degree.

### 3.1.1   Useful Field of View

Useful field of view (UFOV) refers to the size of the visual field that may be perceived in a single glance and is a measure of both processing speed and attention (Owsley et al., 1991; Roenker et al., 2003). That is, UFOV can be thought of as the subset of the total visual field that is available for processing (thus, the similarity with the construct of attention). The UFOV may change within a person, depending on the nature of the task being performed (Owsley et al., 1991). For example, one's UFOV may be larger when driving on a road with no traffic, whereas one may experience the phenomenal sense of constricted vision when driving in the rain and heavy traffic. Research has shown that older adults have a restricted UFOV which has been linked to driving accidents (Owsley et al., 1991; Roenker et al., 2003).

***Design Implications and Suggestions***   The broad implications for changes in UFOV are that practitioners cannot assume that a user will necessarily notice, use, or respond to information falling within the visual field.

Age must be considered, and under some circumstances the UFOV of the user population might need to be assessed directly. With training, as people are able to perform subtasks more efficiently, the UFOV can effectively be increased. Thus, the implications for design are to know the UFOV of the user population in the context of the task, to ensure that stimuli are presented within their UFOV, and if necessary to provide training to increase UFOV for the users.

### 3.1.2 Selective Attention

Selectively attending to goal-relevant stimuli and ignoring goal-irrelevant stimuli are required for efficient performance in any task. Selective attention involves purposefully shifting attention to different stimuli and categories of stimuli in the environment. For example, when driving, a person may be actively searching for certain elements or groups of elements, such as other cars, pedestrians, and traffic signs and signals. Irrelevant stimuli such as a car alarm or brightly colored or waving advertising sign can distract the driver momentarily. The degree of susceptibility to distraction and the duration of the distraction can obviously have severe consequences in this task domain and in others. Older adults are susceptible to distracting effects of irrelevant stimuli in the environment (Rogers and Fisk, 2001).

***Selective Inhibition Deficits*** There has been considerable research in the field of cognitive aging, suggesting that older adults have relatively more difficulty inhibiting irrelevant information (e.g., Hasher and Zacks, 1988; Stoltzfus et al., 1993). However, there is not a general deficit of inhibition because certain inhibitory systems appear unaffected by aging (e.g., Connelly and Hasher, 1993; Kramer et al., 1994). For example, younger and older adults were equally able to adjust their focus of attention to include or exclude information (Hartley et al., 1992), suggesting that older adults, in this case, did not have greater difficulty inhibiting irrelevant information. Older adults were also able to inhibit the location of irrelevant stimuli as well as younger adults, although they were impaired relative to younger adults in their ability to suppress the identities of irrelevant stimuli (Connelly and Hasher, 1993).

This reduced efficiency in selectively deploying attention can have deleterious consequences for performance, especially in highly attention-demanding tasks such as driving. Older adults seem to be inefficient in searching novel visual environments, for example, searching in the same area repeatedly (Maltz and Shinar, 1999), suggesting that they are not monitoring where they have searched previously. In this driving-based task, adults were less likely to maintain attention in a search task, less likely to discriminate previously attended areas, and less likely to attend selectively all relevant areas of the display. It is important to note that after extensive consistent training search performance is still slower for older adults, but the qualitative aspects of the search, such as the learning curve, are similar for younger and older adults (Fisk and Rogers, 1991).

***Attending to Multiple Tasks*** Younger and older adults appear to deploy attention differently across multiple simultaneous tasks after training (Sit and Fisk, 1999). Younger and older participants viewed a screen with a different task in each quadrant and were encouraged (via a points-based system) to attend primarily to one task over the other three (but overall performance was encouraged as well). After training on the multiple-task display, older adults were still performing at a lower level than younger adults, although age-related differences in performance had lessened over training. They were then required to attend more to a different task at transfer—that is, the task demands did not change, but the way in which attention was deployed was changed. Older adults focused primarily on the newly important task to the exclusion of the rest of the tasks. Younger adults were able to perform effectively on three of the four tasks, including the newly important task. The findings suggested: (1) age-related differences in a divided attention task requiring different selective attention allocation strategies are attenuated after training and (2) older adults are less able than younger adults to strategically change the way in which they perform a previously learned task.

***Design Implications and Suggestions*** Age-related changes in selective attention must be carefully considered in any environment in which multiple stimuli are presented. Tasks and environments with multiple displays and controls, such as driving, cockpits, security and surveillance tasks, medical displays, and industrial control panels, all require the user to attend to a subset of a multitude of auditory and visual stimuli in the environment.

One way to improve older adults' ability to select goal-relevant stimuli from a distraction-filled environment is to increase the perceptual salience of the goal-relevant stimuli (or decrease the salience of distracting stimuli) (Shaw, 1990, 1991). When the physical stimulus is improved, the contrast between goal-relevant and goal-irrelevant stimuli is increased, and the relevant cue can be used more efficiently. This form of environmental support can help guide a user's selective attention to relevant stimuli (for review see Morrow and Rogers, 2008).

In addition to increasing the perceptual salience of relevant stimuli, whenever possible, these stimuli should be made salient to the user via explicit means as well. For example, older users should be told how best to differentiate and ignore distracting stimuli (i.e., essentially giving older adult users a specific strategy to follow). Consider using an instructional manual to put together a complex product. The manual may include information about safety, rebates, warranty information, and so on, in addition to the specific, sequential assembly information. In designing this manual, the actual assembly steps should be perceptually salient, relative to other information in the manual. For example, assembly steps should be boxed in a particular color or set off with bold header information. This enables older users to better identify the relevant assembly information in a field of irrelevant information (for the assembly

task). Furthermore, the designers of the manual should explicitly inform the user of the cues they should be looking for, as opposed to requiring users to recognize the relevance or meaning of different cues on their own.

Older adults' attention is more likely to be drawn to perceptually salient stimuli in the task environment. It is important to minimize the attention-attracting nature of irrelevant stimuli. Attention-attracting perceptual characteristics include flashing, moving, bright, loud, and unexpected stimuli. Through training, older adults can improve their ability to successfully select a subset of relevant stimuli. This critical finding suggests that the decrement in selective attention is, at least in part, a more labile age-related change and that, with training, older adults are capable of selecting important stimuli among distracting stimuli. Hence, in situations requiring selective attention, training should be given to users until the criterion performance level is reached.

### 3.1.3  Attentional Capacity

**Divided Attention**   Users can attend to and cognitively process a limited amount of information at a time. The hypothetical construct of attentional resources has been used to explain the capacity to process, think about, and cognitively manipulate information at a given time (Wickens, 1984). Older adults are presumed to have a reduction in the processing resources available to them to perform attention-demanding tasks (e.g., Crossley and Hiscock, 1992). For example, older adults have relatively more difficulty maintaining appropriate levels of performance when required to perform multiple tasks at once (Kramer and Larish, 1996), that is, under divided-attention conditions. Clearly, older adults experienced a greater decrement in performance transferring from single to dual tasks than did younger adults (see, e.g., McDowd and Craik, 1988). In a study of dual-task performance, testing the efficacy of various vehicular travel aids for different age groups (e.g., automated visual map aids, synthetic speech, paper maps), older drivers had more safety-related errors in this dual-task environment than did younger adults (Dingus et al., 1997). However, when redundant auditory guidance was provided in addition to the automated map aid, older adults performed more safely than without the redundant information.

**Visual Clutter**   Even within what seems to be a single task, attention can be overloaded and performance can suffer. In a search-type task performed on a busy, noisy visual display, people are required to orient and reorient their attention as they scan the display. As stimuli become more similar or increase in number, identifying and comparing stimuli become more demanding (Shiffrin, 1988). In general, older adults have more difficulty in higher clutter environments (e.g., Schieber and Goodspeed, 1997). In an investigation of the effect of clutter in a typical driving scene (i.e., a typical two-lane rural highway, a commercial district in a small city, and a downtown metropolis scene), older adults' speed and accuracy at detecting target signs were considerably poorer than younger adults (Schieber and Goodspeed, 1997). This pattern was observed whenever even a small amount of clutter was present.

**Automatic Processing**   With repeated exposure to consistent stimulus mappings, people can be trained to automatically attend to and respond to stimuli that are consistently meaningful in the task (e.g., targets in a search task). This has been termed an automatic attention response (Schneider and Shiffrin, 1977; Shiffrin and Schneider, 1977). Consistent mapping refers to the degree to which stimuli belong to a given category when they appear. In experimental search tasks, consistent mapping occurs when target stimuli never appear as distractor stimuli and distractor stimuli never appear as targets. The concept of consistent mapping (and conversely, varied mapping) extends directly to natural tasks such as driving (brake lights are consistently paired with a slowing of the vehicle) and computing (icons are consistently mapped with the applications they represent, and keyboard keys are consistently mapped with respect to their location and function). For these consistent features in the environment, younger and older users can develop very quick and accurate responses, but older adults' responding will be less efficient than younger adults'.

An automatic attention response can be important in the fast, accurate detection of stimuli in the environment. For example, brake lights may cause a driver automatically to attend to the lights and respond by braking, given enough exposure to consistent instances of brake lights co-occurring with the slowing of the vehicle ahead. Older adults may not develop new automatic attention responses to learned stimuli in visual search tasks, but considerable performance improvements occur under consistent conditions (Fisk et al., 1988; Fisk and Rogers, 1991; Rogers, 1992).

### Design Implications and Suggestions

**Performance Gains with Training**   Consider the plight of novice drivers. In this new environment, with new displays and controls, there are a host of stimuli to monitor, to search for, and perhaps to respond to. Each and every scenario is completely new, and even after some experience, novel instances occur with high frequency (e.g., rarer scenarios such as passing a cyclist or avoiding a large pothole). For these novice drivers, speaking to a passenger (let alone speaking on a mobile phone), adjusting the radio, or waving to someone on the sidewalk can be difficult and error prone (in either the main driving task or the secondary task). However, over time, considerable learning occurs. Similar instances occur multiple times and are remembered, allowing faster responses subsequently. The location of the gas, clutch, and brake pedals are quickly located and never confused. Eventually, attention can safely be divided between driving and other tasks. In the same way, through sufficient, appropriately designed training, performance on a task can become considerably more efficient. Older adults' performance can improve greatly through such experience, and this is the goal of training programs. For example, initial age-related differences in a divided-attention task can be attenuated (though still present) after training (e.g., Rogers et al., 1994; Sit and Fisk, 1999).

**Part-Task Training** To reduce the demands on attention in dual-task conditions, participants can be trained on certain parts of the task at different times before being trained on the whole task—that is, part-task training. For example, participants without computer experience must often undergo mouse training before beginning a study involving computers with mice as the control. Thus, participants are able to devote a majority of their attention to the experimental task instead of dividing their attention between an untrained, novel device and the task. To assess the benefits of part-task training, Kramer et al. (1995) presented a dual-task condition to older adults, with half of the participants required to devote an equal amount of attention to each task and the other half required to pay more attention to one task at certain times and to the other task at other times. Thus, the second group practiced on, essentially, only part of the task at different times throughout training. By the end of training, older adults in the part-task group demonstrated considerably more learning than those in the first group.

**Redundant Information** When information can be presented redundantly in certain dual-task conditions, older adults benefit from the reduction in task demands (e.g., Dingus et al., 1997). Redundant information has been shown to be important in cluttered environments as well. Thus, providing redundant information may be a means of providing environmental support to compensate for age-related reductions in attentional capacity.

**Clutter in the Visual Environment** The effects of clutter in a visual environment such as driving can be very detrimental to performance, particularly for older adults. Older adults spend more time than younger adults searching in a cluttered environment and spend more time making decisions about stimuli (Ho et al., 2001). This can obviously be detrimental to driving performance and to one's safety. However, this situation can be ameliorated if the user is provided attentional cues designed to support the user's performance. This sort of environmental support was tested for younger and older participants in a driving simulation task that involved making quick decisions about whether a left-hand turn could be performed safely (Staplin and Fisk, 1991). When normal driving contextual information was present in the task (i.e., clutter stimuli such as lampposts, trees, and houses), both younger and older adults benefited from receiving a cue about the future status of the intersection, thus aiding their left-hand-turn decision.

### 3.2  Memory

Memory can be divided into three general stages. If information is to be remembered, it must first be encoded, then it must be stored or represented in some way in the brain, and then it must be retrieved from storage. Within this basic framework, researchers have focused on various types of memory, such as memory for information that occurred at a certain time and place, memory for facts, memory for procedures, memory to do something in the future, and memory for the source of information.

**Table 3  Five Conclusions about Memory and Aging**

| Conclusion | Examples |
| --- | --- |
| Older adults maintain their semantic memory, but their ability to retain episodic memories is decreased. | Older adults can remember the names of friends they would like to email, but remembering a new email password will be difficult. |
| Maintenance of episodic memories can be particularly hampered when atypical, distracting elements are present when retrieval is attempted. | Older adults are more likely to forget to pick up milk on the way home if traffic is unusually hectic. |
| The retention of habits, or processes performed automatically, is relatively spared in older adults. | Although learning a new software package may be difficult, older adults retain the learned ability to type. |
| Working memory capacity is reduced in older adults. | Older adults will need to play back a set of auditory instructions more times than younger counterparts. |
| When older adults are required to perform self-initiated processes, memory is more difficult. Environmental support can reduce these difficulties. | Older adults will have difficulty recalling the color, make, and model of a hit-and-run vehicle, but if given a lineup, they will probably be able to identify the correct vehicle. |

*Source*: Adapted from Fisk and Rogers (2002).

Age-related declines in some aspects of memory (such as working memory and episodic memory) are well documented (Zacks et al., 2000). In some cases, these age-related declines can be improved by placing information in the task environment, instead of requiring people to maintain the information in memory (Morrow and Rogers, 2008; see Table 3 for several memory conclusions). In other cases, there are minimal changes across the lifespan, such as in semantic memory (memory for facts) and procedural memory (memory for how to perform an activity or sequence of actions; Zacks et al., 2000).

### 3.2.1  Working Memory

Working memory can be thought of as information that is actively being processed and "used" (Baddeley, 1986). Similar to the concept of capacity limitations on attentional processing, it is typically measured via span tasks, which measure the number of elements that can be kept activated in working memory. Reduced working memory capacity in older adults is a hallmark finding in the cognitive aging literature (see Zacks et al., 2000, for a review), and a meta-analysis of the literature revealed a sizable age-related difference (Verhaeghen et al., 1993).

***Design Implications and Suggestions*** The well-documented age-related decrement in working memory capacity has clear implications for designers. Older users should not be required to keep multiple items in memory. E-commerce websites should provide comparison programs that allow customers to compare similar products as opposed to keeping in memory variables such as price and features. A telephone voice menu should have deep as opposed to broad menu structures, so that users do not have to keep too many options in memory before they make a selection. In general, information should be displayed to the user (i.e., putting the information in the environment), as opposed to requiring users to rely on their working memory.

### 3.2.2   Episodic Memory

Significant age-related differences exist in the ability to recall various events and instances which are referred to as episodic memories. In a typical episodic memory study, participants are shown stimuli and asked to recall them at a later time (see Tulving, 2002, for a review). Age-related deficits in these tasks are commonly found. In this section we discuss two ways that this fundamental age-related difference in memory can be addressed and supported: memory strategies and supportive information.

***Memory Strategies*** When older adults are required to elaborate internally the stimulus to be remembered, it is later recalled with greater accuracy (Park et al., 1990b; Verhaeghen et al., 1993; Dunlosky and Hertzog, 1998). For example, when older adults are required to generate words for later recall as opposed simply to reading a list, they recall more accurately (Hirshman and Bjork, 1988; Johnson et al., 1989). Some studies have demonstrated that older adults may engage in suboptimal encoding strategies, which may explain their relative deficit in remembering (Rogers and Gilbert, 1997; but see Dunlosky and Hertzog, 1998).

In assessing older adults' associative learning ability, Rogers and Gilbert (1997) found that some older adults chose continually to use an inefficient strategy to perform a task, whereas nearly all the younger adults employed the optimal strategy. The task presented a consistent set of word pairs at the top of the screen and a test word pair in the center. Participants were required to indicate whether the test pair was one of the pairs at the top of the screen. The task involved multiple trials, such that it was optimal to attempt to memorize the word pairs above as opposed to searching for a match on each trial. Older adults were less likely than younger adults to adopt optimal strategies spontaneously in this associative learning task, but they were able to use the optimal strategy if encouraged to do so. A larger study of individual differences with this same task replicated the finding that older adults were less likely overall to adopt an optimal strategy (Rogers et al., 2000). However, for those older adults who did adopt the appropriate strategy, the age-related differences in learning were reduced. Taken together, these studies suggest that in certain cases older adults' performance may be improved by providing optimal strategies for performing a task.

***Cognitive Support*** Memory researchers have conducted numerous studies testing the effects of memory cues and aging. Relatively few studies have shown a greater benefit of memory cues for older adults than for younger adults; however, in general, these studies show that older adults' memory can be improved through the use of memory cues. For example, adults of all ages increase their recall when some aspect of the stimulus is present at recall (e.g., a word fragment vs. freely recalling a word); this is the basis for Craik's environmental support framework (Craik, 1986). Memory retrieval has also been shown to increase when stimuli are studied in a visually distinctive context (Park et al., 1990a). Younger and older participants studied a set of different objects on either a colorfully distinctive background or a plain background. They were later asked to place either labeled note cards or the original items back in the original spatial arrangement. Both age groups benefited equally from the distinctive background condition as well from the use of the original items.

In another study, researchers presented target words (i.e., words to be remembered) to older and younger adults, either integrating the target with an object in the environment (e.g., "The key fit the lock on the file cabinet," where a file cabinet was in the test room) or integrating the target with an object not in the environment (e.g., "The key fit the lock on the car," where no car was present in the test room) (Earles et al., 1996). Younger adults benefited more than older adults from this combination of environmental cues and the target word.

***Design Implications and Suggestions***

**Strategy Suggestions** Older adults' episodic recall can be improved through the use of different strategies. One of the more commonly used is a form of encoding elaboration. For example, to remember a pair of words better, it can be helpful to construct a distinctive, imagistic sentence or concept that links the two words (e.g., "dog" and "spoon" → "The dog balanced the spoon on its nose"). Other heuristics can be employed, such as creating acronyms from a series of words to be remembered, repeating a set of instructions, or creating a link between a stimulus and some internal concept. However, many people do not use such memory strategies spontaneously outside the lab. Therefore, a training system should provide strategy suggestions when memory is required.

Although older adults are not as likely as younger adults to adopt optimal strategies spontaneously, they are able to utilize these optimal strategies if encouraged to do so. There are a number of workable methods for encouraging participants to use memory strategies. These include explicitly instructing the participants (Hulicka and Grossman, 1967; Haider and Frensch, 1996; Nichols and Fisk, 2001), pretraining on an orienting task that makes apparent the optimal strategy (Hulicka and Grossman, 1967; Doane et al., 1996; Rogers and Gilbert, 1997), or giving intertask tests that will encourage use of the desired strategy (Rogers and Gilbert, 1997).

**Physical Reminders** Given older adults' knowledge about their memory capabilities, physical reminders can play an important role in older adults' lives. Older adults

reportedly use various physical reminders for medications (Park et al., 1992; Sanchez et al., 2003), and adults over age 71 years benefited most from the physical reminders (Park et al., 1992). Some suggestions for reminders include the following:

1. Physical reminders should be placed in visually salient places, where the person will see them. For example, one can place medications for the following day next to one's toothbrush at night.

2. Given age-related issues in *source monitoring*, the ability to remember the source of an event, the reminder should provide relevant information instead of simply reminding that there is *something* to be remembered.

3. Automated visual and auditory reminders, such as those used in calendaring software, can minimize the need for older adults to search actively for the reminder.

**Environmental Support Framework**  Older adults' memory should benefit from distinctive contexts insofar as the context is present when retrieval occurs. The context serves to provide additional retrieval cues for the person. Thus, if the context is absent, recollection may be harmed. The environmental support framework is based on the notion that older adults appear to have more difficulty with effortful, internally driven processes, such as freely recalling an event or appropriately employing a task strategy (Craik, 1986; McDowd and Shaw, 2000; Morrow and Rogers, 2008). Because of these difficulties, older adults rely more on contextual or environmental information to aid their performance. Thus, providing useful information in the environment of a task can aid performance. However, because there is inevitably distracting stimuli in the environment in addition to the supportive information, older adults may have difficulty ignoring the irrelevant stimuli.

If memory requirements in a task are offset by the existence of readily accessible information in the environment, attentional and memory processes can be allocated elsewhere, to more demanding aspects of a task. For example, in a typical software application, instead of using function buttons with icons only, at least provide the option for turning on text labels, which eliminates the need for novice users to experiment with and memorize the functions of the buttons.

Environmental support can come in various forms:

1. Provide some characteristic of the stimulus to be recalled to aid, or cue, recollection.

2. Provide an outline or map of the material. For Web browsing tasks, navigation aids should be provided if desired by the user. These can provide a visual history of where the user has been, reducing the reliance on memory for the structure of the website.

3. Physical aids constitute a form of environmental support. They serve to remind the user of the previous encoding instance. Specific aids, or cognitive prosthetics, can be used to assist older adults' memory (Morrow, 2003)—for example,

software designed to store grocery lists and to recommend items that have been purchased commonly in the past.

4. Structuring text appropriately can benefit recall of the text.

5. Search and detection of target stimuli benefit from consistent arrays of irrelevant stimuli (Chun and Jiang, 1998; Jiang and Chun, 2001). That is, attention can be guided to targets by the knowledge gathered from the consistent arrays of distractor stimuli.

### 3.2.3  Prospective Memory

Prospective memory involves remembering to do something in the future and is essential in planning and completing general daily activities (e.g., fulfilling appointments, performing household chores). It can also be critically important for safety, such as remembering to take medications or turn off the oven. There are two categories of prospective memory—time based and event-based—which differ in the degree to which the rememberer must rely on self-initiated cues (Einstein et al., 1995; Park et al., 1997). In event-based prospective memory, the person is cued about the to-be-remembered information by some external event or stimulus (e.g., remembering to take a medication when a timer goes off); whereas in time-based prospective memory, the person must remember after some amount of time has passed (e.g., remembering to take a medication at two o'clock in the afternoon). As is typically found in cognitive aging, when self-initiated processing is required, older adults' performance suffers (Craik, 1986), and age-related differences in prospective memory are greater for time-based situations (Einstein et al., 1995).

Although older adults' prospective memory is impaired relative to younger adults (Park et al., 1997; Zacks et al., 2000), the memory phenomenon under heavy task demands may provide a better representation of prospective memory in real tasks. Under these demanding conditions, age-related differences are exacerbated. Einstein and colleagues (1997) investigated prospective memory in the context of additional, distracting activities (essentially, a divided-attention manipulation). They found that older adults had a more difficult time remembering to do the to-be-remembered action than younger adults when attentional demands were high.

***Design Implications and Suggestions***  Older adults' difficulty with prospective memory has important implications for the design of memory aids and reminders. Physical reminders such as cognitive prosthetics are critical in reducing the functional effects of the prospective memory age-related difference. Whenever prospective memory is required (such as in remembering to take one's medications at various points throughout the day), time-based prospective memory tasks should be turned into event-based memory tasks. Alerts can be built into cell phones or personal digital assistants (PDAs) or other small, unobtrusive devices, providing users with a memory aid. Essentially, these

function as environmental supports, reducing older adults' need to rely on self-initiated memory processes.

### 3.2.4 Source Errors

As with all human memory, recall for the context in which a memory was first created is not necessarily reliable (Johnson et al., 1993). For example, one might recall that a car owner's manual provided information about an automotive repair but not in which section of the manual it occurred in (i.e., external source monitoring). Several studies have shown that older adults are poorer than younger adults at source monitoring (Cohen and Faulkner, 1989; Ferguson et al., 1992; Johnson et al., 1995). For example, when the source of information was distinct (i.e., the gender and appearance of two speakers), older adults' source monitoring was on par with younger adults (Ferguson et al., 1992; Johnson et al., 1995). However, older adults were less able to utilize multiple cues to aid their memory for source (Ferguson et al., 1992), and when additional cognitive processing was performed between source and test, older adults had difficulty retaining the link between the distinctive perceptual source information and other aspects of the source, such as what the source said (Johnson et al., 1995).

***Design Implications and Suggestions*** Older adults' memory for source can benefit from perceptual disambiguation. In the repair manual example above, the chapters may be made more distinct by using different-colored paper in each chapter. The memory for a given chapter includes the color of the paper, and this memory for color may be cued when the user flips past that color in the manual. Thus, the user does not need to actively retrieve the section of the manual but, instead, can easily recognize a particularly salient feature.

Because memory for source is impaired in older adults, particularly under intervening cognitively demanding conditions, memory aids should be provided to reduce the need for older adults to rely on self-initiated retrieval processes. For example, an older adult may not recall whether medical advice was given to them by their doctor or by a friend. This problem could be reduced if all medical advice given by their physician was also provided to them via a text file or printed transcript. This added information would serve as a redundant cue to the information source and provide the information in the world, rather than requiring the person to rely on memory for the source.

### 3.2.5 Semantic Memory

Despite the general negative view of aging and memory presented thus far, there are several characteristics of memory and knowledge that remain relatively robust across the lifespan, most notably, semantic memory (memory for facts or general knowledge, sometimes referred to as crystallized knowledge; Cattell, 1963). Designers should take advantage of older adults' relatively preserved semantic memory.

Semantic memory has been used to improve older adults' memory for events in the future, specifically in the domain of medication instructions and health appointments. For example, using older adults' schemata (or crystallized knowledge structures about some domain) about a given memory task can be used to construct aids to help them in the memory task (Morrow et al., 1998, 2000). Younger and older adults share a schema for how medication reminders should be worded and arranged—specifically, they preferred shorter messages, incorporating, in order, time to take a medication, required dosage, duration one should take the medication, health warnings, and side effects (Morrow et al., 2000). When this schema knowledge was incorporated in an automated phone message application to present medication information in a way that followed their schema of presentation or violated the schema, both younger and older adults benefited from the schema-consistent version, such that they recalled the relevant information more accurately.

***Design Implications and Suggestions*** Older adults' knowledge of existing systems and devices can be used as a tool to design systems that can easily be used and understood by older adults. Knowledge engineering, a technique that facilitates the understanding of how tasks are performed by gathering the knowledge used within a specific process, is a critical phase in the design process, particularly with older adults, given their differences in knowledge and experience. Furthermore, older adults' crystallized knowledge may be extended to novel domains, where their knowledge can be transferred to similar novel applications and technology (this idea is discussed in Section 5.2).

## 3.3 Language Comprehension

Language comprehension is heavily reliant on semantic memory (i.e., one's crystallized knowledge base) as well as on working memory capacity. Language comprehension remains a critical function of people's lives as they age. With advancing age, people must be able to efficiently read labels of new medications, the instruction manuals of novel devices such as wheelchairs and health-related devices, and the warning materials that accompany these and other products and systems.

Often, older adults perform well in comprehending spoken and written language (Wingfield and Stine-Morrow, 2000). For example, they are able to comprehend figurative language as well as younger adults (Szuchman and Erber, 1990), and they are able to create an appropriate mental representation of text (Radvansky et al., 1990). However, there are several factors that can negatively influence older adults' comprehension of spoken and written language, more so than for younger adults. In general, these are related to demands on working memory (see Wingfield and Stine-Morrow, 2000, for a review).

### 3.3.1 Sentence Structure

Comprehension can be improved by reducing processing demands on older readers. For example, working memory can be taxed if many words and clauses bisect the subject and verb in a sentence (Norman et al., 1992; Wingfield and Stine-Morrow, 2000). Left-branching sentences contain a particularly difficult clause, as it comes between the subject and verb in the main clause

and requires the maintenance of the subject while simultaneously processing the embedded clause. Left-branching sentences are particularly detrimental to older adults' comprehension (Norman et al., 1992).

***Design Implications and Suggestions***   When designing instructions, warnings, and other text-based materials, the limitations of older adults' working memory should be considered. Comprehension will be improved if subjects and predicates are near each other, minimizing the need to keep the subject of the sentence in working memory for long periods. Following the guidelines put forth by Norman et al. (1992) and Kemper (1987) can greatly improve the readability of text, primarily by reducing the demand on working memory (see Table 4). For clearest writing, subject and predicate should be within close proximity.

### 3.3.2   Inferencing and Figurative Language

Research suggests that older adults are at a disadvantage relative to younger adults in making appropriate inferences (Hamm and Hasher, 1992). Furthermore, this age-related difficulty may be exacerbated in instances where older adults cannot rely on their crystallized knowledge (Arenberg and Robertson-Tchabo, 1985; Hancock et al., 2005). Older adults interpret figurative language well (e.g., metaphors) (Szuchman and Erber, 1990). The use of figurative language taps the rich structure of knowledge that older adults have built across their lifetime and allows them to constrain their inferences with that knowledge.

***Design Implications and Suggestions***   Often, it is necessary to make inferences beyond what is present in the text. For example, if a user manual for a lawn mower tells the user to "disable the starting mechanism before replacing the blade," an important inference one might make would be that one should also disable the starting mechanism before removing debris from the undercarriage of the mower. It is important to minimize the need for inferencing beyond the text. This can be especially critical in the construction of warning materials (Hancock et al., 2005). Older adults' ability to interpret figurative language combined with their high degree of semantic knowledge suggests that considerable information may be communicated through brief figurative text. That is, the figurative text cues older adults' extensive and rich semantic network of knowledge, potentially communicating considerably more information than is solely within the figurative text. This can be especially useful in the design of space-limited text, such as warning labels on pill bottles or cleaners, where space is severely restricted.

## 3.4   Executive Control

The term "executive control" encompasses a number of cognitive abilities related to the maintenance and updating of cognitive and behavioral goals, the planning and sequencing of actions, problem solving, and the inhibition of automatic responses. These abilities tend to demonstrate substantial decline with advancing age, and the brain regions that subserve these functions demonstrate the most dramatic age-related atrophy (Resnick et al., 2003). Additionally, performance on measures of fluid intelligence suffers with increasing age (Bugg et al., 2006). Declining executive control has been demonstrated to take a substantial toll on older adults' ability to function independently (Royall et al., 2004). Thus it is reasonable to assume these changes will influence how older adults interact with products and systems.

***Design Implications and Suggestions***   Age-related declines in executive control suggest older adults may perform especially poorly in multitasking environments and on complex tasks involving the coordination of multiple subtasks. However, although task coordination and multitasking demonstrate age-related impairment, evidence suggests that training can mitigate age effects (Kramer et al., 1999). Variable priority training (VPT), in which learners practice the whole task while at different times emphasizing performance of different task subcomponents, has proven especially effective (Gopher, 2007). A number of studies have found that VPT improves the performance of older adults on complex tasks in which age-related differences in executive control are evident (Bherer et al., 2005; Kramer et al., 1995). As a general rule, the designer should also provide environmental support in the form of salient sensory cues to indicate when a particular task or subtask requires attention to minimize reliance on executive control (Wickens and Seidler, 1997). In dual-task or multitasking environments, consistent mappings between tasks and their responses should be maintained to minimize cognitive effort of switching from one task to another.

**Table 4   Examples of Cognitively Demanding Prose**

| Example | Problem | Revision |
|---|---|---|
| To change the level of the fluid in the round canister above the rotary encoder and the pressure dial, rotate it clockwise. | The reader must identify the referent for the pronoun "it." There are several intervening phrases between the pronoun and referent, causing a load on the reader's working memory. | To change the level of the fluid in the round canister, rotate the canister clockwise. The canister is located above the rotary encoder and the pressure dial. |
| The screw that is used to secure the tray above the cabinets so that the compartment is accessible is located in the yellow bag. | The subject ("screw") and predicate ("is located") are separated by a long clause, causing a load on the reader's working memory. | Locate the screw in the yellow bag. This screw is used to secure the tray above the cabinets so that the compartment is accessible. |

## 4 MOVEMENT AND BIOMECHANICS

### 4.1 Movement Speed

Movement speed is the speed with which a person can make a movement after the requisite cognitive processes to start the movement have occurred (Spirduso, 1995). In general, older adults are slower in their movements than younger adults (Stelmach and Nahom, 1992). Figure 5 shows the typical finding of greater overall time to make a reaching movement for older adults relative to younger adults and the slower peak velocity.

***Design Implications and Suggestions*** Reductions in movement speed are relevant to a wide range of activities and scenarios performed by older adults. Traffic light timers should be set to provide adequate walking time for older pedestrians. The required double-click speed for mouse buttons on public computers should be amenable to older adult users. Other examples include the speed of revolving doors, self-operated credit card terminals, and the time required for text entry in cell phones. Any task that requires relatively rapid movement is potentially difficult for many older users.

### 4.2 Movement Control

In effect, older adults will be slower in tasks that involve grasping (Carnahan et al., 1993), reaching (Seidler-Dobrin and Stelmach, 1998), and continuous movement (Wishart et al., 2000). Their movements involve more submovements and shorter initial primary submovements than those of younger adults (Walker et al., 1997; Seidler-Dobrin and Stelmach, 1998; Smith et al., 1999). This essentially results in slower, more variable movements in older adults. Furthermore, as a movement task becomes more difficult, older adults slow at a greater rate than do younger adults (Ketcham and Stelmach, 2001).

Age-related differences in computer mouse performance have been reported. For example, in four common mouse tasks—pointing, clicking, double clicking, and dragging—older adults were slower in their movements, produced more submovements, and made more

errors, particularly for double-clicking tasks (e.g., moving out of the icon range before it was double clicked) (Smith et al., 1999; see also Walker et al., 1997). Also, older adults are less able to coordinate multiple movements with multiple body parts as well as younger adults, such as bimanual tasks (e.g., twisting off a lid) (Stelmach et al., 1988).

***Design Implications and Suggestions*** The variability in movement control and speed contributes to older adults' greater variability in their use of input devices, both within and between people (Rogers et al., 2005). It is possible that with increased perceptual feedback older adults can be trained to lengthen their initial movements, thereby also reducing the required number of submovements. For example, mice could be operated in conjunction with software that provides tactile feedback as the cursor nears an icon or target. "Sticky" icons have been employed that attract the cursor, within some user-defined sensitivity range. Users should be provided with the option to increase icon size, effectively improving the physical stimulus.

Older adults have been shown to have difficulty decelerating as they approach a target and trouble staying at the target once reached (Smith et al., 1999). This is consistent with the common finding that older adults employ more submovements as they home in on a target (Walker et al., 1997). Discrete hand movements are often required in computer tasks, whether with a mouse, touch screen, light pen, or other input device.

In software design, icons should be a reasonable size, as older adults have difficulty navigating to small targets (Walker et al., 1996). Furthermore, older adults are slowed in their navigation more than younger adults by targets embedded with other stimuli (Rogers et al., 2005). Thus, in addition to making icons and target stimuli easier to select, creating appropriately sized icons can help to separate those relevant targets perceptually from irrelevant or background stimuli.

Older adults' difficulty with coupled movements and other movement coordination tasks can cause
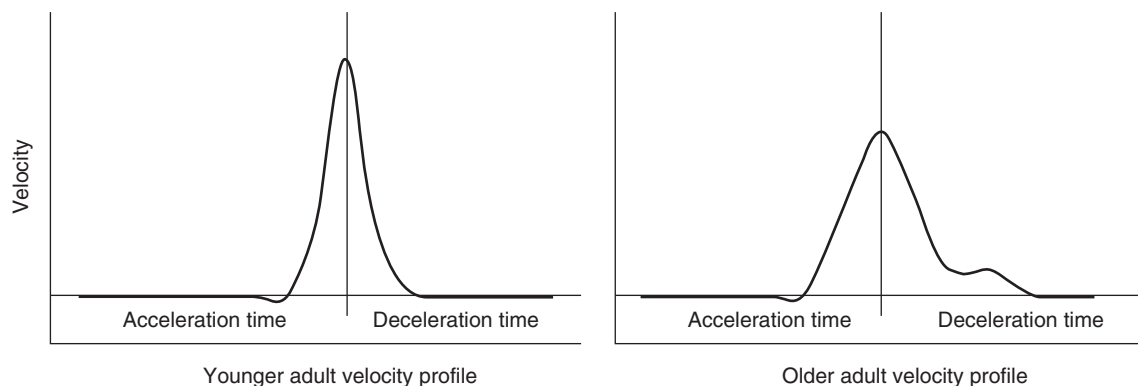


**Figure 5** Examples of a younger adult's and an older adult's velocity profiles in a reaching task. Note the longer time to make the movement and greater variability during the deceleration stage in the older adult profile. (Adapted from Ketcham et al., 2002.)

problems for bimanual tasks, such as opening a twist-top pill bottle or reaching and grasping coordinated movements. Thus, products that require a coordination of multiple movements should be redesigned if possible. For example, pill bottles designed for older adults should employ a tabbed top.

### 4.3 Balance

Falls due to loss of balance are a serious problem for older adults (Horak et al., 1989; Agnew and Suruda, 1993). Postural sway increases with age (Borger et al., 1999; Kristinsdottir et al., 2001), leading to loss of balance and falls. Postural sway is affected by several variables, including poorer vision, which reduces older adults' ability to detect movement cues in the environment that would indicate sway; reduced sensitivity to vibrations in the lower limbs, which reduces the contraction signals sent to muscles (Kristinsdottir et al., 2001); and cognitive demands (Lajoie et al., 1993; Brown et al., 1999; Melzer et al., 2001). It appears that older adults rely more on dynamic visual cues to aid in their balance. When visual cues indicated that they were moving (although they were in fact stationary), older adults were much more likely to exhibit postural sway (Borger et al. 1999). Older adults also tend to have greater postural sway (relative to younger adults) when in a moving visual environment, and when flooring was more compliant (e.g., carpet vs. hardwood flooring), age-related differences in postural sway were greater (Redfern et al., 1997).

Balance is affected by cognitive demands. When participants were asked to perform simple subtraction while stationary and while recovering from a small movement of the platform on which they were standing, older and younger adults performed similarly on the subtraction task prior to perturbations of their platform. However, older adults were repeatedly and significantly slower after a perturbation, suggesting that the cognitive demands of maintaining balance interfered with their ability to perform the subtraction task (Brown et al., 1999).

***Design Implications and Suggestions*** Because older adults may be more prone to losing their balance while getting on and off moving sidewalks, warnings should be provided with a long lead time prior to the need to correct posture. Apparent motion and true motion in the environment can also increase the likelihood that older adults will lose their balance due to their reliance on perceptual cues. Examples include wall-sized animated advertisements and moving subway cars. Environmental support may be provided to reduce this problem. For example, in designing platforms in front of moving subway cars, a row of lights can be provided above the railway cars, resulting in a stationary stimulus to counter the perceptual cues of the train. Cues may be beneficial in other contexts as well. On walls next to stairs and pedestrian ramps, arrows can be displayed indicating the direction of the change of slope in the walkway. For any scenario where balance may be an issue, handrails are highly recommended.

### 4.4 Locomotion

The slower walking speed of older adults is due to several factors: preference/strategy, decreases in strength,

joint motion, and endurance. The decrease in locomotion speed is the primary factor in gait changes as well. Gait changes include shorter steps and an increase in the time that both feet are on the ground, increasing support and balance (Spirduso, 1995; Lockhart, 1997). However, when cognitive demands are increased, walking becomes more demanding, presenting functional problems, particularly for older adults. For example, in one study, age-related differences in a memory task were greater when both age groups were walking than when they were stationary (Li et al., 2001). The researchers interpreted the results in terms of older adults selecting the more important walking task over the memory task (presumably staying balanced was considered more important than memorizing words). Cognitive demands can also result in movement decrement as well, as older adults walked at a slower speed when performing a cognitive secondary task (Lajoie et al., 1996).

***Design Implications and Suggestions*** Older adults walk more slowly than younger adults, and dual-task data suggest that older adults may change their movement speed and their cognitive performance. Slower locomotion for older adults has implications for pedestrian crossings, vehicles with automatic doors (such as elevators and subway trains), and other situations in which older adults are expected to keep pace with younger persons. Attendants should be trained to be prepared for these locomotion limitations. Particularly in instances where cognitive demand is high, such as crossing busy streets, walking in the downtown district of a large city, or any novel walking environment, the walking speed and balance of older adults may be significantly compromised. Although several studies have shown that the cognitive task is compromised as opposed to the walking task, this is probably due to the relative insignificance of the cognitive task. When searching for landmarks in an unfamiliar business district, the cognitive task may be more important, and attention to one's walking may be reduced.

### 4.5 Strength

Muscular strength is maintained through much of adulthood, beginning to fall off around age 60 (Ketcham and Stelmach, 2001). Muscle strength, including hand-grip strength and endurance, decreases with increasing age (e.g., Kallman et al., 1990; Metter et al., 1997). Skin slipperiness also increases with age (Cole, 1991). These changes, along with the onset of disease processes such as arthritis, may reduce older adults' strength in general and for fine motor control tasks. However, strength drops off as a result of muscle mass loss, and appropriate exercise and training regimens can abate strength and muscle losses, at least to some degree.

***Design Implications and Suggestions*** As with changes in speed, changes in strength have widespread effects on the functional limitations of older adults. For those systems and products that may have older adults in the user population, designers must consider the reductions in strength for all tasks that require actions such as pushing, pulling, lifting, twisting, and pressing.

Some pill bottles are now designed to be secured such that minimal force is required to open them, and other products that require actions such as these must be designed with older adults' reductions in strength in mind. If force requirements for a product cannot be reduced, assistive aids should be provided.

### 4.6  Force Control

Controlling one's force correctly can be critical in maintaining one's safety (e.g., holding onto a grab bar in the shower), avoiding embarrassment (e.g., accidentally crushing a beverage-filled Styrofoam cup at a party), or simply manipulating an everyday device (e.g., rotating the jog dial on a PDA). We focus primarily on control of force involving the hand and fingers. Older adults employ a grip force up to twice that of younger adults, and their force beyond that required is also twice that of younger counterparts (Cole, 1991). This indicates both a perceptual component (older adults, particularly over 60 years, do not accurately perceive the friction of the object being gripped) and a strategic component (perhaps aware of their inclination to misgrip objects or misperceive friction characteristics, they overgrip intentionally) (Cole et al., 1999).

***Design Implications and Suggestions***   The ability to control the force employed in a task is decreased significantly as one reaches old age. This has design implications for the structural soundness of products and the design of controls. For example, some computer mice have scroll wheels that can function as a button when clicked. If this button is too sensitive, older adults may have considerably more trouble than younger adults in scrolling without activating the function intended by the button. In designing handles for gripping, extra texture should be provided to offset the perceptual loss experienced by older adults. This can help to reduce the overgripping behavior, which may cause early fatigue when using the product.

## 5   BELIEFS, ATTITUDES, AND MOTIVATION

*Ageism* biases include the belief that older adults are reluctant to interact with technology. Focus group research shows that older adults are actually motivated to use products when they are informed about the benefits (Melenhorst et al., 2001). In fact, reduced usage rates among older adults may be the result of a poor understanding of the benefits of the product, reduced income, and difficulty using certain products (Fisk et al., 2009). For example, many older adults appear interested and motivated to use the Web. A Web usage questionnaire of middle-aged (40–59 years), young–old (60–74 years), and old–old (75–92 years) showed usage proportions of 56, 25, and 10%, respectively (Morrell et al., 2000). When asked about their desires for Web activities, regardless of whether they currently used the Web, older adults' top three preferred activities were to use email, to obtain information on travel or pleasure, and to obtain health-related information. Of the Web users, 66% reported being online for one year or more and were using the Web for an hour or more

several times per week. Of nonusers, middle-aged users expressed the most interest in using the Web, whereas the old–old users expressed the least interest. Relative to other age groups the percentage of older adults who use computers is lower, but their usage rates continue to grow (Kinsella and Velkoff, 2001; Olson et al., 2011).

Experience with new technologies may increase older adults' willingness to use them. For example, after completing an experiment with an automated teller machine (ATM) simulator, the number of older adults who expressed interest in interacting with an ATM increased from 28 to 60% (Rogers et al., 1996). The relatively short experiment provided sufficient exposure to and knowledge about the system to motivate participants to use the technology.

Knowledge and anxiety about computers appear to directly influence interest in computers, mediating the role of age in computer interest as evidenced by significant correlations between age and computer knowledge and computer anxiety in a structural equation model (Ellis and Allaire, 1999). The $r^2$-value for computer interest was 0.49, indicating that nearly half of the variance in interest was captured by the model. The effect of age on computer interest was mediated primarily by computer knowledge and anxiety, although some age-related variance was directly related to computer interest (indicating other possible predictors; see also Czaja et al., 2006).

***Design Implications and Suggestions***   Computer knowledge is highly related to computer anxiety (Czaja et al., 2006; Ellis and Allaire, 1999), suggesting the importance of educating older adults about computers. Understanding the potential benefits of computer technology and reducing anxiety associated with computers may increase older adults' willingness to use computers and, perhaps, technology in general (e.g., Czaja and Sharit, 1998). It is important to understand older adults' attitudes for any system or product they may encounter. Once older adults are made knowledgeable about the system or product (including the knowledge of how to interact with it successfully, the knowledge that they are capable of interacting with it successfully, and the knowledge about how it can benefit them), their interest and motivation to use it will increase, which in turn will result in greater attention to the system or product and lead to more efficient learning and utilization of the technology.

## 6   CASE STUDY: A HAND-HELD GAMING SYSTEM FOR OLDER ADULTS

We next present a hypothetical case study to illustrate the design implications derived from the age-related changes reviewed in this chapter. The example we have chosen to discuss is a hand-held gaming device designed with the purpose of delivering entertainment or cognitive enrichment activities to older adults. We have chosen this system for a number of reasons. First, although far fewer older adults engage in video game play compared to younger adults, gamers who are 65 years of age or older are among the most active gamers, with almost a

third playing nearly every day (Lenhart et al., 2008). The fact that relatively few older adults choose to engage in video game play may be partly due to barriers in the form of poor design. Overcoming these barriers may open up a variety enjoyable activities to the older adult population that have the potential to impact well-being and cognition. Recent evidence suggests that cognitive engagement via commercial video game play can improve a number of perceptual and cognitive abilities, including ameliorating age-related declines in memory, reasoning ability, and multitasking performance (Basak et al., 2008; Green and Bavelier, 2008). Additionally, a number of companies have begun to market software and hardware (including hand-held devices) to older adults with implicit or explicit claims of being able to improve everyday cognitive functioning. The evidence in support of these claims is mixed (Hertzog et al., 2009). However, it is almost certain that the design quality of these technologies will ultimately influence adoption and continued use should they prove effective. For the purposes of this case study, we present a hypothetical hand-held system running a simple operating system that allows users to select from several different video games and cognitive training activities.

## 6.1 Description of the System

The system itself is similar to several commercially available hand-held gaming systems. The hardware component is approximately $5 \times 2.5 \times 0.8$ in., with a 3.5-in. (diagonal) touch-sensitive color liquid crystal display (LCD) screen (Figure 6). A stylus is included which is used to interface with the touch screen. A directional pad is located to the left of the screen and four buttons are located to the right. These serve as game input devices and also allow users to navigate menus without using the stylus. The unit also includes a built-in microphone to accept speech input and speakers to deliver music and auditory feedback. Like many hand-held devices, the system also provides haptic feedback in the form of system vibration. We will assume that this product is novel to older adult users; thus, they must learn how to interact with it efficiently through usable design and effective training.

## 6.2 Visual Information

In this hypothetical system, a small color LCD touch screen is the primary interface presented to the older adult user. Like most technology, a paper operations manual is also provided for the system and games. An icon-based operating system and menu structure allow older adult users to select games and activities to play and to

customize system options. Once a game or activity has been selected, players are presented with the opportunity to navigate menus to change game options. Information such as instructions, tutorials, and performance feedback are also presented on screen, mostly in text and icon form. During game play and tutorials, auditory and haptic feedback accompany certain events.

### 6.2.1 Text

The type of text that is most commonly used in hand-held game systems and games is instructional text, informational text, labels, and performance feedback. That is, verbose and complex writing is typically unnecessary, and clarity and precision should be the primary objectives. In creating instructional text and tutorials for a hand-held game system and accompanying printed materials, organize the relevant information following standard information display guidelines. Avoid creating an interface with multiple frames, segmentation lines, text boxes, headings, icons, and links. The goal is to present minimal distraction and to present relevant information as clearly as possible.

- Group related elements (such as icons for different types of activities).
- Align elements in a list, generally to the left.
- Utilize common symbols to convey meaning efficiently (but avoid novel symbols or highly similar symbols with different meanings).
- Base text on a grid layout. Following this layout throughout an interface creates a common look and feel to the interface.

***Perceptual Considerations*** When presenting textual information on a hand-held system or in print, the text should be presented in high contrast and in an easily readable font (e.g., sans serif fonts; Morrell and Echt, 1997). Text should be divided into sections, with perceptually salient headers or labels. Additionally, strategic use of white space will help to separate sections, reducing the need for visual search. Lines of text should not run across the length of a screen, and horizontal scrolling should be avoided. Text should be centered on the screen if possible. In printed text accompanying the system and games, lines should run for 6–8 in., reducing the need for long visual scanning (Ellis and Kurniawan, 2000). If done consistently, this organization provides support for older adults in scanning to find headings and in reading the text. Text should be presented in 12-point type, sans serif font, and the highest possible contrast should be used (see Figure 7). While color can and should be used to enhance the game experience, important instructional text should be limited to black on white.

***Cognitive Considerations*** Text should not contain left-branching sentences or sentences with many clauses in them, which overload working memory. For adults of all ages, technical text should be written at a sixth-grade level (McLaughlin, 1969). Text should be organized via a small number of organizational principles. For example, in an instructional manual, chapters, subsections, and headers should use consistent conventions (e.g., subsection headings can be presented in a unique font size or in bold type).



**Figure 6** A schematic drawing of the hypothetical hand-held game system and stylus.

| Low-contrast text is difficult to read. |
| :--- |
| *Serif and script fonts are difficult to read.* |
| Eight-point font is too small. |
| This test has appropriate font, size (12 pt), and contrast. |

**Figure 7** Examples of poor and appropriate text presentation. (Adapted from Fisk et al., 2009.)

### 6.2.2 Icons

Icons and symbols are often used in instructional manuals and software applications to convey meaning in minimum physical space. This can lead to small icons that are difficult for older adults to see and confusing icons that are difficult to interpret. In the hypothetical system, icons may be used to depict one of several games or activities available to play and they may be used to denote helpful hints and warnings in the instruction manuals.

***Perceptual Considerations*** Icon size is an obvious consideration when designing for older users, especially on hand-held systems with small screens. Distance of the system from the eyes should be considered as well. Besides icons that are too small, a graphically complex icon may be difficult for lower acuity persons to perceive correctly. Also, if icons and symbols are color coded, care should be taken to avoid multiple colors from the high-frequency end of the color spectrum (e.g., blues, greens) and to use highly distinctive hues to avoid reducing the contrast of the icon.

***Cognitive Considerations*** By serving as graphical representations of concepts and instructions, symbols and icons can be very effective in communicating a large amount of information in a small space. However, evidence suggests that older adults may have more trouble than younger adults in understanding symbols and icons, and usability testing should be conducted with older adults to ensure that all icons are interpretable (Hancock et al., 2004). Older adults may not be familiar with icons familiar to younger gamers and computer users (e.g., a disk icon representing the option to save data or progress). When used, icons should always be accompanied by a textual label and description, at least initially on the screen and always in the manual. Structurally complex icons can also be problematic, not only in terms of comprehension, but also for older adults, in terms of perception.

### 6.2.3 Information Visualization

Like many video games and training activities, feedback regarding performance is essential. Games often allow for multiple related performance variables to be presented at a single time. Data visualization can refer to a simple representation of data in a pictorial fashion, such as a pie chart. In this context, data visualization would require a multivariate display for representing the various relevant characteristics of the data in a multidimensional space. The limiting factor in data visualization can often be traced to the person viewing the representation. An understanding of human cognition and visualization constraints is essential to the development and selection of effective display characteristics. However, there has been little specific focus on age-related differences in information visualization capabilities, per se, but there is evidence that older adults have some difficulties with graph perception (Fausset, 2008). The best design approach here would be iterative user testing of the software with older adults (Fisk et al., 2009).

### 6.3 Auditory Information

Auditory information can be used in a number of ways in our hypothetical game system. For example, game tutorials might have verbal instructions to walk players through how to perform important game actions. Although this may be fine for young adults, previously discussed age-related changes in audition and speech perception should be taken into account when designing for older adults. This is especially true considering the poor sound quality often associated with the small built-in speakers common to hand-held devices and speech compression used to conserve system memory. A common game design practice is to accompany game play with background music. This may be problematic for older adults who may experience more masking of important auditory cues. Ideally, auditory information should be presented with a high signal-to-noise ratio (i.e., minimal background sounds to mask the intended message). Moreover, whenever possible, important system and in-game events should be accompanied by redundant visual or haptic cues.

### 6.4 Haptic Information

Many hand-held devices, including game systems, support haptic feedback in the form of system vibration. For example, in a racing game, a player might be alerted that he or she is off course through the use of haptic feedback. Vibration may also be used as feedback to let the user know an icon has been successfully selected. When using haptic feedback, vibration frequency should be carefully selected as high-frequency vibration is selectively impaired with age. As recommended for auditory cues, haptic cues should not be used alone and should be accompanied by redundant cues from other modalities.

### 6.5 Input Devices

### 6.5.1 Perceptual Considerations

Given the necessarily small size of buttons on hand-held devices, labeling these can be difficult, particularly when considering the reduced acuity of older users. High-contrast symbols are generally necessary as opposed to text, and these should be clearly defined in instructional materials and training programs. Tactile feedback should not be used as the primary feedback for input controls. Older adults are less sensitive than younger adults to tactile feedback (e.g., Thornberry and Mistretta, 1981).

Thus, when a button is pressed, visual or auditory feedback should be the primary means of feedback.

### 6.5.2 Motor Considerations

Hand-held devices necessarily have small controls—small jog dials, small buttons, and small input devices, such as styluses for touch screens. Reductions in force control and movement control are critical considerations in designing these devices.

*Jog Dials* Jog dials, typically used for rotating through options in small hand-held devices, should have sufficiently stiff, discrete stopping points when rotated, given older adults' force control deficiencies. The "teeth" on the dial should provide enough friction to account for older adults' more slippery skin (Cole, 1991). Rotary-type controllers are well suited as input devices for older adult users for certain controls such as sliders and scrolling (Rogers et al., 2005).

*Buttons* Buttons should be far enough apart to minimize accidental activations. Buttons should be firm, so that they are not depressed accidentally when older users rest their finger on them, which may be likely given reductions in force control. Tactile feedback of a button press should be provided as redundant feedback to visual and/or auditory feedback to inform the user that the control has been activated.

*Touch Screens* Touch screens are not optimal for older adults for small targets, due to high variability in movement control. For example, if selecting one of four $\frac{1}{4}$-in.-wide options on a hand-held system, a jog dial should be used to cycle through the options and make a selection, as opposed to requiring the older user to select the option with a touch pen or with a finger. If touch screens are used in small devices, selection areas should be maximized to increase accuracy. In general, touch screens should be used for ballistic movements (particularly when screen real estate is large), but for precise control, indirect pointing devices such as a rotary device or mouse should be used (Charness et al., 2004; Rogers et al., 2005). Older adults have difficulty making accurate, fine motor movements as well as making fast movements. The hypothetical system addresses this by providing navigation alternatives (use of the directional pad and buttons) that are more robust to tremor and motor control issues. If designing a non-hand-held system that includes a mouse, fine mouse movements should be minimized if possible, and the option to control the double-click speed and gain on the mouse should be available.

*Hardware Casing* Age-related changes in muscle strength (e.g., hand-grip strength and endurance), increased slipperiness, and higher likelihood of arthritis together have implications for designing the hardware casing of hand-held devices for older adults, especially gaming systems which may be used for extended periods of time. Nonslip casing materials are preferable and system weight should be minimized. User testing should be conducted to ensure usability and long-term comfort of the system.

### 6.5.3 Voice Command

Many hand-held devices, including game systems, feature voice command options. While this can serve as an alternative in some cases to overcome difficulties in motor control, it is not generally recommended as a primary input method. Even the best speech recognition software may frustrate users if recognition is imperfect, and there is reason to suspect older adults may experience greater errors compared to their younger counterparts. Speech recognition algorithms designed and evaluated using young adults will not be as accurate for older adults as a result of age-related declines in speech quality.

### 6.6 Training and Instructional Materials

Training can cover a multitude of design errors (although it should not be used as a crutch by designers). For example, people are entirely capable of learning the meaning of an obscure icon, provided that it is associated consistently with the same function and is not easily confused with other icons. Older adults may have unique training requirements. For example, older adults may be less familiar with computer and gaming technologies and as a result may require training for basic features of a system. For example, for PC-based gaming systems, mouse training, instruction on windowing, or search tool training may be required before other, higher level aspects of the system are trained. Older adults may also be less confident in their ability to interact successfully with novel technological devices such as PDAs and other hand-held devices (due in part to less experience).

Inexperienced older users may have an incorrect mental model of the system or no model at all, which is particularly likely for gaming devices. Often these models are constructed from repeated interaction with the system; hence, novice users should first be trained to form an appropriate model of the system. Older adults may be able to adopt new strategies that are optimal for the task (Rogers and Gilbert, 1997) and develop new mental models successfully (Gilbert and Rogers, 1999) although there are individual differences between older adults in this ability (Olson et al., 2009).

Training programs and instructional materials should be developed for the use of the hand-held game system and each game. Game and system tutorials should be clearly marked and easily accessible at any time during and after training. In addition, age-specific training might also have to be provided to enable older adults to capitalize on the functionality of the software (Hickman et al., 2007). In designing training programs and materials for the current hypothetical system, several considerations should be made. Many of these are relevant to training adults of all ages.

### 6.6.1 Duration of Training

Based on training research with simple and complex stimulus situations, older adults will require about one and a half to two times the amount of training required by younger adults (Fisk et al., 2004). Extended training (or *overtraining*) can help solidify the process being trained and improve retention (Jones, 1989). Older adults tend to benefit more from overtraining than

younger adults, who presumably reach a more stable area of the learning curve earlier (e.g., Sharit et al., 2003). Short breaks should be provided when training sessions extend for 30 min or more.

### 6.6.2 Format of Training

Video games are complex pieces of software. Appropriately constructed part-task training can be helpful in training complex tasks. Part-task training involves practice on some subset of a task before practice on the whole task (Kramer et al., 1995). Selecting the appropriate subset is, of course, critical. A thorough task analysis is required before a task is divided into parts, particularly in complex tasks. Artificial divisions of the task (especially in the simplest tasks) will be detrimental. For example, imagine a generic action game in which players must race their vehicle around a track and also use weapon systems against competing players. The control aspect of maneuvering around the track could first be trained independently of utilizing weapon systems. When proficiency is reached, the two can be combined and additional training provided. In the present system, a stylus and input device training application should be an optional training program for novice users, to be completed prior to using the system for games. Although part-task training is often beneficial, different aspects of a task will need to be closely integrated, such as related motor movements with different hands, or dividing the task will be detrimental to older adults' performance (Korteling, 1993). In addition to use of a game to provide training and instruction, training should also include clear, comprehensive written manuals for how to use the system and games. Older adults report a preference for written instructional materials when learning how to use new technology (Mitzner et al., 2008).

### 6.6.3 Flexibility of Training

A flexible training program is important for older users, as the variability in experience, skills, and knowledge can be high in an older population. The training application for hand-held gaming systems should assess older adults' system and gaming knowledge and be flexible enough to provide low-level training to novice users but to skip these aspects of the training program for users who demonstrate proficiency.

### 6.6.4 Active Learning

Active training is essential; passive observation leads to little, if any, learning (Schneider, 1985). Training programs and game tutorials should make use of the same interface on which users will be performing the actual tasks or as close as possible. Tutorials for using the gaming software can easily be presented using the same graphical user interface as the actual game. This allows the user to interact actively with the interface, as opposed to reading instructional text or viewing pictures of the interface.

### 6.6.5 Feedback in Training

Feedback should be provided for every interaction with the system (e.g., a button press should result in a corresponding change on the screen as well as an auditory cue), but specifically in training and tutorial environments, feedback is critical in making trainees aware of mistakes and in creating an appropriate mental model. Feedback is particularly critical for older adults (for a review, see Kausler, 1991). Given that they may experience a higher degree of anxiety at learning to interact with novel technology (Fisk et al., 2009), this feedback should be communicated as clearly as possible. In a training tutorial that teaches older adults basic system functions:

- *Feedback should be immediate*. If a user performs steps out of order, the training application should provide immediate feedback to prevent this incorrect order of operations from becoming a learned procedure.
- *Feedback should be specific*. Users should be informed of their incorrect action and shown the correct action.
- *Feedback should be succinct*. Removing the user from the training program for an extended period of time to explain a mistake will prevent the user from quickly learning the correct procedure.

### 6.6.6 Consistency

Learning will not occur for completely inconsistent information, but like younger adults, older adults can learn under partially consistent conditions (Meyer and Fisk, 1998). However, for older adults, consistent relationships between stimuli or aspects of a task or system should be identified explicitly. Because the hypothetical system will run multiple different pieces of game software, it would be to the advantage of designers to have consistent mappings between each system input and game action (e.g., the same buttons used to jump or fire weapons in one game would be used in other games requiring these actions).

### 6.6.7 Importance of Task Analysis

A device that seems simple to designers often is much more complicated and difficult to use for novice users. For example, several commercially available gaming systems require a number of steps before a game can even be started. By performing a comprehensive task analysis, the individual requirements for successful interaction with the system or device will be identified and accounted for in the training and instructional materials. Furthermore, if part-task training is plausible, a careful analysis of all aspects of the task is necessary to segment the task appropriately. Task analysis will result in an understanding of problems and errors that can occur. These issues should be anticipated in instructional manuals; for example, in a "frequently asked questions" section of a manual.

## 7 CONCLUSIONS

Older adults comprise a significant portion of users of technological products, thus demanding the attention of human factors professionals. We have described a wide

array of age-related factors that have functional significance for the performance of older adults and provided design implications and recommendations based on the existing literature. Human factors professionals will benefit by considering the age of their users and designing appropriately.

Several themes in the design guidelines follow from the review of age-related changes and from the case study. The importance of environmental support, or taking cognitively demanding requirements from the user and putting information in the task environment, is pervasive. Especially for older adults, this supportive information can direct attention to relevant stimuli; cue users' memory; support tasks such as reading, visual search, and even balance; and free up valuable cognitive resources that can be applied to other aspects of the task. Resources can also be allocated to other tasks if the stimulus itself is improved. Despite pervasive perceptual declines in the aging system, the perceptual stimulus can be greatly improved by increasing the size or intensity of a stimulus, increasing the lighting of an environment, or minimizing background clutter. The designer must understand the limitations of older users before even such simple design adjustments can be made.

For tasks that older adults continue to have difficulty with after these two guidelines have been followed, older adults may still be capable of attaining proficiency through appropriate training. Older adults are able to achieve skilled performance through training provided that the information to be learned is at least partially consistent. Older adults follow a power law of learning for consistent information similar to that of younger adults. Thus, older adults are certainly not limited in all aspects relative to their younger counterparts. In fact, designers should attempt to take advantage of those areas where older adults surpass the capabilities of younger adults, specifically in their semantic knowledge. However, none of these design guidelines are useful if older adults choose not to use the system or product. Fortunately, data suggest that older adults are willing, and it is critical that the benefits of new systems and products are communicated to potential older users to increase their motivation to learn to use these new technologies.

Designers should also be aware of tools that can simulate and predict older adult performance and, after a task analysis, can help designers choose between alternative designs. One such tool is the goals, operators, methods, and selection (GOMS) rules modeling technique. Originally modeled after the processing speeds and capacities of younger adults, parameters of the model are now available to simulate and predict older adults' interactions with technology (Jastrzembski and Charness, 2007). These updated parameters take into account age-related declines in memory, speed, and motor control. Designers should be cautious when using tools based on young adult data.

Age-related changes in capabilities of older adults have been well studied. An understanding of such differences provides constraints for the design space of new products and new instantiations of existing products. Clearly, iterative design and user testing will always be essential for good design, and older adults must be included in the usability test group. However, our goal in this chapter was to provide a summary of the literature on aging to enable designers to start from an informed position about how systems and products should be designed if they are to be used safely and effectively by older adults.

## ACKNOWLEDGMENTS

## REFERENCES

Abel, S. M., Giguere, C., Consoli, A., and Papsin, B. (2000), "The Effect of Aging on Horizontal Plane Sound Localization," *Journal of the Acoustical Society of America*, Vol. 108, No. 2, pp. 743–752.

Agnew, J., and Suruda, A. (1993), "Age and Fatal Work-Related Falls," *Human Factors*, Vol. 35, pp. 731–736.

Andersen, G. J., Cisneros, J., Saidpour, A., and Atchley, P. (2000), "Age-Related Differences in Collision Detection during Deceleration," *Psychology and Aging*, Vol. 15, pp. 241–252.

Arenberg, D., and Robertson-Tchabo, E. A. (1985), "Adult Age Differences in Memory and Linguistic Integration Revisited," *Experimental Aging Research*, Vol. 11, pp. 187–191.

Atchley, P., and Andersen, G. J. (1998), "The Effects of Age, Retinal Eccentricity, and Speed on the Detection of Optic Flow Components," *Psychology and Aging*, Vol. 13, pp. 297–308.

Baddeley, A. (1986), *Working Memory*, Clarendon, Oxford.

Basak, C., Boot, W. R., Voss, M. W., and Kramer, A. F. (2008). "Can Training in a Real-Time Strategy Video Game Attenuate Cognitive Decline in Older Adults?" *Psychology and Aging*, Vol. 23, No. 4, pp. 765–777.

Bellmann, C., and Holz, F. G. (2001), "Visual Impairment and Functional Deficits in Age-Related Macular Degeneration," in *On the Special Needs of Blind and Low-Vision Seniors*, H. -W. Wahl and H. -E. Schulze, Eds., IOS Press, Amsterdam, pp. 49–57.

Bennett, A. G. (1965), "Ophthalmic Test Types," *British Journal of Physiological Optics*, Vol. 22, pp. 238–271.

Bherer, L., Kramer, A. F., Peterson, M. S., Colcombe, S., Erickson, K., and Becic, E. (2005), "Training Effects on Dual-Task Performance: Are There Age-Related Differences in Plasticity of Attentional Control?" *Psychology and Aging*, Special Issue: Emotion-Cognition Interactions and the Aging Mind, Vol. 20, No. 4, pp. 695–709.

Borger, L. L., Whitney, S. L., Redfern, M. S., and Furman, J. M. (1999), "The Influence of Dynamic Visual Environments on Postural Sway in the Elderly," *Journal of Vestibular Research*, Vol. 9, pp. 197–205.

Brown, L., Shumway-Cook, A., and Woollacott, M. (1999), "Attentional Demands and Postural Recovery: The Effects of Aging," *Journals of Gerontology: Medical Sciences*, Vol. 54A, pp. M165–M171.

Bugg, J. M., Zook, N. A., Delosh, E. L., Davalos, D. B., and Davis, H. P. (2006), "Age Differences in Fluid Intelligence: Contributions of General Slowing and Frontal Decline," *Brain and Cognition*, Vol. 62, pp. 9–16.

Carnahan, H., Goodale, M. A., and Marteniuk, R. G. (1993), "Grasping Versus Pointing and the Differential Use of Visual Feedback," *Human Movement Science*, Vol. 12, pp. 219–234.

Cattell, R. B. (1963), "Theory of Fluid and Crystallized Intelligence: A Critical Experiment," *Journal of Educational Psychology*, Vol. 54, pp. 1–22.

Charness, N., and Dijkstra, K. (1999), "Age, Luminance, and Print Legibility in Homes, Offices, and Public Places," *Human Factors*, Vol. 41, No. 2, pp. 173–193.

Charness, N., Holley, P., Feddon, J., and Jastrzembski, T. (2004), "Light Pen Superiority over a Mouse: Age, Hand, and Practice Effects," *Human Factors*, Vol. 46, pp. 373–384.

Chun, M., and Jiang, Y. (1998), "Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention," *Cognitive Psychology*, Vol. 36, pp. 28–71.

Cohen, G., and Faulkner, D. (1989), "Age Differences in Source Forgetting: Effects on Reality Monitoring and on Eyewitness Testimony," *Psychology and Aging*, Vol. 4, No. 1, pp. 10–17.

Cole, K. J. (1991), "Grasp Force Control in Older Adults," *Journal of Motor Behavior*, Vol. 23, No. 4, pp. 251–258.

Cole, K. J., Rotella, D. L., and Harper, J. G. (1999), "Mechanisms for Age-Related Changes of Fingertip Forces during Precision Gripping and Lifting in an Adult," *Journal of Neuroscience*, Vol. 19, No. 8, pp. 3238–3247.

Connelly, S. L., and Hasher, L. (1993), "Aging and the Inhibition of Spatial Location," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 19, pp. 1238–1250.

Craik, F. I. M. (1986), "A Functional Account of Age Differences in Memory," in *Human Memory and Cognitive Capabilities: Mechanisms and Performances*, F. Klix and H. Hagendorf, Eds., Elsevier North-Holland, Amsterdam, pp. 409–422.

Crossley, M., and Hiscock, M. (1992), "Age-Related Differences in Concurrent-Task Performance of Normal Adults: Evidence for a Decline in Processing Resources," *Psychology and Aging*, Vol. 7, pp. 499–506.

Czaja, S. J., and Sharit, J. (1998), "The Effect of Age and Experience on the Performance of a Data Entry Task," *Journal of Experimental Psychology: Applied*, Vol. 4, pp. 332–351.

Czaja, S. J., Charness, N., Fisk, A. D., Hertzog, C., Nair, S. N., Rogers, W. A., and Sharit, J. (2006), "Factors Predicting the Use of Technology: Findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE)," *Psychology and Aging*, Vol. 21, pp. 333–352.

Desai, M., Pratt, L. A., Lentzner, H., and Robinson, K. N. (2001), "Trends in Vision and Hearing among Older Americans," in *Aging Trends*, Vol. 2, National Center for Health Statistics, Hyattsville, MD.

Dingus, T. A., Hulse, M. C., Mollenhauer, M. A., Fleischman, R. N., McGehee, D. V., and Manakkal, N. (1997), "Effects of Age, System Experience, and Navigation Technique on Driving with an Advanced Traveler Information System," *Human Factors*, Vol. 39, No. 2, pp. 177–199.

Doane, S. M., Alderton, D. L., Sohn, Y. W., and Pellegrino, J. W. (1996), "Acquisition and Transfer of Skilled Performance: Are Visual Discrimination Skills Stimulus Specific?" *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 22, pp. 1218–1248.

Dunlosky, J., and Hertzog, C. (1998), "Aging and Deficits in Associative Memory: What Is the Role of Strategy Use?" *Psychology and Aging*, Vol. 13, pp. 597–607.

Earles, J., Smith, A. D., and Park, D. C. (1996), "Adult Age Differences in the Effects of Environmental Context on Memory Performance," *Experimental Aging Research*, Vol. 22, pp. 267–280.

Einstein, G. O., McDaniel, M. A., Richardson, S. L., Guynn, M. J., and Cunfer, A. R. (1995), "Aging and Prospective Memory: Examining the Influences of Self-Initiated Retrieval Processes," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 21, No. 4, pp. 996–1007.

Einstein, G. O., Smith, R. E., McDaniel, M. A., and Shaw, P. (1997), "Aging and Prospective Memory: The Influence of Increased Task Demands at Encoding and Retrieval," *Psychology and Aging*, Vol. 12, pp. 479–488.

Ellis, R. D., and Allaire, J. (1999), "Modeling Computer Interest in Older Adults: The Role of Age, Education, Computer Knowledge, and Computer Anxiety," *Human Factors*, Vol. 41, No. 3, pp. 345–355.

Ellis, R. D., and Kurniawan, S. H. (2000), "Increasing the Usability of On-Line Information for Older Users: A Case Study in Participatory Design," *International Journal of Human–Computer Interaction*, Vol. 12, pp. 263–276.

Fausset, C. B. (2008). "On Processing Line Graphs: Understanding Aging and the Role of Spatial and Verbal Resources," unpublished master's thesis, Georgia Institute of Technology, Atlanta, GA.

Ferguson, S. A., Hashtroudi, S., and Johnson, M. K. (1992), "Age Differences in Using Source-Relevant Cues," *Psychology and Aging*, Vol. 7, pp. 443–452.

Fisk, A. D. (1999), "Human Factors and the Older Adult," *Ergonomics in Design*, Vol. 7, No. 1, pp. 8–13.

Fisk, A. D., and Rogers, W. A. (1991), "Toward an Understanding of Age-Related Memory and Visual Search Effects," *Journal of Experimental Psychology: General*, Vol. 120, pp. 131–149.

Fisk, A. D., and Rogers, W. A. (2002), "Health Care of Older Adults: The Promise of Human Factors Research," in *Human Factors Interventions for the Health Care of Older Adults*, W. A. Rogers and A. D. Fisk, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 31–46.

Fisk, A. D., McGee, N. D., and Giambra, L. M. (1988), "The Influence of Age on Consistent and Varied Semantic Category Search Performance," *Psychology and Aging*, Vol. 3, pp. 323–333.

Fisk, A. D., Rogers, W. A., Charness, N., Czaja, S. J., and Sharit, J. (2009), *Designing for Older Adults: Principles and Creative Human Factors Approaches*, 2nd ed., CRC Press, Boca Raton, FL.

Fozard, J. L. (1990), "Vision and Hearing in Aging," in *Handbook of the Psychology of Aging*, 3rd ed., J. E. Birren and K. W. Schaie, Eds., Academic, San Diego, CA, pp. 150–170.

Fozard, J. L., and Gordon-Salant, S. (2001), "Sensory and Perceptual Changes with Aging," in *Handbook of the Psychology of Aging*, 5th ed., J. E. Birren and K. W. Schaie, Eds., Academic, San Diego, CA, pp. 241–266.

Gilbert, D. K., and Rogers, W. A. (1999), "Age-Related Differences in the Acquisition, Utilization, and Extension of a Spatial Mental Model," *Journals of Gerontology: Psychological Sciences*, Vol. 54B, pp. P246–P255.

Goldstein, E. B. (1999), *Sensation and Perception*, Brooks/ Cole, Pacific Grove, CA.

Gopher, D. (2007), "Emphasis Change as a Training Protocol for High-Demand Tasks," in *Attention: From Theory to Practice*, A. F. Kramer, D. A. Wiegmann, and A. Kirlik, Eds., Oxford University Press, New York, pp. 209–224.

Green, C. S., and Bavelier, D. (2008), "Exercising Your Brain: A Review of Human Brain Plasticity and Training-Induced Learning," *Psychology and Aging*, Vol. 23, No. 4, pp. 692–701.

Grünwald, M. (2008), *Human Haptic Perception: Basics and Applications*, Birkhäuser, Boston.

Haider, H., and Frensch, P. A. (1996), "The Role of Information Reduction in Skill Acquisition," *Cognitive Psychology*, Vol. 30, pp. 304–337.

Hamm, V. P., and Hasher, L. (1992), "Age and the Availability of Inferences," *Psychology and Aging*, Vol. 7, pp. 56–64.

Hancock, H. E., Fisk, A. D., and Rogers, W. A. (2001), "An Evaluation of Warning Habits and Beliefs Across the Adult Life Span," *Human Factors*, Vol. 43, No. 3, pp. 343–354.

Hancock, H. E., Rogers, W. A., Schroeder, D., and Fisk, A. D. (2004), "Safety Symbol Comprehension: Effects of Symbol Type, Familiarity, and Age," *Human Factors*, Vol. 46, No. 2, pp. 183–195.

Hancock, H. E., Fisk, A. D., and Rogers, W. A. (2005), "Comprehending Product Warning Information: Age-Related Effects of Memory, Inferencing, and Knowledge," *Human Factors*, Vol. 47, pp. 219–234.

Hartley, A. A., Kieley, J., and McKenzie, C. R. M. (1992), "Allocation of Visual Attention in Younger and Older Adults," *Perception and Psychophysics*, Vol. 52, pp. 175–185.

Hasher, L., and Zacks, R. T. (1988), "Working Memory, Comprehension, and Aging: A Review and a New View," in *The Psychology of Learning and Motivation*, G. H. Bower, Ed., Academic, New York, Vol. 22, pp. 193–225.

Hertzog, C., Kramer, A. F., Wilson, R. S., and Lindenberger, U. (2009), "Enrichment Effects on Adult Cognitive Development," *Psychological Science in the Public Interest*, Vol. 9, pp. 1–65.

Hickman, J. M., Rogers, W. A., and Fisk, A. D. (2007), "Training Older Adults to Use New Technology," *Journal of Gerontology: Psychological Sciences*, Vol. 62B, pp. 77–84.

Hirshman, E., and Bjork, R. (1988), "The Generation Effect: Support for a Two-Factor Theory," *Journal of Experimental Psychology: Learning Memory and Cognition*, Vol. 14, pp. 484–494.

Ho, G., Scialfa, C. T., and Caird, J. K. (2001), "Visual Search for Traffic Signs: The Effects of Clutter, Luminance, and Aging," *Human Factors*, Vol. 43, pp. 194–207.

Horak, F. B., Shupert, C. L., and Mirka, A. (1989), "Components of Postural Dyscontrol in the Elderly: A Review," *Neurobiology of Aging*, Vol. 10, pp. 727–738.

Hulicka, I. M., and Grossman, J. L. (1967), "Age Group Comparisons for the Use of Mediators in Paired-Associate Learning," *Journal of Gerontology*, Vol. 22, pp. 46–51.

Jackson, G. R., Owsley, C., and McGwin, G. (1999), "Aging and Dark Adaptation," *Vision Research*, Vol. 39, pp. 3975–3982.

James, W. (1890/1950), *The Principles of Psychology*, Vol. 1, Dover, New York.

Jastrzembski, T. S., and Charness, N. (2007), "The Model Human Processor and the Older Adult: Parameter Estimation and Validation within a Mobile Phone Task," *Journal of Experimental Psychology: Applied*, Vol. 13, pp. 224–248.

Jiang, Y., and Chun, M. M. (2001), "Selective Attention Modulates Implicit Learning," *Quarterly Journal of Experimental Psychology*, Vol. 54A, pp. 1105–1124.

Johnson, M. K., Hashtroudi, S., and Lindsay, D. S. (1993), "Source Monitoring," *Psychological Bulletin*, Vol. 114, pp. 3–28.

Johnson, M. K., De Leonardis, D., Hashtroudi, S., and Ferguson, S. A. (1995), "Aging and Single versus Multiple Cues in Source Monitoring," *Psychology and Aging*, Vol. 10, pp. 507–517.

Johnson, M. M., Schmitt, F. A., and Pietrukowicz, M. (1989), "The Memory Advantages of the Generation Effect: Age and Process Differences," *Journals of Gerontology: Psychological Sciences*, Vol. 44, No. 3, pp. P91–P94.

Johnston, W. A., and Dark, V. J. (1986), "Selective Attention," *Annual Review of Psychology*, Vol. 37, pp. 43–75.

Jones, M. B. (1989), "Individual Differences in Skill Retention," *American Journal of Psychology*, Vol. 102, pp. 183–196.

Kallman, D. A., Plato, C. C., and Tobin, J. D. (1990), "The Role of Muscle Loss in the Age-Related Decline of Grip Strength: Cross-Sectional and Longitudinal Perspectives," *Journal of Gerontology*, Vol. 45, pp. M82–M88.

Kaufman, D. R., Patel, V. L., Hilliman, C., Morin, P. C., Pevzner, J., Weinstock, G. R., Shea, S., and Starren, J. (2003), "Usability in the Real World: Assessing Medical Information Technologies in Patients' Homes," *Journal of Biomedical Informatics*, Vol. 36, pp. 45–60.

Kausler, D. H. (1991), *Experimental Psychology, Cognition, and Human Aging*, Springer-Verlag, New York.

Kemper, S. (1987), "Constraints on Psychological Processes in Discourse Production," in *Psycholinguistic Models of Production*, H. W. Dechert and M. Raupach, Eds., Ablex, Westport, CT, pp. 185–188.

Ketcham, C. J., and Stelmach, G. E. (2001), "Age-Related Declines in Motor Control," in *Handbook of the Psychology of Aging*, 5th ed., J. E. Birren and K. W. Schaie, Eds., Academic, San Diego, CA, pp. 313–348.

Ketcham, C. J., Seidler, R. D., van Gemmert, A. W. A., and Stelmach, G. E. (2002), "Age Related Kinematic Differences as Influenced by Task Difficulty, Target-Size, and Movement Amplitude," *Journals of Gerontology: Psychological Sciences and Social Sciences*, Vol. 57B, pp. P54–P64.

Kinsella, K., and Velkoff, V. A. (2001), *An Aging World: 2001*, U.S. Census Bureau, Series P95/01-1, U.S. Government Printing Office, Washington, DC.

Kline, D. W., Buck, K., Sell, Y., Bolan, T., and Dewar, R. E. (1999), "Older Observers' Tolerance of Optical Blur: Age Differences in the Identification of Defocused Text Signs," *Human Factors*, Vol. 41, No. 3, pp. 356–364.

Knoblauch, K., Saunders, F., Kusuda, M., Hynes, R., Podgor, M., Higgens, K. E., and deMosasterio, M. (1987), "Age and Illuminance Effects in the Farnsworth–Munsell 100-Hue Test," *Applied Optics*, Vol. 26, pp. 1441–1448.

Korteling, J. E. (1993), "Effects of Age and Task Similarity on Dual-Task Performance," *Human Factors*, Vol. 35, No. 1, pp. 99–113.

Kraft, J. M., and Werner, J. S. (1999), "Aging and the Saturation of Colors, 1: Colorimetric Purity Discrimination," *Journal of the Optical Society of America A*, Vol. 16, pp. 223–230.

Kramer, A. F., and Larish, J. (1996), "Aging and Dual-Task Performance," in *Aging and Skilled Performance*, W. A. Rogers, A. D. Fisk, and N. Walker, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 83–112.

Kramer, A. F., Hahn, S., and Gopher, D. (1999), "Task Coordination and Aging: Explorations of Executive Control Processes in the Task Switching Paradigm." *Acta Psychologica*, Vol. 101, pp. 339–378.

Kramer, A. F., Humphrey, D. G., Larish, J. F., Logan, G. D., and Strayer, D. L. (1994), "Aging and Inhibition: Beyond a Unitary View of Inhibitory Processing in Attention," *Psychology and Aging*, Vol. 9, pp. 491–512.

Kramer, A. F., Larish, J. F., and Strayer, D. L. (1995), "Training for Attentional Control in Dual-Task Settings: A Comparison of Young and Old Adults," *Journal of Experimental Psychology: Applied*, Vol. 1, pp. 50–76.

Kristinsdottir, E. K., Fransson, P. -A., and Magnusson, M. (2001), "Changes in Postural Control in Healthy Elderly Are Related to Vibration Sensation, Vision and Vestibular Asymmetry," *Acta Otolaryngologica*, Vol. 121, pp. 700–706.

Lajoie, Y., Teasdale, N., Bard, C., and Fleury, M. (1993), "Attentional Demands for Static and Dynamic Equilibrium," *Experimental Brain Research*, Vol. 97, No. 1, pp. 139–144.

Lajoie, Y., Teasdale, N., Bard, C., and Fleury, M. (1996), "Attentional Demands for Walking: Age-Related Changes," in *Changes in Sensory Motor Behavior in Aging*, A. Ferrandez and N. Teasdale, Eds., Elsevier, Amsterdam, Vol. 114, pp. 235–256.

Legge, G. E., Mullen, K. T., Woo, G. C., and Campbell, F. W. (1987a), "Tolerance to Visual Defocus," *Journal of the Optical Society of America*, Vol. A4, pp. 851–863.

Legge, G. E., Rubin, G. S., and Luebker, A. (1987b), "Psychophysics of Reading, V: The Role of Contrast in Normal Vision," *Vision Research*, Vol. 27, pp. 1165–1171.

Lenhart, A., Jones, S., and Rankin Macgill, A. (2008), "PEW Internet Project Data Memo: Adults and Video Games," Pew Internet & American Life Project, available: http://www.pewinternet.org/index.asp, accessed June 20, 2010.

Li, K. Z. H., Lindenberger, U., Freund, A. M., and Baltes, P. B. (2001), "Walking While Memorizing: Age-Related Differences in Compensatory Behavior," *Psychological Science*, Vol. 12, pp. 230–237.

Lockhart, T. E. (1997), "Ability of Elderly People to Traverse Slippery Walking Surfaces," in *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 125–129.

Lorenzi, C., Gatehouse, S., and Lever, C. (1999), "Sound Localization in Noise in Hearing-Impaired Listeners," *Journal of the Acoustical Society of America*, Vol. 105, No. 6, pp. 3454–3463.

Maltz, M., and Shinar, D. (1999), "Eye Movements of Younger and Older Drivers," *Human Factors*, Vol. 41, No. 1, pp. 15–25.

McDowd, J. M., and Craik, F. I. M. (1988), "Effects of Aging and Task Difficulty on Divided Attention Performance," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 14, pp. 267–280.

McDowd, J., and Shaw, R. J. (2000), "Attention and Aging," in *The Handbook of Aging and Cognition*, 2nd ed., F. I. M. Craik and T. A. Salthouse, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 221–292.

McLaughlin, A. C., Rogers, W. A., and Fisk, A. D. (2002), "Effectiveness of Audio and Visual Training Presentation Modes for Glucometer Calibration," in *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA.

McLaughlin, G. (1969), "SMOG Grading: A New Readability Formula," *Journal of Reading*, Vol. 12, No. 8, pp. 639–646.

Melenhorst, A. S., Rogers, W. A., and Caylor, E. C. (2001), "The Use of Communication Technologies by Older Adults: Exploring the Benefits from the User's Perspective," in *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA.

Melzer, I., Benjuya, N., and Kaplanski, J. (2001), "Age Related Changes of Postural Control: The Effect of Cognitive Task," *Gerontology*, Vol. 47, pp. 189–194.

Metter, E. J., Conwit, R., Tobin, J., and Fozard, J. L. (1997), "Age-Associated Loss of Power and Strength in the Upper Extremities in Women and Men," *Journals of Gerontology: Biological Sciences*, Vol. 52A, pp. B267–B276.

Meyer, B., and Fisk, A. D. (1998), "Toward an Understanding of Age-Related Use of Incidental Consistency," in *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA.

Mitzner, T. L., and Rogers, W. A. (2003), "Age-Related Differences in Reading Text Presented with Degraded Contrast," in *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA.

Mitzner, T. L., Fausset, C. B., Boron, J. B., Adams, A. E., Dijkstra, K., Lee, C. C., Rogers, W. A., and Fisk, A. D. (2008), "Older Adults' Training Preferences for Learning to Use Technology," in *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 2047–2051.

Morrell, R. W., and Echt, K. V. (1997), "Designing Written Instructions for Older Adults: Learning to Use Computers," in *Handbook of Human Factors and the Older Adult*, A. D. Fisk and W. A. Rogers, Eds., Academic, San Diego, CA, pp. 335–361.

Morrell, R. W., Mayhorn, C. B., and Bennett, J. (2000), "A Survey of World Wide Web Use in Middle-Aged and Older Adults," *Human Factors*, Vol. 42, pp. 175–182.

Morrow, D. G. (2003), "Technology as Environmental Support for Older Adults' Daily Activities," in *The Impact of Technology on Successful Aging*, N. Charness and W. Schaie, Eds., Springer-Verlag, New York, pp. 290–305.

Morrow, D. G., and Rogers, W. A. (2008), "Environmental Support: An Integrative Framework," *Human Factors*, Vol. 50, pp. 589–613.

Morrow, D. G., Leirer, V. O., Carver, L. M., and Tanke, E. D. (1998), "Older and Younger Adult Memory for Health Appointment Information: Implications for Automated Telephone Messaging Design," *Journal of Experimental Psychology: Applied*, Vol. 4, No. 4, pp. 352–374.

Morrow, D. G., Carver, L. M., Leirer, V. O., and Tanke, E. D. (2000), "Medication Schemas and Memory for Automated Telephone Messages," *Human Factors*, Vol. 42, pp. 523–540.

Moscicki, E. K., Elkins, E. F., Baum, H. M., and McNamara, P. M. (1985), "Hearing Loss in the Elderly: An Epidemiologic Study of the Framingham Study Cohort," *Ear and Hearing*, Vol. 6, No. 8, pp. 184–190.

Mykityshyn, A. L., Fisk, A. D., and Rogers, W. A. (2002), "Learning to Use a Home Medical Device: Mediating Age-Related Differences with Training," *Human Factors*, Vol. 44, pp. 354–364.

Nichols, T. A., and Fisk, A. D. (2001, August), "Age-Related Differences in Using Environmental Support: Resources or Strategies?" paper presented at the 109th Annual APA Convention, San Francisco.

Nichols, T. A., Rogers, W. A., and Fisk, A. D. (2003), "Do You Know How Old Your Participants Are? Recognizing the Importance of Participant Age Classifications," *Ergonomics in Design*, Vol. 11, pp. 22–26.

Nichols, T. A., Rogers, W. A., and Fisk, A. D. (2006), "Design for Aging," in *Handbook of Human Factors and Ergonomics*, 3rd ed., G. Salvendy Ed., Wiley, Hoboken, NJ, pp. 1418–1445.

Noble, W., Byrne, D., and Lepage, B. (1994), "Effects on Sound Localization of Configuration and Type of Hearing Impairment," *Journal of the Acoustical Society of America*, Vol. 95, No. 2, pp. 992–1005.

Norman, S., Kemper, S., and Kynette, D. (1992), "Adults' Reading Comprehension: Effects of Syntactic Complexity and Working Memory," *Journal of Gerontology*, Vol. 47, No. 4, pp. 258–265.

Olson, K. E., Fisk, A. D., and Rogers, W. A. (2009), "Collaborative Automated Systems: Older Adults' Mental Model Acquisition and Trust in Automation," in *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting*, San Antonio, TX, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1704–1708.

Olson, K. A., O'Brien, M. A., Rogers, W. A., and Charness, N. (2011), "Diffusion of Technology for Younger and Older Adults," *Ageing International*, Vol. 36, pp. 123–145.

Owsley, C., Ball, K., Sloane, M., Roenker, D., and Bruni, J. (1991), "Visual/Cognitive Correlates of Vehicle Accidents in Older Drivers," *Psychology and Aging*, Vol. 6, pp. 403–415.

Parasuraman, R., and Davies, D. R., Eds. (1984), *Varieties of Attention*, Academic, New York.

Park, D. C., Cherry, K. E., Smith, A. D., and LaFronza, V. N. (1990a), "Effects of Distinctive Context on Memory for Objects and Their Locations in Young and Older Adults," *Psychology and Aging*, Vol. 5, pp. 250–255.

Park, D. C., Smith, A. D., Morrell, R. W., Puglisi, J. T., and Dudley, W. N. (1990b), "Effects of Contextual Integration on Recall of Pictures in Older Adults," *Journals of Gerontology: Psychological Sciences*, Vol. 45, pp. P52–P58.

Park, D. C., Morrell, R. W., Frieske, D., and Kincaid, D. (1992), "Medication Adherence Behaviors in Older Adults: Effects of External Cognitive Supports," *Psychology and Aging*, Vol. 7, pp. 252–256.

Park, D. C., Morrell, R., Hertzog, C., Kidder, D., and Mayhorn, C. (1997), "Effect of Age on Event-Based and Time-Based Prospective Memory," *Psychology and Aging*, Vol. 12, pp. 314–327.

Pinckers, A. (1980), "Color Vision and Age," *Ophthalmologica*, Vol. 181, pp. 23–30.

Radvansky, G. A., Gerard, L. D., Zacks, R. T., and Hasher, L. (1990), "Younger and Older Adults Use of Models as Representations for Text Materials," *Psychology and Aging*, Vol. 5, pp. 209–214.

Redfern, M. S., Moore, P. L., and Yarsky, C. M. (1997), "The Influence of Flooring on Standing Balance among Older Persons," *Human Factors*, Vol. 39, No. 3, pp. 445–455.

Resnick, S. M., Pham, D. L., Kraut, M. A., Zonderman, A. B., and Davatzikos, C. (2003), "Longitudinal Magnetic Resonance Imaging Studies of Older Adults: A Shrinking Brain," *Journal of Neuroscience*, Vol. 23, pp. 3295–3301.

Roenker, D. L., Cissell, G. M., Ball, K. K., Wadley, V. G., and Edwards, J. D. (2003), "Speed-of-Processing and Driving Simulator Training Result in Improved Driving Performance," *Human Factors*, Vol. 45, pp. 218–233.

Rogers, W. A. (1992), "Age Differences in Visual Search: Target and Distractor Learning," *Psychology and Aging*, Vol. 7, pp. 526–535.

Rogers, W. A., and Fisk, A. D. (2001), "Understanding the Role of Attention in Cognitive Aging Research," in *Handbook of the Psychology of Aging*, 5th ed., J. E. Birren and K. W. Schaie, Eds., Academic, San Diego, CA, pp. 267–287.

Rogers, W. A., and Gilbert, D. K. (1997), "Do Performance Strategies Mediate Age-Related Differences in Associative Learning?" *Psychology and Aging*, Vol. 12, pp. 620–633.

Rogers, W. A., Bertus, E. L., and Gilbert, D. K. (1994), "A Dual-Task Assessment of Age Differences in Automatic Process Development," *Psychology and Aging*, Vol. 9, pp. 398–413.

Rogers, W. A., Fisk, A. D., Mead, S. E., Walker, N., and Cabrera, E. F. (1996), "Training Older Adults to Use Automatic Teller Machines," *Human Factors*, Vol. 38, pp. 425–433.

Rogers, W. A., Meyer, B., Walker, N., and Fisk, A. D. (1998), "Functional Limitations to Daily Living Tasks in the Aged: A Focus Group Analysis," *Human Factors*, Vol. 40, pp. 111–125.

Rogers, W. A., Hertzog, C., and Fisk, A. D. (2000), "Age-Related Differences in Associative Learning: An Individual Differences Analysis of Ability and Strategy Influences," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 26, pp. 359–394.

Rogers, W. A., Mykityshyn, A. L., Campbell, R. H., and Fisk, A. D. (2001), "Analysis of a 'Simple' Medical Device," *Ergonomics in Design*, Vol. 9, pp. 6–14.

Rogers, W. A., Fisk, A. D., McLaughlin, A. C., and Pak, R. (2005), "Touch a Screen or Turn a Knob: Choosing the Best Device for the Job," *Human Factors*, Vol. 47, pp. 271–288.

Royall, D. R., Palmer, R., Chiodo, L. K., and Polk, M. J. (2004), "Declining Executive Control in Normal Aging Predicts Change in Functional Status: The Freedom House Study," *Journal of the American Geriatrics Society*, Vol. 52, No. 3, pp. 346–352.

Said, F. S., and Weale, R. A. (1959), "The Variation with Age of the Spectral Transmissivity of the Living Human Crystalline Lens," *Gerontologica*, Vol. 3, pp. 213–231.

Sanchez, J., Nichols, T. A., Mitzner, T. L., Rogers, W. A., and Fisk, A. D. (2003), "Medication Adherence Strategies in Older Adults," in *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA.

Schieber, F., and Goodspeed, C. H. (1997), "Nighttime Conspicuity of Highway Signs as a Function of Sign Brightness, Background Complexity and Age of Observer," in *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1362–1366.

Schneider, B., and Pichora-Fuller, M. K. (2000), "Implications of Sensory Deficits for Cognitive Aging," in *The Handbook of Aging and Cognition*, 2nd ed., F. I. M. Craik and T. Salthouse, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 155–219.

Schneider, B., Daneman, M., Murphy, D., and See, S. (2000), "Listening to Discourse in Distracting Settings: The Effects of Aging," *Psychology and Aging*, Vol. 15, pp. 110–125.

Schneider, W. (1985), "Toward a Model of Attention and the Development of Automatic Processing," in *Attention and Performance XI*, M. Posner and O. S. Marin, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 475–492.

Schneider, W., and Shiffrin, R. M. (1977), "Controlled and Automatic Human Information Processing, I: Detection, Search, and Attention," *Psychological Review*, Vol. 84, pp. 1–66.

Seidler-Dobrin, R. D., and Stelmach, G. E. (1998), "Persistence in Visual Feedback Control by the Elderly," *Experimental Brain Research*, Vol. 119, pp. 467–474.

Sharit, J., Czaja, S. J., Nair, S., and Lee, C. C. (2003), "Effects of Age, Speech Rate, and Environmental Support in Using Telephone Voice Menu Systems," *Human Factors*, Vol. 45, No. 2, pp. 234–251.

Shaw, R. J. (1990), "Older Adults Sometimes Benefit from Environmental Support: Evidence from Reading Distorted Text," in *Proceedings of the Human Factors Society 34th Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 168–172.

Shaw, R. J. (1991), "Age-Related Increases in the Effects of Automatic Semantic Activation," *Psychology and Aging*, Vol. 6, pp. 595–604.

Shiffrin, R. M. (1988), "Attention," in *Stevens' Handbook of Experimental Psychology*, 2nd ed., R. C. Atkinson, R. J. Herrnstein, G. Lindzey, and R. D. Luce, Eds., Wiley, New York, pp. 739–811.

Shiffrin, R. M., and Schneider, W. (1977), "Controlled and Automatic Human Information Processing, II: Perceptual Learning, Automatic Attending, and a General Theory," *Psychological Review*, Vol. 84, pp. 127–190.

Sit, R. A., and Fisk, A. D. (1999), "Age-Related Performance in a Multiple-Task Environment," *Human Factors*, Vol. 41, pp. 26–34.

Smith, M. W., Sharit, J., and Czaja, S. J. (1999), "Aging, Motor Control, and the Performance of Computer Mouse Tasks," *Human Factors*, Vol. 41, No. 3, pp. 389–396.

Spirduso, W. W. (1995), *Physical Dimensions of Aging*, Human Kinetics, Champaign, IL.

Staplin, L. K., and Fisk, A. D. (1991), "Left-Turn Intersection Problems: A Cognitive Engineering Approach to Improve the Safety of Young and Old Drivers," *Human Factors*, Vol. 33, pp. 559–571.

Stelmach, G. E., and Nahom, A. (1992), "Cognitive–Motor Abilities of the Elderly Driver," *Human Factors*, Vol. 34, No. 1, pp. 53–65.

Stelmach, G. E., Amrhein, P. C., and Goggin, N. L. (1988), "Age Differences in Bimanual Coordination," *Journals of Gerontology: Psychological Sciences*, Vol. 43, pp. P18–P23.

Stoltzfus, E. R., Hasher, L., Zacks, R. T., Ulivi, M. S., and Goldstein, D. (1993), "Investigations of Inhibition and Interference in Younger and Older Adults," *Journal of Gerontology: Psychological Sciences*, Vol. 48, pp. P179–P188.

Szuchman, L. T., and Erber, J. T. (1990), "Young and Older Adults' Metaphor Interpretation: The Judgments of Professionals and Nonprofessionals," *Experimental Aging Research*, Vol. 16, No. 2, pp. 62–72.

Thornberry, J. M., and Mistretta, C. M. (1981), "Tactile Sensitivity as a Function of Age," *Journal of Gerontology*, Vol. 36, No. 1, pp. 34–39.

Tran, D. B., Silverman, S. E., Zimmerman, K., and Feldon, S. E. (1998), "Age-Related Deterioration of Motion Perception and Detection," *Graefe's Archive of Clinical Experimental Ophthalmology*, Vol. 236, No. 4, pp. 269–273.

Trick, G. L., and Silverman, S. E. (1991), "Visual Sensitivity to Motion: Age-Related Changes and Deficits in Senile Dementia of the Alzheimer Type," *Neurology*, Vol. 41, No. 9, pp. 1437–1440.

Tulving, E. (2002), "Episodic Memory: From Mind to Brain," *Annual Review of Psychology*, Vol. 53, pp. 1–25.

U.S. Bureau of the Census (1996), *Population Projections of the United States by Age, Sex, Race, and Hispanic Origin: 1995–2050*, Current Population Reports P25–1130, U.S. Government Printing Office, Washington, DC.

Verdonck-de Leeuw, I. M., and Mahieu, H. F. (2004), "Vocal Aging and the Impact on Daily Life: A Longitudinal Study," *Journal of Voice*, Vol. 18, No. 2, pp. 193–202.

Verhaeghen, P., Marcoen, A., and Goossens, L. (1993), "Facts and Fiction about Memory Aging: A Quantitative Integration of Research Findings," *Journals of Gerontology: Psychological Sciences*, Vol. 48, pp. P157–P171.

Voelcker-Rehage, C., Stronge, A. J., and Alberts, J. A. (2006), "Age-Related Differences in Working Memory and Force Control Under Dual-Task Conditions," *Aging, Neuropsychology, and Cognition*, Vol. 13, pp. 366–384.

Walker, N., Millians, J., and Worden, A. (1996), "Mouse Accelerations and Performance of Older Computer Users," in *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 151–154.

Walker, N., Philbin, D. A., and Fisk, A. D. (1997), "Age-Related Differences in Movement Control: Adjusting Submovement Structure to Optimize Performance," *Journal of Gerontology: Psychological Sciences*, Vol. 52B, pp. P40–P52.

Wickens, C. D. (1984), "Processing Resources in Attention," in *Varieties of Attention*, R. Parasuraman and D. R. Davies, Eds., Academic, Orlando, FL, pp. 63–102.

Wickens, C. D., and Seidler, K. S. (1997), "Information Access in a Dual-Task Context: Testing a Model of Optimal Strategy Selection," *Journal of Experimental Psychology: Applied*, Vol. 3, No. 3, pp. 196–215.

Wickens, C. D., Gordon, S., and Liu, Y. (1998), *Introduction to Human Factors Engineering*, 1st ed., Addison-Wesley-Longman, New York.

Wingfield, A., and Stine-Morrow, E. A. L. (2000), "Language and Speech," in *Handbook of Aging and Cognition*, 2nd ed., F. I. M. Craik and T. A. Salthouse, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 359–416.

Wishart, L. R., Lee, T. D., Murdoch, J. E., and Hodges, N. J. (2000), "Aging and Bimanual Coordination: Effects of Speed and Instructional Set on In-Phase and Anti-Phase Patterns," *Journals of Gerontology: Psychological Sciences*, Vol. 55B, pp. P85–P94.

Zacks, R. T., Hasher, L., and Li, K. Z. H. (2000), "Human Memory," in *Handbook of Aging and Cognition*, 2nd ed., T. A. Salthouse and F. I. M. Craik, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 293–357.

# CHAPTER 53

# DESIGNING FOR CHILDREN

**Valerie J. Berg Rice**
**General Ergonomics**
**Selma, Texas**

*There can be no keener revelation of a society's soul than the way in which it treats its children.*

  Nelson Mandela

## 1  INTRODUCTION: HOW DESIGNING FOR CHILDREN IS DIFFERENT FROM DESIGNING FOR ADULTS

Children are not merely diminutive adults. Their bodies and minds are still developing, and their physical and mental activities impact their development. The tools they use, the songs they sing, the games they play, and even the chairs they sit on can influence their growth. Those who design for children must consider how their designs will impact children's maturation, a consideration that is unnecessary when designing for adults.

Children lack the experience of an adult. They are less aware of dangers and are not fully conscious of the consequences of their own actions. They need to be protected by concerned, well-informed adults. The design of the products they use and the places where they work and play must incorporate protective measures, more so than typical when designing for adults.

Designing for children differs from designing for adults in context and content (Box 1). While safety is an important consideration during product design for adults, it receives greater emphasis when designing for children, as they are more vulnerable. Their vulnerability is a result of their cognitive abilities and understanding, their size and strength, and their being in the process of physical growth and development.

Annually worldwide nearly a million children die of injuries, while over 10 times that number receive care for nonfatal injuries [World Health Organization (WHO), 2008]. Many children are left with permanent disabilities. Younger children (under age 5) and boys are at greater risk (WHO, 2008). The most commonly occurring injuries result from traffic accidents, near drowning, burns, falls, and poisoning (WHO, 2008).

Children do not have to go far to encounter hazards, as the most frequently cited physical location at the time of injury is in or around their own homes (Brown and Beran, 2008). Most importantly, nearly all injuries are from accidents that could potentially be prevented through the thoughtful design of products, places, and processes and the design and implementation of educational, guidance, and supervisory programs. The responsibility for safe and useful designs falls not to any single group or profession but also to many, including families, local communities, and governments.

## 2  PRINCIPLES OF DESIGNING FOR CHILDREN

*You know the only people who are always sure about the proper way to raise children? Those who've never had any.*

  *Bill Cosby*

  Consider the Goal
  Consider the Target Audience

A similar sentiment may apply to those who design for children. While having children of one's own may not be necessary to design for children, certainly in-depth, first-hand knowledge and understanding of the target audience are requirements. Those who design for children must be intimately aware of how children think and act and possess a competent knowledge of children's capabilities and limitations at each age or stage of development. Having an expert in child development on the design team will enhance the application of design principles in accordance with the primary target audience—children (Lueder and Rice, 2008a).

Well-designed products, places, and processes fit the target audience who will use them and are effective, efficient, and safe (Kroemer and Grandjean, 2009).

**Figure 1** Two-year-old Maddie uses her mothers' decorative stickers in an unexpected way. (Photo courtesy of Carrie Vita.)



**Figure 2** Sophie explores a ''touch and feel'' book with her tongue, instead of her hands. Young children explore both with their mouths and their hands; both are rich in nerve endings, giving the child feedback as to the shape and consistency of the object. (Photo courtesy of Tim Vita.)

Yet, well-known principles of functional human factors design have slightly different implications and are applied differently when designing for children (Table 1). For example, when designing work processes for adults, the design process is often driven by the need to increase profits and decrease associated costs; thus the more effective and efficient the design, the better. The overriding management mission is to increase the company income. Safety, while important, historically has taken a "relative back seat," emerging when the cost of injuries and need for meeting standards are emphasized by management. This is not the case when designing for children. Product safety is considered paramount, especially by the secondary target audience and primary purchaser, the caregivers.

The parent or caregiver should have influence in designs for children. Their goals for the products, places, and processes their children use reflect their aspirations

**Table 1  Functional Design Principles and How They Differ When Designing for Children**

| Principle | How the Application Differs When Designing for Children |
|---|---|
| Fitting the Task to the Person | Not all strength measures are available for children of each age and gender.<br>Not all static and dynamic anthropometric measures are available for children of each age and gender.<br>Not all children develop at the same rate and change occurs rapidly.<br>Children of the same age and gender may vary significantly in their abilities.<br>Design for children:<br>• Focuses on more than the physical tasks and anthropometrics often considered for adult applications.<br>• Must take all types of development into account (physical, cognitive, emotional, social, speech/language, cultural, and the integration of each such as perceptual-motor, motor planning, etc.).<br>• Must simultaneously fit the current abilities of the child, while pulling them into the next phase of their development.<br>• Must ''fit''' a number of children at different ages and developmental stages.<br>• Must consider caregiver and parental concerns and goals for children, not just the child user. |
| Design for Efficacy | Finance: Effectiveness influences the financial ''bottom line'' in many designs for adults but does not similarly influence sales of products designed for children.<br>Importance: Effectiveness is essential in designs for adults but may be considered less important in designs for children.<br>Advocacy: Neither children nor caregivers may demand a product or process ''work'' exactly the way it should and when it should, while adults will actively advocate for effectiveness in products for adults.<br>Exploration: Some product design is purposely less than effective (less defined) to encourage imagination or multifunctional use by children.<br>Effectiveness assists with teaching children ''cause and effect.'' |
| Design for Efficiency | Finance: Efficiency influences the financial bottom line in many designs for adults but does not similarly influence sales of products designed for children.<br>Importance: Efficiency is extremely important when designing for adults but may be less so when designing for children.<br>Advocacy: Children are thought to have more available time and speed is less of an issue than exploration, problem solving, and imaginative use when designing for children.<br>Exploration: Some leeway in efficiency can lead children to explore alternate methods of use to discover effective, efficient, and fun techniques that are applicable in differing circumstances. This encourages creativity and problem solving. |
| Design for Safety | Safety is the priority focus in designing for children. This is especially the case if an injury link between product use and physical harm is readily apparent.<br>Safety can be ''down-played,'' if the injury link is less apparent. This is because adults may be unaware of injury potential or may believe symptoms among children are short-lived or will be outgrown. For example, many adults are not aware of the potential impact of overuse injuries among children.<br>Product safety is often taken for granted. Caregivers assume products for children are safe and have been safety tested before a product is offered for sale.<br>Parents tend to believe their own children are more mature and will make better decisions than other children of the same age (Harshman and Murphy, 2005). Thus, parents may not adhere to age-related recommendations for products or may encourage activities beyond their children's maturity level. |
| Intuitive Use and Ease-of-Use | Designs should be easy to use, but<br>• Children's physical size and attributes change rapidly.<br>• Children have little experience to draw upon.<br>• Children use items in unexpected ways.<br>• Cognitive development changes rapidly, for example, what is intuitive for an 8-year-old may not be intuitive for a 6-year-old or even for another 8-year-old with a different background or maturity level.<br>Young children explore a range of potential uses for an object.<br>A child's developmental skill levels differ in each type of development. For example, a child may be more physically developed than others their age yet less socially developed.<br>While a design should be easy to use, it is also necessary to offer supplemental challenges to help a child develop and hone their skills, especially with designing items such as playgrounds, toys, and learning products. |

for the children themselves. First and foremost, they want their children to be safe. They also want them to grow, develop, and mature, preferably at a rate similar to (or ahead of) their peers. They want their children to achieve, contribute, and have fun. Caregivers and parents want their children to "fit in," be happy, and at times simply to be entertained. They want the items their children interact with to:

- Be harmless
- Encourage and extend their developmental progress
- Permit their children to gain mastery over their environment
- Teach them important lessons necessary for their future
- Align with their own values and beliefs
- Bring them joy
- Keep them amused (Figure 3)

In addition, parents also recall their own childhood and these memories can influence their purchases. As noted by Walt Disney, "You are dead if you only aim for kids. Adults are grown up kids."

Just as caregiver's desires for designs[*] for children are multifaceted, the design process is also complex. Designing for children is not simple, as a design must be safe and fit children's current abilities, allowing them to experience success and accomplishment while simultaneously "pulling" them into the next stages



**Figure 3** Sometimes parents simply want their baby to be amused by the products (toys) they use. This baby plays happily while her mother shops for groceries. (Photo by Valerie Rice.)

---

[*] Design refers to the design of a product, place, or process.

of development. Designs must protect while encouraging a degree of risk. Designs for children should promote exploration even while they meet a child's expectations and "do what they are designed to do." The former fuels a child's thinking and investigation, while the latter helps a child learn about cause and effect.

A child's desires for designs of toys or products may be less complicated (compared with those of caretakers and business owners). Children want to have fun, for a design to reflect their taste in terms of color and preferred (often popular) images, and, in general, for an item to function as it should (Rice et al., 2008). They are less concerned with safety, learning, and their own development! For example, students in an elementary school noted their primary motives for selecting their backpack were the color and characters displayed on the pack (Rice et al., 2008).

Manufacturer's goals are aligned with the purpose and mission of their business, which is to show a financial profit. Therefore, they will often focus on making their design appealing to children, who then petition their caregiver (or another adult) to purchase the item (Box 2). Most also want their product to attract caregivers, who are the primary purchasers of children's products. Although the belief may not be in line with parental desires for their children, manufacturers often sell based on "age compression," that is, they believe the children of today mature at an earlier age than they did years ago. Therefore, the designs may try to capture the fun, but in a more sophisticated manner than years ago. Thus, the display of Barbie Dolls at FAO Swartz does not look much like a toy store; instead it has the appearance of a boutique (Fishel, 2001), and BRATZ Dolls wear clothing appropriate for an 18 year old but are popular with girls half that age.

Obviously, the principal user of children's products and environments are the children themselves. Other users are caregivers, family members, and friends. Caregivers can include parents, grandparents, nannies, baby-sitters, daycare providers, and teachers. Each of these individuals must keep the children in their care safe and must provide learning opportunities as well. The wise designer is aware of caregivers' goals and values. For example, toy designers may blend the old with the new in using toy designs that invoke nostalgia among caregivers (thus encouraging them to identify with and purchase a toy), while keeping the product fresh and in-line with children's current likes and dislikes.

### 2.0.1 What Does a Designer Need to Know?

While the item under design will dictate the core focus, in general designers should consider all aspects of child development. Without this broader consideration, oversights can easily ensue. For example, some playgrounds have signs describing them as accessible for children with disabilities. However, the single feature of accessibility may be a ramp with blocks a child can spin at the top platform. The designer did not seem to understand that a physical disability does not necessarily equate with a cognitive or social disability, nor does accessibility to children with disabilities merely mean being able to be "present" on a playground construed by the participant as being part of the "playing field." An accessible playground should have numerous ways for those who have disabilities, and those who do not have disabilities, to play together (Figures 4 and 5).

Another example demonstrating the importance of considering multiple perspectives of child development occurs with the design of clothing for toddlers. Many younger children cannot easily zip their jackets or button the small buttons on their shirts, as they are still developing their fine motor skills. Imagine the gains if a clothing manufacturer considered fine motor skills and psychosocial development, along with anthropometrics, in designing children's clothing. They could provide suitable clothing, offer early training to hone fine motor skills, and promote self-esteem through successful, independent dressing.



**Figure 4** Morgan's Wonderland contains playgrounds with a series of ramps as well as swings, cars, and a small train that can accommodate wheelchairs. On the playgrounds are games such as those shown here, where a child can move a person or car through a maze, learn Braille, make designs with sand or beads, or play music. (Photos by Valerie Rice.)



**Figure 5** At Morgan's Wonderland, children with disabilities can control water spouts, direct remote control boats, use water to shoot at targets, fish, and see themselves on a "weather channel." The entire park is accessible. (Photo by Valerie Rice.)

The provision of information on all aspects of child development is beyond the scope of this chapter (see Table 2). Entire books are dedicated to specific aspects of child development, such as motor skills (Liddle and Yorke, 2004) or language (Pinker, 1996). Other texts cover the full gamut of development, but may be limited to an age range (Brazelton and Sparrow, 2006; Berk, 2009). Frequently, texts contain developmental charts or tables listing and depicting milestones for skills and abilities of children. These describe the approximate time period during which the skills typically develop (Brown and Beran, 2008).[*]

Until recently, designers had to peruse a vast set of professional journals to obtain the most current information on designing for children. To address this issue, Lueder and Rice (2008a) compiled a unique, seminal text on designing for children. The text includes basic information on child development as well as specific chapters on product design, warnings, wayfinding, playgrounds, and designing for school and home environments, among others.

### 2.1 Does a Better Design Succeed in Preventing Injuries?

Protecting children through product, place, and process design can reduce injuries and even death. For example, the U.S. Consumer Product Safety Commission (CPSC) demonstrated that child-resistive packaging decreased prescription-related deaths by 45% (Rodgers, 1996), while a similar study on child-resistive packaging for aspirin reduced aspirin-related deaths by 34% (Rodgers, 1997, 2002). Numerous evidence-based studies have demonstrated other design-related success stories regarding injury and mortality prevention. Table 3

---

[*] Online developmental milestones are available in Lueder and Rice, (2008a) and at http://www.cdc.gov and http://www.dars .state.tx.us/ecis/resources/developmentmilestones.shtml

**Table 2 Child Development of Importance for Designers**

| Aspects of Development | Considerations |
|---|---|
| *Physical Development*<br>Body size (anthropometrics) and body composition; muscle, neuron, and skeletal growth; gross and fine motor skills; reflexes; vision, hearing, touch, taste and smell, complex skills such as visual–motor skills, etc. | Gender<br>Hormonal influences<br>Variations (heredity, environment, activities) |
| *Cognitive Development*<br>Neuropsychological, attention, concentration, planning, memory and memory retrieval, reasoning, problem solving, intelligence, etc. | Piaget's Cognitive Development Stages<br>Information processing perspectives<br>Brain-based learning<br>Metacognition |
| *Emotional Development*<br>Function, expressing, understanding, etc. | Interaction with cognition, health, and social behaviors<br>Self-regulation |
| *Social Development*<br>Self-awareness, self-concept, self-esteem, self-identify, relationships with others, moral development and reasoning, etc. | Vygotsky's Sociocultural Theories<br>Social problem solving<br>Coping skills, conflict resolution |
| *Speech and Language Development*<br>Phonics, semantics, grammar, sociolinguistics, etc. | Biological-Native Development<br>Behavioral learning<br>Interaction-based learning |
| *Cultural Understanding*<br>Social contexts — family, peers, teams, school, community, country, etc. | Gender<br>Context<br>Media<br>Groups |

**Table 3 Six Basic WHO Principles That Underlie Successful Child Injury Prevention Programs**

Legislation and regulations and their enforcement
Product modification
Environmental modification
Supportive home visits
Promotion of safety devices
Education and the teaching of skills

*Source*: WHO, 2008.

iterates basic WHO principles that underlie successful child injury prevention programs.

Even conveying risk and prevention information to caregivers and children can help, although it is not the sole answer to preventing injuries. In a literature review, Bass and colleagues (1993) found informing children and caregivers about potential injuries and injury prevention helped to reduce injuries. A high-quality, effective injury prevention program needs both a micro- and macroapproach and requires a variety of new and altered designs. Table 4 iterates WHO key strategies to reduce road traffic injuries and deaths among children, highlighting those that use or require design solutions.

Sweden has realized considerable success in reducing unintentional child injuries and deaths. Over the last 30 years, Sweden has reduced death rates due to injury from 24 to 5 (per 100,000) for boys and from 11 to 3 (per 100,000) for girls (WHO, 2008). Sweden has achieved these remarkable reductions using techniques such as (WHO, 2008):

- Environmental planning
  - Designing changes that divert traffic from residential areas and towns, permitting children to walk to school and play
  - Designing "safe communities"
- Home safety measures including home visits by health care professionals
- Product safety standards improvements
- School-based safety programs/measures
- Traffic safety measures including helmets and child restraint systems
- Water safety instruction

It also helps when success is expressed in terms of financial savings. For example, it is estimated that for every U.S. dollar spent on a child car seat, there is a savings of $29 in direct and indirect health care costs and other costs (WHO, 2008). Similar methods of reporting design and programmatic victories may assist with convincing governments to intervene on behalf of children.

Make safety a priority.

## 2.2 Special Design Considerations

### 2.2.1 Safety

Ensuring safety requires the designer to think about the probable results that may occur during reasonable,

**Table 4  WHO Key Strategies That Prevent Road Traffic Injuries among Children**

| Key Strategies | Effective | Insufficient Evidence | Ineffective | Harmful |
|---|---|---|---|---|
| Introducing and enforcing minimum legal drinking age laws | X | | | |
| *Setting (and enforcing) lower blood alcohol concentration limits for novice drivers and zero tolerance for offenders* | X | | | |
| ***Using appropriate child restraints and seatbelts*** | | | | |
| ***Wearing motorcycle and bicycle helmets*** | X | | | |
| *Forcing a reduction of speed around schools, residential areas, play areas* | X | | | |
| Separating different types of road user | X | | | |
| ***Introducing (and enforcing) daytime running lights for motorcycles*** | X | | | |
| Introducing graduated driver licensing systems | X | | | |
| *Implementing designated driver programs* | | X | | |
| ***Increasing the visibility of pedestrians*** | | X | | |
| *Introducing instruction in schools on the dangers of drunk driving* | | X | | |
| *Conducting school-based driver education* | | | X | |
| Putting babies or children on a seat with an air bag | | | | X |
| Licensing novice teenage drivers | | | | X |

*Source*: WHO, 2008.
NOTE: Bold italics designates obvious design needs, nonbold italics designates secondary design needs, such as newly designed signage, spaces, and programs/education.

normal use or misuse of a product. In the United States, the CPSC has detailed test methods to identify possible hazards, such as reasonably foreseeable misuse, damage, or abuse of the product, as well as established safety standards and guidelines [CPSC, 2010; 16 Code of Federal Regulations (CFR) §§ 1500.50–1500.53].

Determining what is "reasonable" for children as they explore their world using all of their senses of taste, touch, smell, hearing, and sight is not straightforward (Figure 6). It requires a thorough knowledge of child development, a "foreseeabilty conference" involving closely scrutinized observations of children using the product, and a team approach that involves both professionals and caregivers. In addition, there is no guarantee children will solely use those products designed for their age and maturity level. Parents can overestimate a child's maturity level and physical abilities (Schwebel and Bounds, 2003), younger siblings or playmates may pick up toys that do not belong to them, sales personnel may give out toys without checking for appropriate age ranges (such as fast food restaurants), and caregivers may fail to notice when a child is given a toy designed for an older child. Children also overestimate their own ability to perform physical tasks (Plumert, 1995; Schwebel and Bounds, 2003), and

Plumert (1995) found a relationship between six year olds overestimation of their abilities and their injuries.

It is important to consider safety when designing for children of all ages, not only younger children. Adolescents engage in risky behaviors (Steinberg, 2004; Vrendenburgh et al., 2010) and the highest unintentional injury rates occur among 15–18 year olds (Grossman, 2000). What is reasonable for an adolescent, may be less so for an adult and vice versa (Vrendenburgh et al., 2010). For example, the combined influence of wanting to look attractive and believing they are less vulnerable to injury may contribute to adolescent girls not using personal protective equipment (Vrendenburgh et al., 2010). Caregivers may warn their children about risk-taking behaviors, but they also underestimate the risk-taking behaviors of their children (Stanton et. al., 2000).

Removing a hazard through design is the best method of eliminating a hazard, thus removing the potential for human error (Kalsher and Wogalter, 2008; Rice and Lueder, 2008). The hierarchy for eliminating hazards includes (Table 5):

- Change the physical design so hazards are eliminated.

**Figure 6** Cole chooses ''pan play'' over play with his toys. Thus, he chooses the dull color of a pan over the primary colors of his playthings, the sound of a clanging pan over the melodies of his electronic toys or blocks, and the more-difficult-to-open cabinet over an easy-to-reach toy-basket. Such choices might seem counterintuitive, but children enjoy exploring and using objects in novel ways. (Photo by Jerry Duncan.)

- Add guards against hazards that cannot be eliminated through design.
- Provide warnings about hazards that cannot be eliminated through design and/or guarding.

### 2.2.2 Warnings

As noted above, warnings are necessary because not every hazard can be eliminated through design. Warnings for children's products target caregivers because caregivers purchase the majority of child products, supervise child activities, and guide and direct children's pursuits.

Caregivers are not likely to have access to the same information about a product that a manufacturer does, including the results of user testing, product construction (strength, component parts, etc.), and reports of misuse or breakage. Warnings assist manufacturers in alerting users about products, their proper use, breakage, and potential risks associated with improper or accidental misuse. Armed with such information, caregivers can make informed decisions about the products their children use, including the instructions they should

provide, how much supervision may be necessary, and whether personal protective equipment is needed and must be fit to the child.

The International Organization for Standardization (ISO) publishes safety label standards. The ISO standards deemphasize the use of text and encourage the use of pictorials as a means of broadening warning applications beyond the language of the text. The American National Standards Institute (ANSI) Z535.4 provides guidelines for warnings in the United States. The ANSI standards require testing before substituting a pictorial for text and testing must reveal less than 5% misunderstanding of the critical message and 85% comprehension.

One use of warnings is in poison control. In the United States, warnings must alert potential buyers if packaging does not include a child-resistant feature (or tampering with the product might have occurred) [Food and Drug Administration (FDA), 2005; Federal Register, 1998; Poison Prevention Packaging Act, 1970]. The warnings must be conspicuous and prominent (FDA 2005; Federal Register, 1998; Poison Prevention Packaging Act, 1970). Bix and colleagues (2009) found that adults (those with and without children in the home) spent little time attending to child-resistant warnings; instead the most time was spent attending to the brand name. Study results also revealed little recall of warning information and that child-resistant and tampering warnings were less legible than other gaze zones such as the brand name, facts about the drug, and claims (Bix et al., 2009). This is important because having non-child-resistant packaging for over-the-counter medications may contribute to unintentional poisonings (CPSC, 2005).

Table 6 gives both suggestions for constructing warnings and considerations important when designing warnings for children. For a detailed chapter on warnings for children, see Kalsher and Wogalter (2008).

Warnings are also constructed specifically for children, such as "Mr. Yuk." Mr. Yuk uses line drawings to depict a face with eyes closed and tongue protruding on a green background. The poison control design appears on stickers, typically with phone numbers of a local or national poison control center. The Mr. Yuk pictorial is trademarked and protected by copyright. The Pittsburgh Poison Center of Children's Hospital developed Mr. Yuk and tested its effectiveness with children at daycare centers, yet subsequent testing has not found conclusive evidence to substantiate the findings (Fergusson et al., 1982; Vernberg et al., 1984). Other picture-based warnings have been developed and tested for children, although they have not gained the notoriety of Mr. Yuk (Kalsher and Wolgalter, 2008; Wolgalter and Laughery, 2005).

Although warnings can be present, legible, and easily understood, if users do not read them, the effect is nil. In a recent survey, Vrendenburgh and colleagues (2010) investigated adolescent risk-taking behavior and use of instructions and warnings. They found 27% of participating adolescents rarely or never read instructions, and 35% rarely or never read warnings.

**Table 5  Hierarchy to Eliminate Hazards and Child Ergonomic Examples**

| Hierarchy | Example |
|---|---|
| Design to Eliminate hazards | **Crib Rails**<br>**Purpose:** Prevent strangulation (a child's body can slip an opening while their head cannot).<br>**Design Guidelines:** No more than a $2\frac{3}{8}$-in. space can exist between slats, spindles, corner posts, and rods at any point. A rectangular block $2\frac{1}{2}$ in. x $3\frac{1}{4}$ in. x $3\frac{1}{4}$ in. inserted in any position shall not be allowed to pass through any space between contoured or irregular slats or spindles. See CPSC (2001) for additional information. |
| Guard against hazards that cannot be eliminated through design | **Crib Rail Cover**<br>**Purpose**:<br>• Primary — Prevent injury hazard incurred by a child using a crib rail as "teething tool" or accidently bumping teeth or head on the crib rail.<br>• Secondary — May include attachment for teething toys, making them readily available for children.<br><br>**Design:** Crib rail guard/cover. Covered batting pads that cover wood crib railings, protecting the child, should he or she use the crib rail during teething (potential wood splinters or paint/varnish ingestion) or fall against the hardwood railing. |
| Warn about hazards that cannot be eliminated through design and/or guarding | **Crib and Bed Sheets**<br>**Purpose:** To prevent suffocation or strangulation of babies who may become entangled in crib or bed sheets.<br>**Future Design Guidelines/Requirements for Fitted Crib Sheets**<br><br>• Warning labels stressing the importance of a secure fit of sheets on crib mattresses. Anticipated Warning:<br><br>⚠ **WARNING**<br>Prevent suffocation or entanglement. Never use crib sheet unless it fits securely on crib mattress.<br>• Improvement in industry standards regarding the fit of crib sheets on mattresses.<br>• Current precautions offered to caregivers (CPSC Safety Alert, Crib Sheets, CPSC Document #5137): *To prevent tragedies, parents and caretakers can take the following precautions to ensure a safer sleeping environment for their young children.*<br>• *Make sure crib sheets fit snugly on a crib mattress and overlap the mattress so it cannot be dislodged by pulling on the corner of the sheet.*<br>• *Never use an adult sheet on a crib mattress; it can come loose and present an entanglement hazard to young children.*<br>• *Place a baby on his or her back on a firm, tight-fitting mattress in a crib meeting current safety standards.*<br>• *Remove pillows, quilts, comforters, and sheepskins from the crib.* |

## 3  CONCLUSIONS

Each new invention brings further design challenges for human factors engineers. Videogaming brings questions of exposure to violence and opportunities for producing games that will incorporate our culture and values, while children learn and still have fun. Social media (along with gaming) brings fears of a new generation of young adults who have difficulty interacting socially when in the physical presence of another person. They also bring us the chance to provide innovative ways to integrate learning in a virtual world with actions in the physical world. Extended computer use has us questioning the physical abilities of our youth and voicing concern over the detrimental effects of a sedentary lifestyle, including childhood obesity. Yet, it also introduces the prospects of shared exercise with friends who are miles away—at the click of a button. In short, there are vast opportunities to improve current designs for children, and new prospects are always on the horizon.

Designing for children is different than designing for adults. It requires additional knowledge of child development for a design to "pull" a child into a new developmental stage, while allowing them to succeed with their current skills. Toys, games, and playthings must be safe, even while providing an element of risk.

Most importantly, designing for children can serve to protect children from accidental injury and death. Each of the major causes of child injury and death (fire, drowning, falls, and traffic-related accidents) can be lessened through a multidisciplinary approach that

**Table 6 Suggestions for Constructing Warning Labels and How They May Differ for Children**

| Suggestions for Constructing Warnings (for Everyone) | Additional Considerations for Warnings for Children. |
|---|---|
| Warnings should be:<br><br>• Easy to understand<br>• Written in simple, uncomplicated language<br>• Uncluttered (easy to see and quickly comprehend)<br>• Placed where they are easily seen<br>• Durable | In addition, testing should include:<br><br>• Detection of the warning (did they see it?)<br>• Perception of the warning (did they read it?)<br>• Comprehension of the warning (did they understand it?)<br><br>Additional suggested testing includes whether the warning:<br><br>• Is recognized at a later date/time<br>• Is memorable<br>• Impacts decisions and/or actions<br><br>User testing should include children of the age and abilities expected to use the product and whether they find the warning (Steward and Martin, 1994; MacKinnon et al., 1993):<br><br>• Believable<br>• Important<br><br>Warnings should be specific, rather than general (Federal Trade Commission, 1981). |
| In compiling written warnings, designers should consider readers'<br><br>• Primary language(s)<br>• Reading ability<br>• Visual acuity in differing conditions (rain, low light, etc.)<br>• Visual field (so the warning is noticed) | In compiling written warnings, designers should consider child readers' abilities to:<br><br>• Understand the concept(s) being conveyed<br>• Problem solve<br>• Link warnings with their own behaviors and activities<br>• Understand the consequences of their own behaviors and deeds<br><br>In compiling written warnings, designers should consider caregivers' tendencies to overestimate their child's abilities.<br>An uncluttered visual field is less distracting. Adolescents demonstrate greater attention to and recall of warnings on generic (plain) packaging, than those containing brand imagery (Beede and Lawson, 1992) |
| If a picture is used, it should<br><br>• Provide a quick understanding of the issue<br>• Reinforce the textual message | If a picture is used:<br><br>• Pictures are better for very young children (Kalsher and Wogalter, 2008).<br>• Picture warnings may need an adult caregiver to explain them to children (DeLoache, 1991).<br>• The effectiveness of picture warnings differs according to the child's age. |

[a]*Note*: The same suggestions for constructing warnings "for everyone" (shown on the left of this table) apply to warnings for children (shown on the right of this table).

includes design modifications to products, places, and processes/procedures. Even the process of educating our youth and teaching injury prevention skills can be enhanced through informed design, such as brain-based learning, where knowledge of human abilities and limitations informs design.

It has been said that if one chooses not to pay for prevention, one will eventually pay in another way (medical costs). In terms of designing for children, particularly for injury prevention, the current needs include:

• More trained practitioners to identify needs, design improvements, implement interventions, and assess the effectiveness of such interventions
• Greater funding for design-related injury prevention programs and research
• Annotation of the cost effectiveness of design solutions
• Shared definitions of injuries and methods for collecting injury data

- Epidemiological injury data on risks, causes, and interventions
- Establishment of child injury prevention strategies based on injury data

There is much to do in meeting the needs to design for children. There is no shortage of opportunity. There is only the challenge to unite, to progress, and to care for our children, whatever their circumstance or background.

## REFERENCES

Bass, J. L., Chstofell, K. K., Widome, M., Boyle, W., Scheidt, P., Stanwick, R., and Roberts, K. (1993, October), "Childhood Injury Prevention Counseling in Primary Care Settings: A Critical Review of the Literature," *Pediatrics*, Vol. 92, No. 4, pp. 544–550.

Beede, P. and Lawson, R. (1992), "The Effect of Plain Packages on the Perception of Cigarette Health Warnings," *Public Health*, Vol. 106, pp. 315–322.

Berk, L. (2009), *Child Development*, 8th ed., Pearson, NJ.

Bix, L., Bello, N. M., Auras, R., Ranger, J., and Lapinski, M.K. (2009), "Examining the Conspicuousness and Prominence of Two Required Warnings on OTC Pain Relievers," *Proceedings of the National Academy of Science*, Vol. 106, No. 16, pp. 6550–6555.

Boucher, B. H. (2008), "Physical Education Programs," in *Ergonomics for Children: Designing Products and Places for Toodlers to Teens*, R. Lueder and V. J. B. Rice, Eds., Taylor & Francis, New York.

Brazelton, T., and Sparrow, J. D. (2006), *Touchpoints—Birth to Three*, 2nd ed., Da Capo, New York.

Brown, T., and Beran, M. (2008), "Developmental Stages of Children," in *Ergonomics for Children: Designing Products and Places for Toodlers to Teens*, R. Lueder and V. J. B. Rice, Eds., Taylor & Francis, New York.

Consumer Product and Safety Commission (CPSC) (2001), "U.S. Consumer Product Safety Commission Office of Compliance Requirements1 for Full Size Baby Cribs," 16 C.F.R. Part 1508, available: http:/www.cpsc.gov.

Consumer Product and Safety Commission (CPSC) (2005), "Poison Prevention Packaging: A Guide for Healthcare Professionals, CPSC Brochure 384, available: http:/www.cpsc.gov.

Consumer Product and Safety Commission (CPSC) (2010), available: http://www.cpsc.gov/.

DeLoache, J.S. (1991), "Symbolic Functioning in Very Young Children: Understanding of Pictures and Models," *Child Development*, Vol. 62, No. 4, pp. 736–752.

Federal Register (1998), "Tamper-Evident Packaging Requirements for Over-the-Counter (OTC) Human Drug Products," Food and Drugs Chapter, Subchapter C, *Federal Register*, 63, 213, pp. 59463–59471.

Federal Trade Commission (FTC) (1981), "Staff Report on the Cigarette Advertising Investigation," FTC, Washington, DC.

Fergusson, D. M., Horwood, L. J., Beautrais, A. L., and Shannon, F. T. (1982), "A Controlled Field Trial of a Poisoning Prevention Method," *Pediatrics*, Vol. 69, pp. 515–520.

Fishel, C. (2001), *Designing for Children: Marking Design That Speaks to Kids*, Rockport, Gloucester, MA.

Food and Drug Administration (2005), "Guidance for Industry: Labeling OTC Human Drug Products Questions and Answers," available: http://fda-drugs.legalview.info/.

Grossman, D. C. (2000), "The History of Injury Control and the Epidemiology of Child and Adolescent Injuries," *The Future of Children*, Vol. 10, No. 1, pp. 23–52.

Harshman, W. C., and Murphy, D. J. (2005), "ATVs and Youth: Matching Children and Vehicles," Publication No. E-45, National Agriculture Safety Database.

Kalsher, M. J., and Wogalter, M. S. (2008), "Warnings: Hazard Control Methods for Caregivers and Children," in *Ergonomics for Children: Designing Products and Places for Toodlers to Teens*, R. Lueder and V. J. B. Rice, Eds., Taylor & Francis, New York.

Kroemer, K. H. E., and Grandjean, E. (2009), *Fitting the Task to the Human: A Textbook of Occupational Ergonomics*, Taylor & Francis, New York.

Liddle, T. L., and Yorke, L. (2004), *Why Motor Skills Matter*, McGraw Hill, New York.

Lueder, R., and Rice, V. J. B., Eds. (2008a), *Ergonomics for Children: Designing Products and Places for Toodlers to Teens*, Taylor & Francis, New York.

Lueder, R., and Rice, V. J. B. (2008b), "Physical Development in Children and Adolescents and Age-Related Risks," in *Ergonomics for Children: Designing Products and Places for Toodlers to Teens*, R. Lueder and V. J. B. Rice, Eds., Taylor & Francis, New York.

MacKinnon, D., Pentz, M., and Stacey, A. (1993), "The Alcohol Warning Label and Adolescents," *American Journal of Public Health*, Vol. 83, pp. 585–587.

Peden, M., Oyegbite, K., Ozanne-Smith, J., Hyder, A. A., Branche, C., Rahman A. K. M. F., Rivara, F., and Bartholomeos, K., Eds. (2008), "World Report on Child Injury Prevention," World Health Organization (WHO) and UNICEF, WHO, Geneva.

Pinker, S. (1996), *Language Learnability and Language Development*, Harvard University Press, Harvard, MA.

Plumert, J.M. (1995), "Relations between Children's Overestimation of Their Physical Abilities and Accident Proneness," *Developmental Psychology*, Vol. 31. No. 5, pp. 866–876.

Rice, V., and Lueder, R. (2008), "Children and Injuries," in *Ergonomics for Children: Designing Products and Places for Toodlers to Teens*, R. Lueder and V. J. B. Rice, Eds., Taylor & Francis, New York.

Rice, V. J. B., DeVilbiss, C., and Bazley, C. (2008), "An Educational Exercise on Backpacks for School Children: Including Children, Faculty, and Parents," in *Applied Ergonomics International Conference Proceedings*, Ceasars Palace, Las Vegas, NV.

Rodgers, G. (1996), "The Safety Effects of Child-Resistant Packaging for Oral Prescription Drugs, Two Decades of Experience," *Journal of the American Medical Association*, Vol. 275, pp. 1661–166.

Rodgers, G. (1997), "Let's Welcome a New Generation of Child-Resistant Packaging," *Contemporary Pediatrics*, pp. 57–72.

Rodgers, G. (2002), "The Effectiveness of Child-Resistant Packaging for Aspirin," *Archives of Pediatric and Adolescent Medicine*, Vol. 156, pp. 929–933.

Schwebel, D. C., and Bounds, M. L. (2003), "The Role of Parents and Temperament on Children's Estimation of Physical Ability: Links to Unintentional Injury Prevention," *Journal of Pediatric Psychology*, Vol. 28, No. 7, pp. 505–516.

Shaffer, D. R., and Kipp, K. (2010), *Developmental Psychology: Childhood and Adolescence*, Wadsworth, CA.

Stanton, B. F., Li, X., Galbraith, J., Cornick, G., Feigelman, S., Kaljee, L., and Zhou, Y. (2000), "Parental Underestimates of Adolescent Risk Behavior: A Randomized, Controlled Trial of a Parental Monitoring Intervention," *Journal of Adolescent Health*, Vol. 26, No. 1, pp. 18–26.

Steinberg, L. (2004), "Risk Taking In Adolescence: What Changes And Why," *Adolescent Brain Development: Vulnerabilities and Opportunities*, Vol. 1021, pp. 51–58.

Steward, D., and Martin, I. (1994), "Intended and Unintended Consequences of Warning Messages: A Review and Synthesis of Empirical Research," *Journal of Health Policy Marketing*, Vol. 13, No. 1, pp. 1–19.

Vernberg, K., Culver-Dickinson, P., and Spyker, D. A. (1984), "The Deterrent Effect Of Poison-Warning Stickers," *American Journal of Diseases of Children*, Vol. 138, pp. 1018–1020.

Vrendenburgh, M. J., Zackowitz, I. B., Spencer, D. R., DeTaboada, M. R., and Vrendenburgh, A. G. (2010), "What Constitutes Typical Adolescent Behavior and How Does It Differ from Adult Conduct?" in *Applied Human Factors and Ergonomics Conference Proceedings*, Applied Human Factors and Ergonomics, Louisville, KY.

Wogalter, M. S., and Laughery, K. R. (2005), "Warnings," in *Handbook of Human Factors/Ergonomics*, 3rd ed., G. Salvendy, Ed., Wiley, Hoboken, NJ.

# CHAPTER 54

# DESIGN FOR ALL: COMPUTER-ASSISTED DESIGN OF USER INTERFACE ADAPTATION

**Constantine Stephanidis**
Institute of Computer Science Foundation for Research and Technology—Hellas (FORTH)
Heraklion, Crete, Greece
Department of Computer Science
University of Crete
Crete, Greece


**Margherita Antona**
Institute of Computer Science Foundation for Research and Technology—Hellas (FORTH)
Heraklion, Crete, Greece


**Anthony Savidis**
Institute of Computer Science Foundation for Research and Technology—Hellas (FORTH)
Heraklion, Crete, Greece
Department of Computer Science
University of Crete
Crete, Greece


**Nikolaos Partarakis, Konstantina Doulgeraki, and Asterios Leonidis**
Institute of Computer Science Foundation for Research and Technology—Hellas (FORTH)
Heraklion, Crete, Greece

## 1  INTRODUCTION

The increased importance of user interface design methodologies, techniques, and tools in the context of the development and evolution of the information society has been widely recognized in the recent past in the light of the profound impact that interactive technologies are progressively acquiring on everybody's life and activities and of the difficulty in developing usable and attractive interactive services and products (e.g., Winograd, 2001). As the information society further develops, the issue of human–computer interaction (HCI) design becomes even more prominent when considering the notions of universal access (Stephanidis, 2001a) and universal usability (Shneiderman, 2000), aiming at the provision of access to anyone, from anywhere and at anytime, through a variety of computing platforms and devices to diverse products and services. Design for universal access in the information society has often been defined as design for diversity, based on the consideration of the several dimensions of diversity that emerge from the broad range of user characteristics, the changing nature of human activities, the variety of contexts of use, the increasing availability and diversification of information, the variety of knowledge sources and services, and the proliferation of diverse technological platforms that occur in the information society.

These issues imply an explicit design focus to systematically address diversity, as opposed to afterthoughts or ad hoc approaches, as well as an effort toward reconsidering and redefining the concept of design for all in the context of HCI (Stephanidis, 2001a). In the emerging information society, therefore, universal access becomes predominantly an issue of design, and the question arises of how it is possible to design systems that take into account diversity and satisfy the variety of implied requirements. Research work in the past two decades has highlighted a shift of perspective and reinterpretation of HCI design, in the context of universal access, from current artifact-oriented practices toward a deeper and multidisciplinary understanding of the diverse factors shaping interaction with technology, such as users' characteristics and requirements and contexts of use, and has proposed methods and techniques that enable to proactively take into account and appropriately address diversity in the design of interactive artifacts (Stephanidis, 2001a). In the framework of such efforts, the concept of design for all has been reinterpreted and redefined in the domain of HCI. One of the main concepts proposed in such a context as a solution for catering for the needs and requirements of a diverse user population in a variety of context of use is that of automatic user interface adaptation (Stephanidis, 2001b).

Despite the progress that has been made, however, the practice of designing for diversity remains difficult, due to intrinsic complexity of the task, the current limited expertise of designers and practitioners in designing interfaces capable of automatic adaptation, as well as the current limited availability of appropriate supporting tools.

The rationale behind this chapter is that the wider practice and adoption of an appropriate design method, supported through appropriate tools, have the potential to contribute to overcoming the above difficulties. Toward this end, this chapter, after highlighting the main issues involved in the effort of designing for diversity, briefly describes a design method, the unified user interface design method, which has been developed in recent years to facilitate the design of user interfaces with automatic adaptation behavior (Savidis and Stephanidis, 2009a). Subsequently, the chapter discusses a series of tools and components which have been developed and applied in various development projects. These tools are targeted to support the design and development of user interfaces capable of adaptation behavior and in particular the conduct and application of the unified user interface development approach. Over the years, these tools have demonstrated the technical feasibility of the approach and have contributed to reducing the practice gap between traditional user interface design and design for adaptation. They have been applied in a number of pilot applications and case studies.

## 2  DESIGN FOR ALL: OVERVIEW OF APPROACHES, METHODS, AND TECHNIQUES

Universal access implies the accessibility and usability of information society technologies by anyone, anywhere, anytime, with the aim to enable equitable access and active participation of potentially all citizens in existing and emerging computer-mediated human activities by developing universally accessible and usable products and services which are capable of accommodating individual user requirements in different contexts of use and independently of location, target machine, or run time environment. The origins of the concept of universal access are to be identified in early approaches to accessibility, mainly targeted toward providing access to computer-based applications by users with disabilities.

Subsequently, accessibility-related methods and techniques have been generalized and extended toward more generic and inclusive approaches. HCI design approaches targeted to support universal access are often grouped under the term *design for all*.

### 2.1  Reactive versus Proactive Strategies

Accessibility in the context of HCI refers to the access by people with disabilities to information and communication technologies (ICT). Interaction with ICT may be affected in various ways by the user's permanent, temporary, or contextual individual abilities or functional limitations. For example, someone with limited seeing functions will not be able to use an interactive system which only provides visual output, while someone with limited bone or joint mobility or movement functions which affect the upper limbs will encounter difficulties in using an interactive system which only accepts input through the standard keyboard and mouse. Accessibility in the context of HCI aims to overcome such barriers by making the interaction experience of people with diverse functional or contextual limitations as near as possible to that of people without such limitations.

In traditional efforts to improve accessibility, the main direction followed has been to enable disabled users to access interactive applications originally developed for able-bodied users through appropriate adaptations.

Two main technical approaches to adaptation have been followed. The first is to treat each application separately and take all the necessary implementation steps to arrive at an alternative accessible version—product-level adaptation. Product-level adaptation practically often implies redevelopment from scratch. Due to the high costs associated with this strategy, it is considered the least favorable option for providing alternative access. The second alternative is to "intervene" at the level of the particular interactive application environment (e.g., MS-Windows) in order to provide appropriate software and hardware technology so as to make that environment alternatively accessible (environment-level adaptation). The latter option extends the scope of accessibility to cover potentially all applications running under the same interactive environment, rather than a single application.

The above approaches have given rise to several methods for addressing accessibility, including techniques for the configuration of input/output at the level of the user interface and the provision of assistive technologies. Popular assistive technologies include screen readers and Braille displays for blind users, screen magnifiers for users with low vision, alternative input and output devices for motor-impaired users (e.g., adapted keyboards, mouse emulators, joystick, binary switches), specialized browsers, and text prediction systems.

Despite progress, prevailing practices aimed at providing alternative access systems, either at the product or environment level, have been criticized for their essentially reactive nature (Emiliani, 2009). Although the reactive approach to accessibility may be the only viable solution in certain cases, it suffers from some serious shortcomings, especially when considering the radically changing technological environment and, in particular, the emerging information society technologies. The critique is grounded in two lines of argumentation. The first is that reactive solutions typically provide limited and low-quality access.

The second line of critique concerns the economic feasibility of the reactive approach to accessibility. Reactive approaches, based on a posteriori adaptations, though important to partially solve some of the accessibility problems of people with disabilities, are not viable in sectors of the industry characterized by rapid technological change. By the time a particular access problem has been addressed, technology has advanced to a point where the same or a similar problem reoccurs. The typical example that illustrates this state of affairs is the case of blind people's access to computers. Each generation of technology (e.g., DOS environment, windowing systems, and multimedia) caused a new "generation" of accessibility problems to blind users, addressed through dedicated techniques, such as text-to-speech translation for the DOS environment, off-screen models, and filtering for the windowing systems.

In some cases, adaptations may not be possible without loss of functionality. For example, in the early versions of windowing systems, it was impossible for the programmer to obtain access to certain window functions, such as window management. In subsequent versions, this shortcoming was addressed by the vendors of such products allowing certain adaptations on interaction objects on the screen.

Finally, adaptations are programming intensive and, therefore, are expensive and difficult to implement and maintain. Minor changes in product configuration, or the user interface, may require substantial resources to rebuild the accessibility features.

From the above, it becomes evident that the reactive paradigm to accessible products and services does not suffice to cope with the rapid technological change and the evolving human requirements. At the same time, the proliferation of interactive products and services in the information society, as well as of technological platforms and access devices, brought about the need to reconsider the issue of access under a proactive perspective, resulting in more generic solutions. This entails an effort to build access features into a product starting from its conception throughout the entire development life cycle. In the context of the emerging information society, therefore, universal access becomes predominantly an issue of design, and the question arises of how it is possible to design systems that permit systematic and cost-effective approaches to accommodating all users. Toward this end, the concept of design for all has been revisited in the context of HCI (Stephanidis et al., 1998, 1999).

In the context of universal access, design for all in the information society has been defined as a general framework catering for conscious and systematic efforts to proactively apply principles, methods, and tools in order to develop information society technology (IST) products and services that are accessible and usable by all citizens, thus avoiding the need for a posteriori adaptations or specialized design. Design for all, or universal design, is well known in several engineering disciplines, such as, for example, civil engineering and architecture, with many applications in interior design, building, and road construction. In the context of universal access, design for all either subsumes or is a synonym for terms such as *accessible design*, *inclusive design*, *barrier-free design*, and *universal design*, each highlighting different aspects of the concept. Through the years, the concept of design for all has assumed various connotations:

- Design of interactive products, services, and applications which are suitable for most of the potential users without any modifications. Related efforts mainly aim to formulate accessibility guidelines and standards in the context of international collaborative initiatives (see Section 2.2).
- Design of products which have standardized interfaces capable of being accessed by specialized user interaction devices (Zimmermann et al., 2002).

- Design of products which are easily adaptable to different users by incorporating adaptable or customizable user interfaces (Stephanidis, 2001b). This entails an effort to build access features into a product starting from its conception throughout the entire development life cycle.

## 2.2 Accessibility Guidelines and De Facto Standards

Guidelines play a key role in the adoption of Web accessibility and usability by industries and society. In essence, they constitute a rapidly evolving medium for transferring established and de facto knowledge (know-how) to various interested parties.

Concerning accessibility, a number of guidelines have been developed (Vanderheiden et al., 1996; Pernice and Nielsen, 2001). In particular, the Web Content Accessibility Guidelines (WCAG) [World Wide Web Consortium (W3C), 1999] explains how to make Web content accessible to people with disabilities. Web "content" generally refers to the information on a Web page or Web application, including text, images, forms, and sounds. WCAG 1.0 provides 14 guidelines that are general principles of accessible design. Each guideline has one or more checkpoints that explain how the guideline applies in a specific area. WCAG foresees three levels of compliance, A, AA, and AAA. Each level requires a stricter set of conformance guidelines, such as different versions of Hypertext Markup Language (HTML) (transitional vs. strict) and other techniques that need to be incorporated into code before accomplishing validation. In addition to WCAG 1.0, in December 2008, the W3C announced a new version of the guidelines, targeted to help Web designers and developers create sites that better meet the needs of users with disabilities and older users. Drawing on extensive experience and community feedback, WCAG 2.0 (W3C, 2008) improves upon WCAG 1.0 and applies to more advanced technologies.

In general, for a website to comply with accessibility guidelines, it should have at least the following characteristics:

- (X)HTML Validation from the W3C for the page content
- Cascading style sheet (CSS) validation from the W3C for the page layout
- At least WAI-AA (preferably AAA) compliance with the Web Accessibility Initiative (WAI) WCAG
- Compliance with all guidelines from Section 508 of the U.S. Rehabilitation Act
- Access keys built into the HTML
- Semantic Web Markup
- A high-contrast version of the site for individuals with low vision
- Alternative media for any multimedia used on the site (video, flash, audio, etc.).

The usage of guidelines is today the most widely adopted process by Web authors for creating accessible Web content. This approach has proven valuable for bridging a number of barriers faced today by people with disabilities.

Additionally, guidelines constitute de facto standards as well as the basis for legislation and regulation related to accessibility in many countries (Kemppainen et al., 2009).

For example, U.S. government Section 508 of the U.S. Rehabilitation Act (Rehabilitation Act Amendments, 1998) provides a comprehensive set of rules designed to help Web designers make their sites accessible.

Unfortunately, however, many limitations arise in the use of guidelines due to a number of reasons, including the difficulty in interpreting and applying guidelines, which require extensive training. Additionally, the process of using, or testing conformance to, widely accepted accessibility guidelines is complex and time consuming. To address this issue, several tools have been developed enabling the semiautomatic checking of HTML documents. Such tools make easier the development of accessible Web content, especially due to the fact that the checking of conformance does not rely solely on the expertise of developers. Developers with limited experience in Web accessibility can use such tools for evaluating Web content without the need to go through a large number of checklists.

As a final consideration, guidelines provide a "one-size-fits-all" approach to accessibility, which, while ensuring a basic level of accessibility for users with various types of disabilities, does not support personalization and improved interaction experience.

## 2.3 Design for All as User Interface Adaptation Design

In the light of the above, it appears that single artifact-oriented design approaches offer limited possibilities of addressing the requirements posed by universal access. A critical property of interactive artifacts becomes, therefore, their capability for automatic adaptation and personalization (Stephanidis, 2001b).

Methods and techniques for user interface adaptation meet significant success in modern interfaces. Some popular examples are the desktop adaptations in Microsoft Windows XP, offering, for example, the ability to hide or delete unused desktop items. Microsoft Windows Vista and Seven (7) also offer various personalization features of the desktop based on personal preferences of the user, by adding helpful animations, transparent glass menu bars, and live thumbnail previews of open programs and desktop gadgets (clocks, calendars, weather forecast, etc.). Similarly, Microsoft Office applications offer several customizations, such as toolbars positioning and showing/hiding recently used options. However, adaptations integrated into commercial systems need to be set manually and mainly focus on aesthetic preferences. In terms of accessibility and usability, for instance to people with disability or older people, only a limited number of adaptations are available, such as keyboard shortcuts, size and zoom options, changing color and sound settings, and automated tasks.

On the other hand, research efforts in the past two decades have elaborated more comprehensive and

systematic approaches to user interface adaptations in the context of universal access and design for all. The unified user interface methodology was conceived and applied (Savidis and Stephanidis, 2009a) as a vehicle to efficiently and effectively ensure, through an adaptation-based approach, the accessibility and usability of user interfaces (UIs) to users with diverse characteristics, supporting also technological platform independence, metaphor independence, and user-profile independence. In such a context, automatic UI adaptation seeks to minimize the need for a posteriori adaptations and deliver products that can be adapted for use by the widest possible end-user population (adaptable user interfaces). This implies the provision of alternative interface manifestations depending on the abilities, requirements, and preferences of the target user groups as well as the characteristics of the context of use (e.g., technological platform, physical environment). The main objective is to ensure that each end user is provided with the most appropriate interactive experience at run time.

The scope of design for diversity is broad and complex, since it involves issues pertaining to context-oriented design, diverse user requirements, as well as adaptable and adaptive interactive behaviors. This complexity arises from the numerous dimensions that are involved and the multiplicity of aspects in each dimension. In this context, designers should be prepared to cope with large design spaces to accommodate design constraints posed by diversity in the target user population and the emerging contexts of use in the information society. Therefore, designers need accessibility knowledge and expertise. Moreover, user adaptation must be carefully planned, designed, and accommodated into the life cycle of an interactive system, from the early exploratory phases of design through evaluation, implementation, and deployment.

Therefore, a need arises of providing computational tools which can support the design of user interface adaptation. In the past, the availability of tools was an indication of maturity of a sector and a critical factor for technological diffusion. As an example, graphical user interfaces became popular once tools for constructing them became available, either as libraries of reusable elements (e.g., toolkits) or as higher level systems (e.g., user interface builders and user interface management systems). As design methods and techniques for addressing diversity are anticipated to involve complex design processes and have a higher entrance barrier with respect to more traditional artifact-oriented methods, it is believed that the provision of appropriate design tools can contribute overcoming some of the difficulties that hinder the wider adoption of design methods and techniques appropriate for universal access, in terms of both quality and cost, by making the complex design process less resource demanding. The main objective in this respect is to offer tools which reduce the difference in practice between conventional user interface development and development for adaptation.

Finally, another prominent challenge in the context of universal access has been identified as the need for developing large-scale case study applications providing instruments for further experimentation and ultimately improving the empirical basis of the field by collecting knowledge on how design for diversity may be concretely practiced. Such case studies should aim to not only demonstrate technical feasibility but also assess the benefits of the overall approach as well the applied methods and tools.

## 3 UNIFIED USER INTERFACES

The *unified user interface development* methodology provides a complete technological solution for supporting universal access of interactive applications and services through a principled and systematic approach toward coping with diversity in the target user requirements, tasks, and environments of use (Savidis and Stephanidis, 2009a). A unified user interface comprises a single (unified) interface specification that exhibits the following properties:

1. It embeds representation schemes for user and usage context parameters and accesses user and usage context information resources (e.g., repositories, servers) to extract or update such information.

2. It is equipped with alternative implemented dialogue artifacts appropriately associated with different combinations of values for user and usage context–related parameters. The need for such alternative dialogue patterns is identified during the design process, when, given a particular design context, for differing user and usage context attribute values, alternative design artifacts are deemed as necessary to accomplish optimal interaction.

3. It embeds design logic and decision-making capabilities that support activating, at run time, the most appropriate dialogue patterns according to particular instances of user and usage context parameters and is capable of interaction monitoring to detect changes in parameters.

As a consequence, a unified user interface realizes:

- User-adapted behavior (user awareness), that is, the interface is capable of automatically selecting interaction patterns appropriate to the particular user.

- Usage context–adapted behavior (usage context awareness), that is, the interface is capable of automatically selecting interaction patterns appropriate to the particular physical and technological environment.

From a user perspective, a unified user interface can be considered as an interface tailored to personal attributes and to the particular context of use, while from the designer perspective it can be seen as an interface design populated with alternative designs, each alternative addressing specific user and usage context parameter values. Finally, in an engineering perspective, a unified user interface is a repository of implemented

dialogue artifacts, from which the most appropriate according to the specific task context are selected at run time by means of an adaptation logic supporting decision making.

At run time, the adaptations may be of two types:

a) Adaptations driven from initial user and context information known prior to the initiation of interaction

b) Adaptations driven by information acquired through context and interaction monitoring

The former behavior is referred to as *adaptability* (i.e., initial automatic adaptation) reflecting the interface's capability to automatically tailor itself initially to each individual end user in a particular context. The latter behavior is referred to as *adaptivity* (i.e., continuous automatic adaptation) and characterizes the interface's capability to cope with the dynamically changing or evolving user and context characteristics.

The concept of unified user interface is supported by a specifically developed architecture (Savidis and Stephanidis, 2009b). This architecture consists of independent communicating components, possibly implemented with different software methods and tools (see Figure 1). Briefly, a user interface capable of adaptation behavior includes (i) information regarding user and context characteristics (user and context profile), (ii) a decision-making logic, and (iii) alternative interaction widgets and dialogues.

The storage location, origin, and format of user-oriented information may vary. For example, information may be stored in profiles indexed by unique user identifiers, may be extracted from user-owned cards, may be entered by the user in an initial interaction session, or may be inferred by the system through continuous interaction monitoring and analysis. Additionally,

usage context information, for example, user location, environment noise, and network bandwidth, is normally provided by special-purpose equipment, such as sensors, or system-level software. In order to support optimal interface delivery for individual user and usage context attributes, it is required that for any given user task or group of user activities the implementations of the alternative best-fit interface components are appropriately encapsulated.

At design time, the design space is captured through a task hierarchy representation which allows for explicitly assigning alternative designs to node elements, called polymorphic task hierarchy (see Figure 2). Alternatives designs, call styles, can affect the syntactic level (i.e., alternative task decompositions) or the lexical level (i.e., alternative (i.e., alternative physical designs, such as layout appearances and widgets). Adaptation relations are established among alternative design styles for each node in the hierarchy. These relations define the run time adaptation behavior of the user interface, thus providing the adaptation decision-making logic. They include exclusion (two styles are never active at the same time), compatibility (two styles may be active at the same time), substitution (one style is deactivated and the second one is activated), and augmentation (the second style is activated, keeping the first active).

Upon start-up and during run time, the software interface relies on the particular user and context profiles to assemble the user interface on the fly, collecting and gluing together the constituent interface components required for the particular end user and usage context. In this context, run time adaptation-oriented decision making is engaged, so as to select the most appropriate interface components for the particular user and context profiles, for each distinct part of the user interface. The role of the decision making in UI adaptation is to effectively drive the interface assembly process
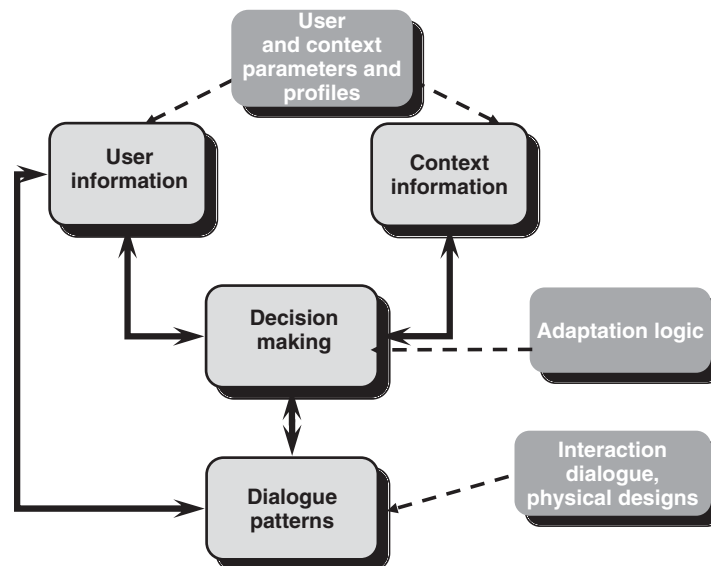


**Figure 1**   Unified user interface architecture.
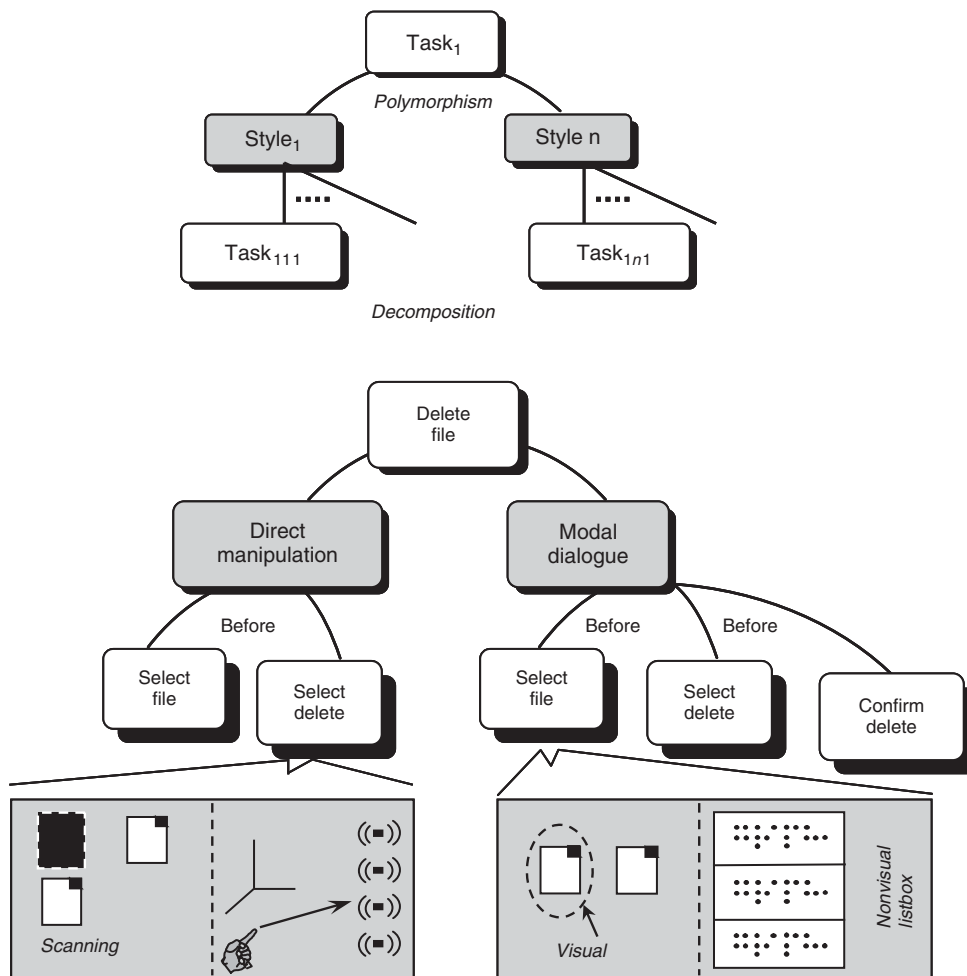
**Figure 2**  Example of polymorphic task hierarchy.

by deciding which interface components need to be selectively activated. The interface assembly process has inherent software engineering implications on the software organization model of interface components. For any component (i.e., part of the interface to support a user activity or task) alternative implemented incarnations may need to coexist, conditionally activated during run time due to decision making. In other words, there is a need to organize interface components around their particular task contexts, enabling them to be supported in different ways depending on user and context parameters. This contrasts with traditional nonadapted interfaces in which all components have singular implementations.

The unified user interface development method is not prescriptive regarding how each component is to be implemented (Savidis and Stephanidis, 2009c). For example, the alternative ways of representing user-oriented information and decision-making mechanisms may be employed. Also, the method does not affect the way designers will create the necessary alternative artifacts (e.g., through prototyping).

Since its beginning, the unified user interface development methodology has been accompanied by tools targeted to facilitate its employment. Early tools developed in this context are discussed in detail elsewhere (Stephanidis, 2001a). The next sections of this chapter focus on more recent tools which have been applied in a variety of case studies and have proved to contribute to a more effective and efficient application of the unified user interface concept, with particular focus on design.

## 4   TOOLS FOR DESIGN OF USER INTERFACE ADAPTATIONS

Tools developed in recent years to support user interface adaptation design include facilities for specifying

decision-making rules, adaptation design tools, adaptable widget toolkits for various interaction platforms, and user interface prototyping facilities.

## 4.1 Decision-Making Specification Language

The role of decision making in user interface adaptation is to effectively drive the interface assembly process by deciding which interface components need to be selectively activated. The Decision Making Specification Language (DMSL) (Savidis et al., 2005) is a rule-based language specifically designed and implemented for supporting the specification of adaptations. DMSL supports the effective implementation of decision making and has been purposefully elaborated to be easier for designers to directly assimilate and deploy, in comparison to programming-based approaches using logic-based or imperative-oriented programming languages.

In DMSL, the decision-making logic is defined in independent decision "if–then–else" blocks, each uniquely associated to a particular dialogue context. The individual end-user and usage context profiles are represented in the condition part of DMSL rules using an attribute values notation. Three types of design parameters values are allowed: (i) enumerated, that is, values belong to a list of (more than two) strings specified by the designer; (ii) boolean, that is, values True or False; and (iii) integer, which are specified by supplying minimum and maximum bounds of an integer range allowed as a value. Value ranges define the space of legal values for a given attribute. The language is equipped with three primitive statements: (a) dialogue, which initiates evaluation for the rule block corresponding to dialogue context value supplied; (b) activate, which triggers the activation of the specified component(s); and (c) cancel, which, similar to activate, triggers the cancellation of the specified component(s). These rules are compiled in a tabular representation that is executed at run time. Figure 3 provides an example DMSL rule. The representation engages simple expression evaluation trees for the conditional expressions.

The decision-making process is performed in independent sequential decision sessions, and each session is initiated by a request of the interface assembly module for execution of a particular initial decision block. In such a decision session, the evaluation of an arbitrary decision block may be performed, while the session completes once the computation exits from the initial decision block. The outcome of a decision session is a sequence of activation and cancellation commands, all of which are directly associated with the task context of the initial decision block. Those commands are posted back to the interface assembly module as the product of the performed decision-making session.

## 4.2 MENTOR Tool for User Interface Adaptation

The unified user interface design is recognized to require a higher initial effort and investment than traditional HCI design approaches, as it involves the identification of relevant design parameters, the design of alternative interface instances, and the delivery of an interface adaptation logic. MENTOR (Antona et al., 2006) is a support tool for the process of unified user interface design, which has been developed in order to address the following objectives:

- Provision of practical integrated support for all phases of unified user interface design by appropriately guiding the process and structuring the outcomes of creative design steps through appropriate editing facilities.
- Provision of practical support for a "smooth transition" from design to development of unified user interfaces through availability of automated verification mechanisms for the designed adaptation logic as well as the automated generation of "ready-to-implement" interface specifications, including the adaptation logic.
- Provision of support for reusing and extending (parts of) past design cases.

MENTOR targets the community of interface designers and do not assume deep knowledge of the unified user interface design method or particular HCI modeling techniques while, on the other hand, also supporting designers more experienced in adaptation design in effectively performing their work.

Figure 4 depicts the overall interactive environment of MENTOR, comprising four main editing environments:

- *Design Parameters Editor* (1 in Figure 4). The design parameters editor supports the encoding of design parameter attributes and related value spaces. These constitute the "vocabulary" for

---

| 1 | If | [Elderly user's age = 1 or 2 or 3] or [Elderly user 's life situation = 2 or 3] or [Elderly user's computer literacy level = 0] or [ Vision impairment = 1 or 2 or 3] |
| | Then | Resolution 640*480 pixels |

| 2 | If | [Elderly user's life situation = 1] or [Elderly user's computer literacy level = 1] |
| | Then | Resolution 800*600 pixels |

**Figure 3**  Example DMSL rule.
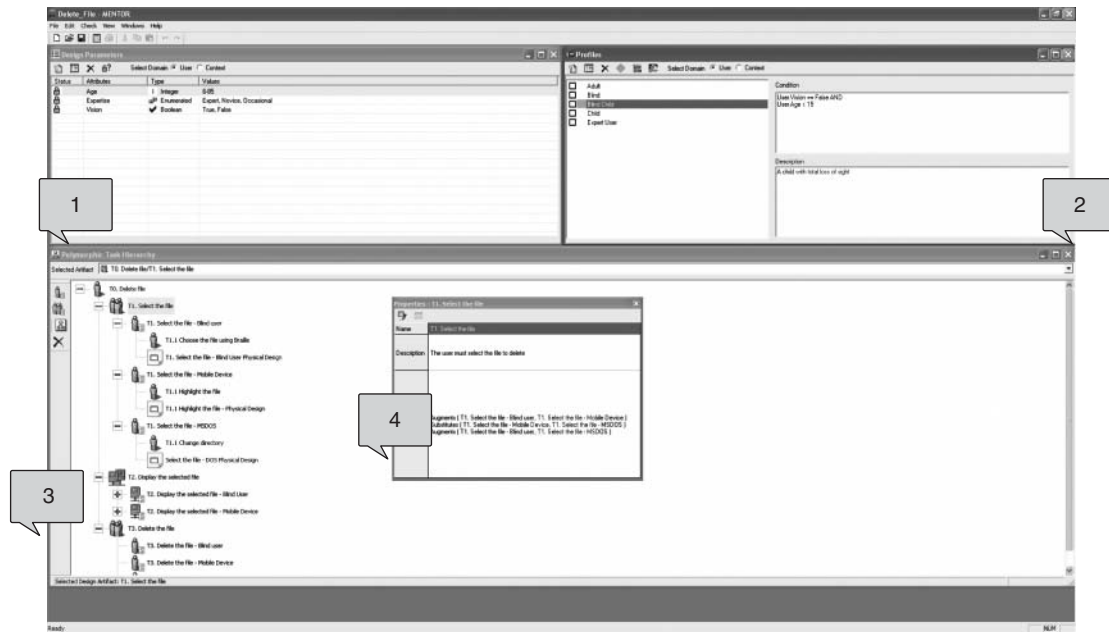
**Figure 4**  Overall MENTOR interactive environment: (1) design parameters editor; (2) profile editor; (3) polymorphic task hierarchy editor; (4) properties editor.

defining the "adaptation space" of the user interface under design. Parameters can belong to the user domain (i.e., user characteristics) or to the context domain (i.e., characteristics of the context of use and of the interactive platform). The editor also supports importing existing design parameters, applying the necessary consistency checking.

- *Profile Editor* (2 in Figure 4). User and context profiles can be defined by setting design parameter values in the profile editor. Existing profiles can be imported if consistent with the current design case.

- *Polymorphic Task Hierarchy Editor* (3 in Figure 4). The polymorphic task hierarchy editor allows designers to perform polymorphic task decomposition and encode the results in a hierarchy. The editor guides the decomposition process through decomposition steps.

- *Properties Editor* (4 in Figure 4). This editor allows assigning specific properties to the artifacts in the polymorphic hierarchy. Different categories of artifacts involve different properties. The most important piece of information to be attached to styles concerns the user and context parameter instantiations that define the style appropriateness at run time. Style conditions in MENTOR are formulated in the condition fragment of DMSL. For polymorphic artifacts, adaptation relations between children styles also need to be specified (selecting among incompatibility, compatibility, augmentation, and substitution).

Automated verification facilities for DMSL conditions are also included in MENTOR. These include the verification of the lexical and syntactic correctness as well as the verifiability of each DMSL expression separately. Additionally, hierarchical relations among styles in the polymorphic task hierarchy are also checked. MENTOR also supports verifying that the conditions on two styles related through a particular relation are compatible with the type of the relation. For example, if two styles are defined as incompatible, their conditions must not be consistent. These verification facilities ensure that the resulting run time adaptation logic is semantically sound and does not contain ambiguities which could cause problems when applying adaptations.

MENTOR also produces textual documentation of designs which can be used for several purposes, such as reviewing and evaluation, interface documentation, and, most importantly, implementation. The design report contains the project's design parameters and defined profiles, a textual representation of the polymorphic task hierarchy, the properties of each designed artifact, and the designed adaptation logic in the form of DMSL rules automatically produced by the tool. The DMSL rules produced by MENTOR can be directly embedded in the decision-making component of the designed user interface.

MENTOR has been validated in a number of design case studies, including the design of a unified user interface in the context of a health telematics scenario as well as the design of a shopping cart (Antona et al., 2006). These case studies have confirmed its overall usefulness and its advantages compared to "paper-based" adaptation design. The designers involved in

the case studies were able to rapidly acquire familiarity with the unified user interface design method and with the tool itself and expressed the opinion that the tool appropriately reflects and complements the method and significantly simplifies the conduct of polymorphic task decomposition. The verification facilities have also been found particularly effective in helping the designer to detect and correct inconsistencies or inaccuracies in the style conditions. Furthermore, the tool has been considered as particularly useful in providing the automatic generation of the DMSL adaptation logic, which, in the case of the shopping cart case study, has been directly integrated in the prototype implementation of the component.

### 4.3 Interaction Toolkits

User interface adaptation necessitates alternative versions of interaction artifacts to be created and coexist in the eventual design space. At the lexical level of interaction this can be achieved through software toolkits capable of dynamically delivering an interface instance that is lexically adapted to a specific user in a specific context of use. Such toolkits are essentially software libraries encompassing alternative versions of interaction elements and common dialogues, each version designed in order to address particular values of the user and usage context parameters. The run time adaptation-oriented selection of the most appropriate version, according to the end-user and usage context profiles, is the key element in supporting a wide range of alternative interactive incarnations. It should be noted that the presence and management of the alternative versions is fully transparent to toolkit clients. The latter provides the behavior of a smart toolkit capable of adaptively delivering its interaction elements so as to fit the current usage profile.

#### 4.3.1 EAGER

In order to support unified Web user interfaces, the combination of user-centered design, user interface prototyping, and design guidelines is applied together with unified user interface design. The proposed methodology (Partarakis et al., 2010a) is derived from the unified user interface software architecture and is instantiated in the EAGER software toolkit. In particular, EAGER integrates a design repository of:

- Alternative primitive UI elements with enriched attributes (e.g., buttons, links, radios)
- Alternative structural page elements (e.g., page templates, headers, footers, containers)
- Fundamental abstract interaction dialogues in multiple alternative styles (e.g., navigation, file uploaders, paging styles, text entry)

The EAGER Designs Repository is an extensible collection of implemented and ready-to-use alternative interaction elements which are organized around a polymorphic task hierarchy (Savidis and Stephanidis, 2009a). Each alternative element version, called a style following the terminology of Savidis and Stephanidis

(2009a), is purposefully designed to address the requirements of specific user and context parameter values. Alternative styles have been designed following typical user-centered design, user interface prototyping, and adoption of design guidelines. Additionally, EAGER design alternatives not only integrate current accessibility guidelines but also provide a suitable approach to personalized accessibility. In this respect, the EAGER Designs Repository can be viewed as encompassing consolidated adaptation design knowledge, thus greatly facilitating designers in the choice of suitable adaptations according to user-related or context-related parameters.

The Designs Repository component of EAGER provides the designs of alternative dialogue controls in the form of abstract interaction objects and task-level polymorphism. For each alternative version, the respective adaptation rationale is also recorded, including the profile parameters which are adaptively addressed.

An example is provided by images. Blind or low-vision users are interested not in viewing images but in reading the alternative text that describes the image. In order to facilitate blind and low-vision users, two design alternatives were produced, which are presented in Figure 5.

The text representation of the image presents not the image but only a label with the prefix "Image:" followed by the alternative text of the image. The second representation, targeted to users with visual impairments, is same as the first with the difference that, instead of a label, a link is included that leads to the specific image giving the ability of saving the image. In particular, a blind user may not wish to view an image but may wish to save it to a disk and use it properly. In addition to the above, another design was produced that can be selected as a preference by Web portal users in which the images are represented as a thumbnail bounding the size that holds on the Web page. A user who wishes to view the image in normal size may click on it. In Table 1, the design rationale of the alternative images design is presented, including its adaptation logic.

EAGER allows Microsoft® .NET developers to create interfaces that conform to the W3C accessibility guidelines (W3C, 1999) and which are able to adapt to the interaction modalities, metaphors, and user interface elements most appropriate to each individual user, according to profile information (user and context). The process of employing EAGER is significantly less demanding in terms of time, experience, and skills required from the developer than the typical process of developing Web interfaces for the "average" user. The benefits gained by using the EAGER toolkit lie on a number of dimensions, including:

- The time required for designing a Web application and the detail of design information needed
- The time required for designing the front end of the application to be used by end users
- The developer effort for setting up the application

Through EAGER, the complexity of the UI design effort is radically reduced due to the flexibility provided
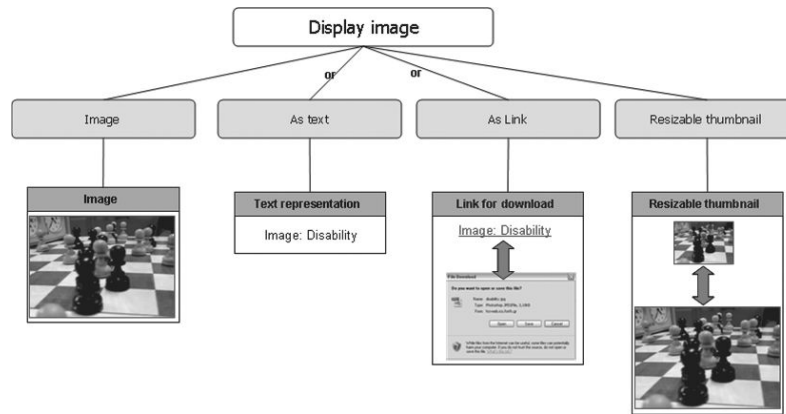
**Figure 5**   Alternative image hierarchy in EAGER. (From Partarakis et al., 2010a.)

**Table 1  Design Rationale of Alternative Image Styles (from 0)**

| | | Task: Display Image | | |
|---|---|---|---|---|
| Style | Image | As Text | As Link | Resizable Thumbnail |
| Targets | — | Facilitate screen reader and low-vision users in order not to be in difficulties with image viewing | Facilitate screen reader and low-vision users in order not to be in difficulties with image viewing but with the capability to save or view an image | Viewing images in small size in order not to hold large size on the Web page with the capability to enlarge the image to normal size when it is necessitated |
| Parameters | User (default) | User (blind or low vision) | User (blind or low vision) and user preference | User (preference) |
| Properties | View image | Read image alternative text | Read image alternative text and/or select link named as the image alternative text to save or view the image | View image thumbnail and select it to view it in normal size |
| Relationships | Exclusive | Exclusive | Exclusive | Exclusive |

by the toolkit for designing interfaces at an abstract task-oriented level. Therefore, designers are not required to be aware of the low-level details introduced in representing interaction elements; rather they should be aware of only the high-level structural representation of a task and its appropriate decomposition into subtasks, each of which represents a basic UI and system function.

On the other hand, the process of designing the actual front end of the application using a mark-up language is radically decreased in terms of time, due to the fact that developers initially have to select among a number of interface components each of which represents a far more complex facility. Additionally, developers do not have to spend time editing the presentation characteristics of the high-level interaction element due to the internal styling behavior.

The actual process of transforming the initial design into the final Web application using traditional UI controls introduces a lot of coding. However, when using EAGER, the amount of code required is significantly reduced due to the fact that the developer has the option to use a number of plug-and-play controls each of which

represents a complex user task. These controls are contained in the advance UI library of EAGER consisting of a total number of 55K pure code lines. Furthermore, the incorporation of EAGER's higher level elements make the code more usable, more readable, and especially safe due to the fact that each interaction component introduced is designed separately and developed and tested introducing a high level of code reuse, efficiency, and safety.

### 4.3.2   JMorph

The JMorph adaptable widget library (Leonidis et al., in press) is another example of a toolkit that inherently supports the adaptation of user interface components. It contains a set of adaptation-aware widgets designed to satisfy the needs of various target devices—Swing-based components for PC and Adaptive Window Toolkit- (AWT-) based components for Windows Mobile devices. Adaptation is completely transparent to developers, who can use the widgets as typical UI building blocks.

JMorph instantiates a common look and feel across the applications developed using it. The implemented

adaptations are meant to address the interaction needs of older users (Leuteritz et al., 2009) and follow specific guidelines which have been encoded into DMSL rules (Savidis et al., 2005). This approach is targeted to novice developers of adaptable user interfaces, as it relieves developers from the task of reimplementing or modifying their applications to integrate adaptation-related functionality.

The developed widgets are built in a modular way that facilitates their further evolution by offering the necessary mechanism to support new feature additions and modifications. Therefore, more experienced developers can use their own adaptation rules to modify the adaptation behavior of the interactive widgets.

The library's implementation using the Java programming language ensures the development of portable UIs that can run unmodified with the same look regardless of the underlying operating system (OS). Apart from OS independence, the proposed framework offers a solution that targets mobile devices running Windows Mobile.

The JMorph library provides the necessary mechanisms to support the alternative look and feel of either the entire environment (i.e., skins) or individual applications.

For that to be achieved, every widget initially follows the general rules to ensure that the common look and feel invariant will be met and then applies any additional presentation directives declared as "custom" look and feel rules. A custom rule could affect either an individual widget (e.g., the OK button that appears in the confirmation dialog of a specific application) or a group of widgets; therefore, entire applications can be fully skinned since their widgets inherently belong to a group defined by the application itself (e.g., all the buttons that belong to a specific application). The look-and-feel implementation of JMorph is based on the Synth technology (Sun Microsystems, 2010a).

Every adaptable widget in JMorph extends the relevant primitive Java component (i.e., AdaptableButton extends Java's Swing JButton) to provide its typical functionality, while the adaptation-related functionality is exposed via a straightforward application programming interface (API), the AdaptableWidget API. The API declares one main and two auxiliary methods: the adapt and the get/set function methods. Application developers can apply adaptation by simply calling the adapt method. The notion of the adapt method and the augmented set/get attribute methods has been originally proposed and implemented in the context of the PIM language-based generator of multiplatform adaptable toolkits (Savidis et al., 1997). This zero-argument method is the key method of the whole API, as it encapsulates the essential adaptation functionality and every adaptation-aware widget implements it accordingly. The global adaptation process includes first the evaluation of the respective DMSL rules that define the appropriate style and size and then their application through Synth's region matching mechanism.

For a local look and feel to be applied, the adapt method additionally utilizes the function getter method. The function attribute can be set manually by the designer/developer and is used on the one hand to decide whether and which transformations should be applied and on the other hand to define the group (i.e., all the buttons appear in the main navigation bar) or the exact widget (i.e., the OK button in a specific application) where they should be applied utilizing Synth's name-matching mechanism.

The adaptable widgets currently implemented in JMorph include label, button, check box, list, scrollbar, textbox, text area, drop-down menu, radio button, hyperlink, slider, spinner, progress bar, tabbed pane, menu bar, menu, menu item, and tooltip. Complex widgets such as date and time entry have also been developed. Adaptable widget attributes include background color/image, widget appearance and dimensions, text appearance, cursor's appearance on mouse over, highlighting of currently selected items or options, orientation options (vertical or horizontal), and explanatory tooltips.

Figure 6 shows some of the available widgets. Adaptable attributes for each widget are summarized in Table 2. All widgets in the library also include a text description which allows easy interoperation with speech-based interfaces, thus offering also the possibility to deploy a nonvisual instance of the developed interfaces.

## 4.4 Adaptive User Interface Prototyping

Popular user interface builders provide graphical environments for user interface prototyping, usually following a WYSIWYG ("what you see is what you get")



**Figure 6** Examples of adaptable widgets.

**Table 2  Adaptation Features of Widgets in Adaptable Widget Library**

| | |
|---|---|
| Buttons | Associated icon when idle, mouse over it, clicked or disabled |
| | Shortcut key |
| | Text |
| | Status (Enabled, Disabled) |
| | Tooltip's text and colors (foreground and background) |
| | Border |
| | Background and foreground color when clicked or idle |
| | Button's font |
| | Cursor appearance on mouse over i(e.g., hand cursor) |
| | Access key |
| | Vertical and horizontal text alignment |
| | Free space (gap) between button's icon and text |
| Check box | Associated icon when enabled and checked, enabled and unchecked, disabled and checked, or disabled and unchecked |
| | Shortcut key to check/uncheck the checkbox |
| | Text |
| | Status (Enabled, Disabled) |
| | Tooltip's text and colors (foreground and background) |
| | Border |
| | Background and foreground color |
| | Checkbox's font |
| | Cursor appearance on mouse over (e.g., hand cursor) |
| | Access key |
| | Vertical and horizontal text alignment |
| Drop-down menu | Background and foreground color of available and highlighted choices |
| | Choices' font |
| List | List orientation (vertical or horizontal) |
| | Background and foreground color of available and highlighted choices |
| | Choices' font |
| | Border around list component |
| | Tooltip text either one common for the list itself or a different one for each choice |
| | Cursor appearance on mouse over (e.g., hand cursor) |
| | Access key |
| Text box | Text |
| | Maximum number of characters per line |
| | Text's font |
| | Background color when this component is on or out of focus |
| | Border around text box |

**Table 2  (Continued)**

| | |
|---|---|
| | Foreground color of the text when either enabled or disabled |
| | Cursor appearance when user hovers mouse over it (e.g., hand cursor) |
| | Access key that facilitates traversal using keyboard (e.g., right arrow instead of tab) |
| | Status (Enabled, Disabled) |
| | Editable status, whether the user can alter the contents of this text box |
| | Tooltip's text and colors (foreground and background) |
| | Highlight text color when selected by mouse or due to search facility |
| Password text box | Text |
| | Maximum number of characters per line |
| | Text's font |
| | Background color when this component is on or out of focus |
| | Border around text box |
| | Foreground color of the text when either enabled or disabled |
| | Cursor appearance when user hovers mouse over it (e.g., hand cursor) |
| | Access key that facilitates traversal using keyboard (e.g., right arrow instead of tab) |
| | Status (Enabled, Disabled) |
| | Editable status, whether the user can alter the contents of this text box |
| | Tooltip's text and colors (foreground and background) |
| | Highlight text color when selected by mouse or due to search facility |
| Text area | Text |
| | Maximum number of characters per line |
| | Text's font |
| | Background (focused, not focused) |
| | Border around text box |
| | Foreground color of the text when either enabled or disabled |
| | Cursor appearance when user hovers mouse over it (e.g., hand cursor) |
| | Access key that facilitates traversal using keyboard (e.g., right arrow instead of tab) |
| | Status (Enabled, Disabled) |
| | Editable status, whether the user can alter the contents of this text box |
| | Tooltip's text and colors (foreground and background) |
| | Highlight text color when selected by mouse or due to search facility |
| | Maximum number of lines |

**Table 2** *(Continued)*

|  |  |
|---|---|
|  | Type of text wrapping when text area is not wide enough |
| Radio buttons | Text |
|  | Background and foreground color when selected or not |
|  | Border around radio button |
|  | Radio button's text font |
|  | Cursor appearance when user hovers mouse over it (e.g., hand cursor) |
|  | Status (Enabled, Disabled) |
|  | Tooltip's text and colors (foreground and background) |
|  | Foreground color when disabled |
|  | Associated icon when enabled and selected, enabled and unselected, disabled and selected, or disabled and unselected |
|  | Shortcut key to select the radio button |
|  | Foreground and background color when radio button is on or off focus |
| Hyperlink / label | Text |
|  | Foreground color when enabled |
|  | Background color (inherited by parent container) |
|  | Hyperlink's font |
|  | Tooltip's text and colors (foreground and background) |
|  | Cursor appearance when user hovers mouse over it (e.g., hand cursor) |
|  | Border around hyperlink |
|  | Status (Enabled, Disabled) |
| Table | Row height and margin between rows |
|  | Foreground and background color of currently selected cell |
|  | Show grid (horizontal, vertical lines) |
|  | Column width |
|  | Border around table cells |
|  | Background color of table cell |
|  | Tooltips' text and colors (foreground and background) |
|  | Background and foreground color of the table |
|  | Text's font |
| Slider | Slider's orientation (horizontal or vertical) |
|  | Minimum and maximum value that user could select using slider |
|  | Label for each discrete slider value (e.g., start — end, 0 — 100) |
|  | Visibility status of labels, major and minor ticks |
|  | Background color of slider component |
|  | Status (Enabled, Disabled) |
|  | Border around slider component |

**Table 2** *(Continued)*

|  |  |
|---|---|
|  | Tooltip's text and colors (foreground and background) |
|  | Cursor appearance when user hovers mouse over it (e.g., hand cursor) |
|  | Major and minor tick spacing (e.g., every fifth tick should be large — major-, while any other should be small — minor) |
|  | Foreground color of ticks (little vertical lines below slider) |
|  | Visibility status of the track |
|  | Invert start with end (e.g., on vertical orientation start is the top of the slider while end is the bottom) |
|  | Snap to ticks (limit user's selection only to ticks, e.g., when user slides cursor to 4.6, cursor should automatically be ''attracted'' to 5 |
| Spinner | Background and foreground color |
|  | Border around spinner |
|  | Cursor appearance when user hovers mouse over it (e.g., hand cursor) |
|  | Status (Enabled, Disabled) |
|  | Text's font |
|  | Tooltip's text and colors (foreground and background) |
| Menu bar | Background color |
| Menu | Background color (inherited from menu bar) |
|  | Foreground color |
|  | Text |
|  | Border around menu |
|  | Associated icon when menu is opened and closed or enabled and disabled |
| Menu item | Background and foreground color of each menu item |
|  | Associated icon when component is enabled or disabled or when user hover its mouse over it |
| Progress bar | Background and foreground color |
|  | Progress's text font |
|  | Border around bar |
|  | Status (Enabled, Disabled) |
|  | Tooltip's text and colors (foreground and background) |
|  | Minimum and maximum value of the bar |
|  | Progress bar's orientation (horizontal or vertical) |
|  | Progress's text visibility status |
| Tooltips | Background and foreground color |
|  | Text |
|  | Text's font |
|  | Border around tooltip |
|  | Cursor appearance when user hovers mouse over it (e.g., hand cursor) |
|  | Status (Enabled, Disabled) |

*(continued overleaf)*

**Table 2** *(Continued)*

| Tabs | Tab layout policy |
|---|---|
| | Tab placement |
| | Border around tab |
| | Tab's label font |
| | Cursor appearance on mouse over (e.g., hand cursor) |
| | Mnemonic (visible and functional per tab) |
| | Status (Enabled, Disabled) either for all tabs or for a specific one |
| | Tooltip's text and colors (foreground and background) |
| | Associated icon with each tab when enabled or disabled or when selected or unselected |
| | Tab's color when selected or not |

design paradigm. Available WYSIWYG editors offer graphical editing facilities that allow designers to perform rapid prototyping visually. Such editors may be stand alone or embedded in integrated environments (IDEs), that is, programming environments which allow developing application functionality for the created prototypes directly. Commonly used IDEs are Microsoft Visual Studio, NetBeans, and Eclipse. IDEs are very popular in application development because they greatly simplify the transition from design to implementation, thus speeding up considerably the entire process. However, currently available tools do not integrate adaptable widgets or provide any support for developing user interface adaptations. Therefore, prototyping alternative design solutions for different needs and requirements using prevalent prototyping tools may become a complex and difficult task if the number of alternatives to be produced is large and no specific support is provided for structuring and managing the design space. In order to facilitate the employment of the JMorph adaptable widget library described in the previous section toward rapid development of adaptable UIs, it has been integrated into the NetBeans graphical user interface (GUI) Builder (version 8.0, see Figure 7). The result is claimed to be the first and so far unique tool which supports rapid prototyping of adaptable user interfaces, with the possibility of immediately previewing adaptation results.

The choice of NetBeans was based on a thorough survey to identify the most suitable available IDE candidates to incorporate the adaptable widget library into their GUI Builder. NetBeans was preferred to Eclipse, which offers almost equivalent facilities, because it is better supported and more extensible, as its GUI Builder offers the essential mechanisms that facilitate the integration of custom widgets. The library's integration into the NetBeans built-in tool offers prototyping functionalities such as live "UI" preview as well as automatic application of specific sizing directives according to the OASIS styleguide. Moreover, NetBeans facilitates the implementation of the application's logic associated with the UI, thus offering not only a prototyping tool but

also a complete framework supporting the entire application development life cycle (design, development, and maintenance).

The NetBeans GUI Builder contains a palette that displays all the available widgets, initially only Java's built-in widgets, while the designers/developers, experienced or not, are familiar with its straightforward drag-and-drop functionality to add widgets on a "screen." The palette can only contain widgets that adhere to the JavaBean specification (Sun Microsystems, 2010b). JavaBeans are reusable software components for Java that can be manipulated visually in a builder tool. Practically, they are classes written in the Java programming language conforming to a particular convention.

They are used to encapsulate many objects into a single object (the bean), so that they can be passed around as a single bean object instead of as multiple individual objects. The integration of JMorph into NetBeans was achieved by implementing every AWT widget as a JavaBean.

To prototype a user interface, the designer will create the application's main window and will add the common containers (e.g., menu panels, status bar, header) by placing AdaptivePanels where appropriate. The necessary widgets (e.g., menu buttons, labels, text fields) will then be dragged from the palette and dropped into the design area of the builder.

To customize widgets, the typical process is to manually set the relevant attributes for each widget using the designer's "property sheets." To apply the same adjustment to other widgets, one can either copy/paste them or iteratively set them manually. In the adaptation-enabled process, using the function attribute, the process is slightly different. First, one needs to set the function attribute, then define the required style (e.g., colors, images, fonts), and finally define the rule (in a separate rule file) that maps the newly added style to the specific function. Whenever the same style should be applied, it is sufficient to simply set the function attribute respectively (CSS-like).

In some cases, more radical adaptations are required with respect to widget customization, as the same physical UI design cannot be applied as is.

In these cases, alternative dialogues can be designed by creating a container to host the different screens. The JMorph library offers the means to dynamically load different UI elements on demand, providing the functionality through adaptation rules and utilizing Java's reflection (introspection) capabilities.

The drag-and-drop selection and placement of widgets follow a conventional WYSIWYG approach. However, in the specific case, what you see is one instance of what you get, as all adaptation alternatives can be produced in the preview mode of the builder by simply setting some user-related variables (e.g., selecting a profile). During preview, a set of sizing rules are automatically applied to ease the design process. The obtained prototypes can easily be used for testing and evaluation purposes.

The result is a tool which offers the possibility of prototyping adaptable interfaces following standard practices without the need of designing customized
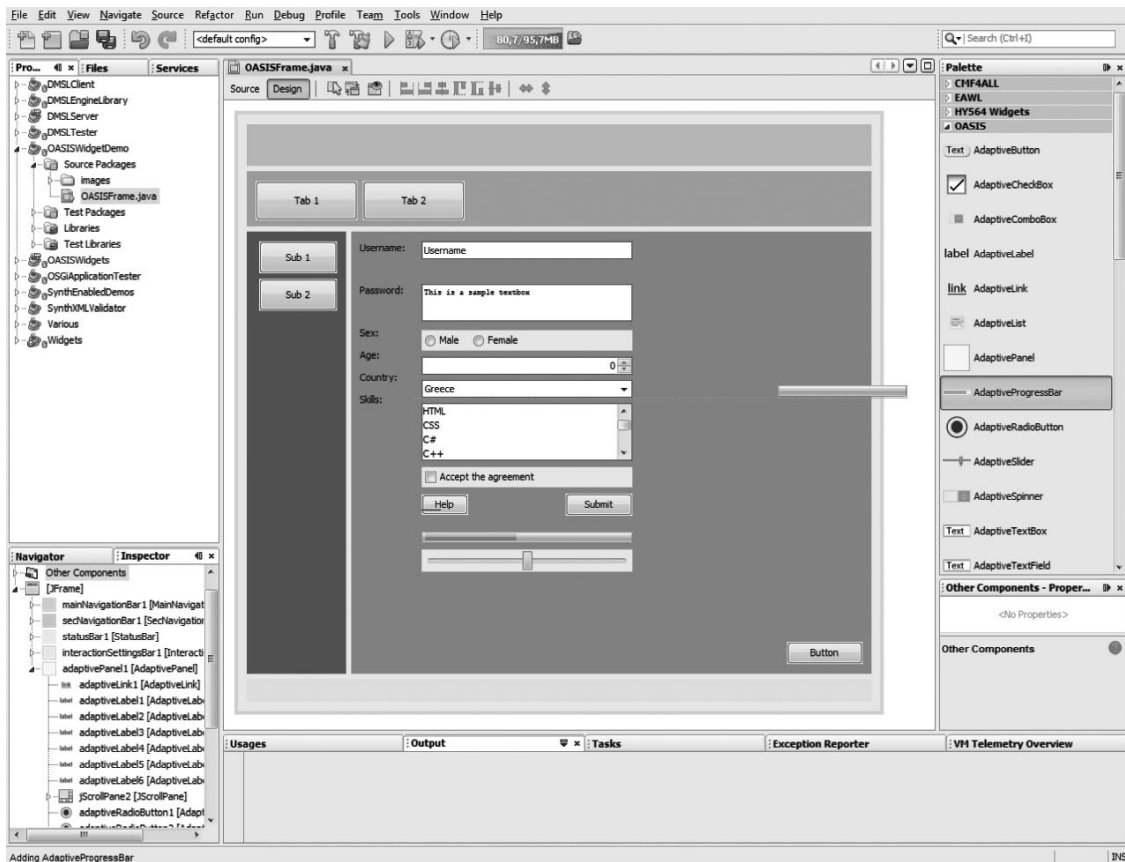
**Figure 7** Using the adaptable widget library integrated in the NetBeans IDE.

widget alternatives, which are included in the adaptable widgets, or to specify adaptation rules (which are predefined). Through the prototyping tool, it is also possible to preview how adaptations are applied for the defined user profiles. However, more expert designers can easily modify the DMSL adaptation rules, which are stored in a separate editable file, and experiment with new adaptations and varying look and feels. Finally, besides appearance adaptations, the overall approach allows to implement more complex forms of adaptation (e.g., dialogue adaptation) by prototyping alternative dialogues and introducing the respective adaptation logic.

## 5 PROTOTYPE APPLICATIONS AND CASE STUDIES

The methods and tools described in the previous sections have been employed over the years in the development of several prototype applications and services in various domains. These efforts demonstrate both the technical feasibility of the adaptation-based approach and the

progress achieved toward simplifying and improving development practices of user interface adaptation.

### 5.1 AVANTI and PALIO

The AVANTI universally accessible Web browser[*] and the PALIO tourist information system[†] (Stephanidis et al., 2011) constitute the first large applications applying the concepts and methods of unified user interfaces (see Section 3) as well as the first applications of the DMSL language for the implementation of the decision-making component in their architectures (see Section 4.1). While AVANTI constituted an adaptable and adaptive content-viewing application that can view any type of content, adapted or not, PALIO supported the creation of adaptable and adaptive content that can be viewed with any kind of browser. Thus, the two complemented each other.

---

[*] The AVANTI Web browser has been developed in the context of the ACTS AC042—AVANTI project (see Acknowledgments).

[†] The PALIO tourist information system has been developed in the context of the IST-1999-20656—PALIO project (see Acknowledgments).

The AVANTI browser provides an accessible and usable interface to a range of user categories, irrespective of physical abilities or technology expertise. Moreover, it supports various differing situations of use. The end-user groups targeted in AVANTI, in terms of physical abilities, include (i) "able-bodied" people, assumed to have full use of all their sensory and motor communication "channels"; (ii) blind people; and (iii) motor-impaired people, with different forms of impairments in their upper limbs, causing different degrees of difficulty in employing traditional computer input devices, such as a keyboard and/or a mouse. In particular, in the case of motor-impaired people, two coarse levels of impairment were taken into account: "light" motor impairments (i.e., users have limited use of their upper limps but can operate traditional input devices or equivalents with adequate support) and "severe" motor impairments (i.e., users cannot operate traditional input devices at all). Furthermore, since the AVANTI system was intended to be used both by professionals (e.g., travel agents) and by the general public (e.g., citizens, tourists), the users' experience in the use of, and interaction with, technology was another major parameter that was taken into account in the design of the user interface. Thus, in addition to the conventional requirement of supporting novice and experienced users of the system, two new requirements were put forward: (a) supporting users with any level of computer expertise and (b) supporting users with or without previous experience in the use of Web-based software.

In terms of usage context, the system was intended to be used both by individuals in their personal settings (e.g., home, office) and by the population at large through public information terminals (e.g., information kiosks at a railway station, airport). Furthermore, in the case of private use, the front end of AVANTI was intended to be appropriate for general Web browsing, allowing users to make use of the accessibility facilities beyond the context of a particular information system.

Users were also continuously supported as their communication and interaction requirements changed over time due to personal or environmental reasons (e.g., stress, tiredness, or system configuration). This entailed the capability, on the part of the system, to detect dynamic changes in the characteristics of the user and the context of use (either of temporary or of permanent nature) and cater for these changes by appropriately modifying itself.

The above requirements dictated the development of a new experimental front end which would not be based on existing Web browser technology or designed following traditional techniques oriented to the "typical user." In fact, the accessibility requirements posed by the user categories addressed in AVANTI could not be met either by existing customizability features supported by commercial Web browsers or through the use of third-party assistive products.

The unified user interface development approach was adopted to address the above requirements, as it provides appropriate methodologies and tools to facilitate the design and implementation of user interfaces that cater for the requirements of multiple, diverse end-user categories and usage contexts.

Following the unified user interface design method, the design of the user interface followed three main stages: (a) identification of different design alternatives to cater for the particular requirements of the users and the context of use; (b) integration of the designed alternatives into a polymorphic task hierarchy; and (c) development and documentation of the adaptation logic that drives the run time selection between the available alternatives.

AVANTI also constituted the first application of the DMLS language. Figure 7 shows an example adaptation decision block in the context of AVANTI. Such a decision block is targeted to selecting the best alternative interface components for the "link" task context. The interface design relating to this adaptation decision logic is provided in Figure 8.

Building on the results and findings of AVANTI, PALIO set out to address the issue of access to communitywide services by anyone, from anywhere, by proposing a hypermedia development framework supporting the creation of adaptive hypermedia systems. PALIO supported the provision of tourist services in an integrated, open structure while it constituted an extension of previous efforts, as it accommodated a broader perspective on adaptation and covered a wider range of interactive encounters beyond desktop access and advanced the current state of affairs by considering novel types of adaptation based on context and situation awareness. The PALIO framework was based on the concurrent adoption of the following concepts: (a) integration of different wireless and wired telecommunication technologies to offer services through both fixed terminals in public places and mobile personal terminals [e.g., mobile phones, personal digital assistants (PDAs), laptops]; (b) location awareness to allow the dynamic modification of information presented (according to user position); (c) adaptation of the contents to automatically provide different presentations depending on user requirements, needs, and preferences; (d) scalability of the information to different communication technologies and terminals; and (e) interoperability between different service providers in both the envisaged wireless network and the World Wide Web.

In the context of PALIO, DMSL has been effectively employed not only for user interface adaptation but also for adaptable information delivery over mobile devices to tourist users. The decision-making process was based on parameters such as nationality, age, location, interests or hobbies, time of day, visit history, and group information (i.e., family, friends, couple, colleagues, etc.). The information model reflected a typical relational database structure while content retrieval was carried out using Extensible Markup Language (XML)–based Structured Query Language (SQL) queries. In this context, in order to enable adapted information delivery, instead of implementing hard-coded SQL queries, query patterns have been designed, with specific polymorphic placeholders filled in by dynamically decided concrete subquery patterns. For instance, as seen in Figure 10, particular data categories or even query operations may be left

```
taskcontext link [
    evaluate linktargeting;
    evaluate linkselection;
    evaluate loadconfirmation;
]
taskcontext linktargeting [
    if (user.abilities.pointing == accurate) then
            activate "manual pointing";
    else
            activate "gravity pointing";
]
taskcontext linkselection [
    if (user.webknowledge in {good, nornal}) then
            activate "underlined text";
    else
            activate "push button";
]
taskcontext loadconfirmation [
    if (user.webknowledge in {low, none} or context..net==low) then
            activate "confirm dialogue";
    else
            activate "empty";
]
```

**Figure 8** DMSL decision block for adaptation of links in the AVANTI browser. (From Savidis et al., 2005.)



**S1**: Requires load confirmation. It is designed for:
1. Users with limited web experience;
2. Users that get tired and show high error rates during interaction;
3. Low-bandwidth networks.

**S2**: Link selection is done as far as the mouse cursor is inside the rectangular area of the link and the left mouse button is pressed. Designed for frequent and/or web-expert users.

**S3**: Link selection is done via typical GUI button press (i.e., press while cursor inside, release while cursor inside). In comparison to S2, it allows cancellation in the middle of the action (by releasing mouse button while cursor is outside).

**S4**: Gravity support for link targeting. If mouse cursor is inside gravity zone of the link, it is automatically positioned at the center of the link. Designed for users that cannot perform accurate mouse positioning.

**Figure 9** Link adaptation design in AVANTI. (From Savidis et al., 2005.)
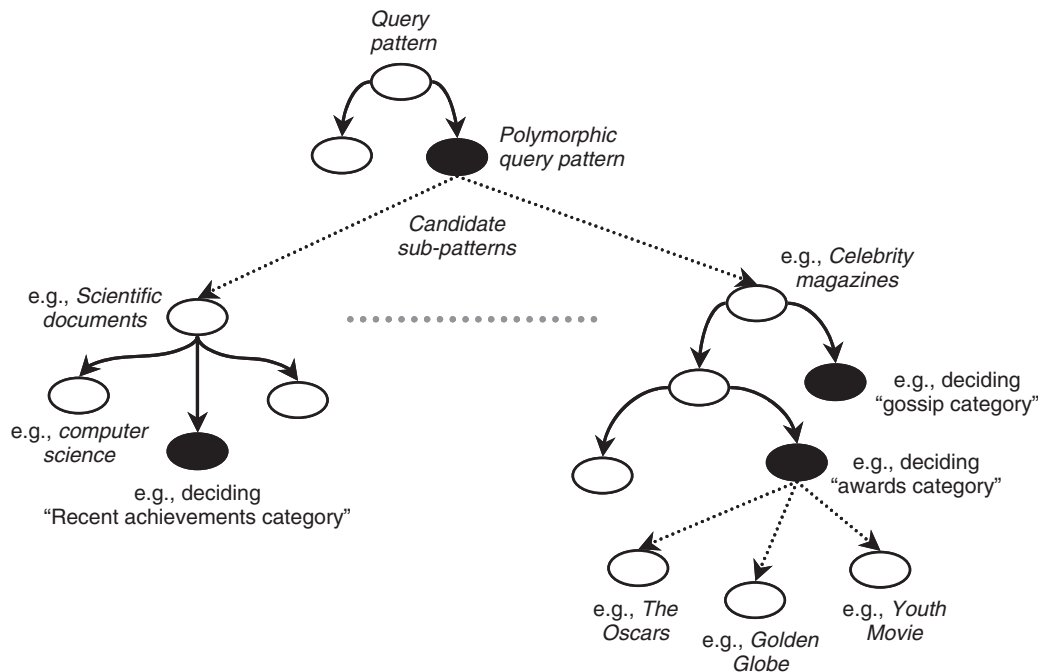
**Figure 10** Query adaptation using DMSL in PALIO. (From Savidis et al., 2005.)

"open," with multiple alternatives, depending on run time content adaptation decision making.

The experience gained in the development of AVANTI and PALIO has demonstrated the effectiveness of the use of adaptation-based methodologies, techniques, and tools toward the achievement of access to the World Wide Web by a wide range of user categories, irrespective of physical abilities or technology expertise, in a variety of contexts of use and through a variety of access devices, going far beyond previous approaches that rely on assistive or dedicated technologies.

Both AVANTI and PALIO make it possible to adopt a stepwise introduction of adaptation at different stages of development, thus enabling the progressive introduction of complex accessibility features and facilitating the incorporation of new user groups with distinct requirements in terms of accessibility.

### 5.2  EDeAN Web Portal

The portal of the European Design for All e-Accessibility Network (EDeAN)[*] (Partarakis et al., 2010b) was developed as a proof of concept by means of the EAGER toolkit (see Section 4.3.). As this was the redevelopment of an existing portal, it provided the opportunity to identify and compare the advantages of using EAGER, both at the developer's site, in terms of developer's performance, and at the end-user site, in terms of user experience improvement.

The new EDeAN portal disseminates information about the scope, objectives, and outcomes of the EDeAN networking activities. Through the portal public area (Figure 11) a number of facilities can be accessed, such as information about EDeAN, resources from a dedicated resource center, news and announcements, frequently asked questions, statistics regarding the networking activities, and surveys for collecting user feedback. The portal area for subscribed users is intended to support the actual networking activities and therefore provides a number of communication and collaboration facilities.

The users of the portal have the option to access the portal settings and alter them in order to match their personal characteristics and the characteristics of the context of use. A number of parameters can be set, such as language, device and display resolution, assistive technology, input device, disability, and Web familiarity. Additionally, to allow users to quickly alter their settings, the quick-settings option can be used, offering a number of predefined user profiles.

Adaptations can also take place based in interaction preferences. More specifically, interaction preferences settings can alter the interaction elements used for performing fundamental operations, such as browsing content and images or uploading files. The changes made to these settings are propagated to all portal modules. By manually altering these settings, the default adaptation logic that occurs based on the user basic setting is enriched.

Finally, adaptations can take place based on accessibility preferences. Custom accessibility includes all the

(silver-black)

(crystal-blue)

(cyan-blue)
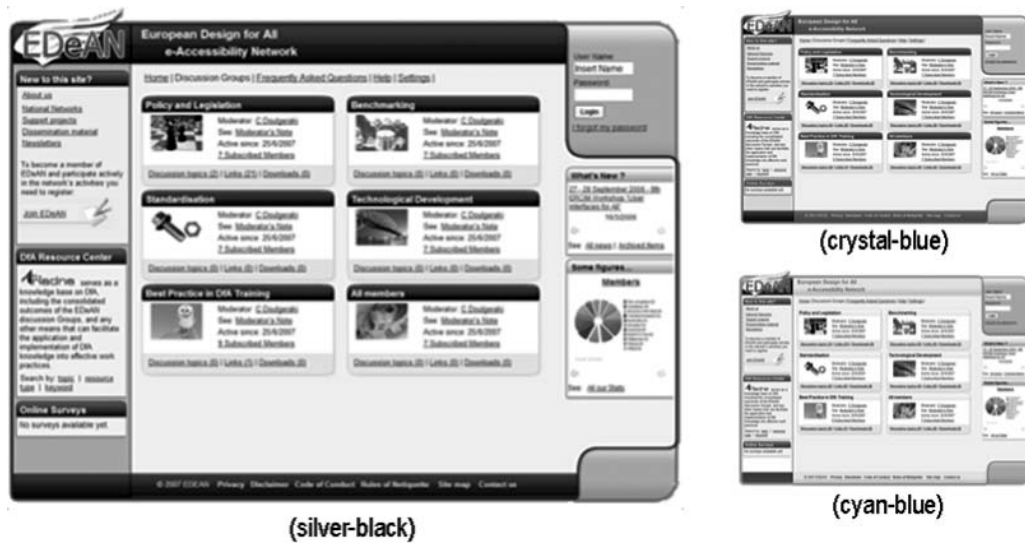
**Figure 11**   Main page of the EDeAN Web portal.



**Figure 12**   Example of widget adaptation in the ASK-IT Home Automation Application.

settings that can be altered to enhance the accessibility characteristics of the final user interface. Although each user interface is already compliant with the W3C accessibility guidelines, theses settings can further enhance the actual system accessibility and the perceived quality of interaction.

### 5.3   Applications of the JMorph Library

The JMorph adaptable widget library (see Section 4.3.2) has been used in the context of a number of development projects through which it is continuously refined and enhanced.

The first such example was the ASK-IT Home Automation Application,[*] which facilitates remote overview and control of home devices through the use of a

portable device. The user interface of the application has the ability to adapt itself according to user needs (vision and motor impairments), context of use (alternative display types and display devices), and presence of assistive technologies (alternative input devices).

Figure 12 presents an example of adaptation in the screen of the application which supports room selection on a PDA device. In the left part of the figure, the interface displays a color combination, while in the right part a grey scale is used for enhanced contrast.

The version of JMorph used to develop this application included simple graphics and adaptation rules and the widgets needed to be used programmatically.

Subsequently, the library has been enriched with more simple and complex widgets specifically designed for older users (Leuteritz et al., 2009). Currently, JMorph is being used in the development of the OASIS service suite (Bekiaris and Bonfiglio, 2009), comprising

---
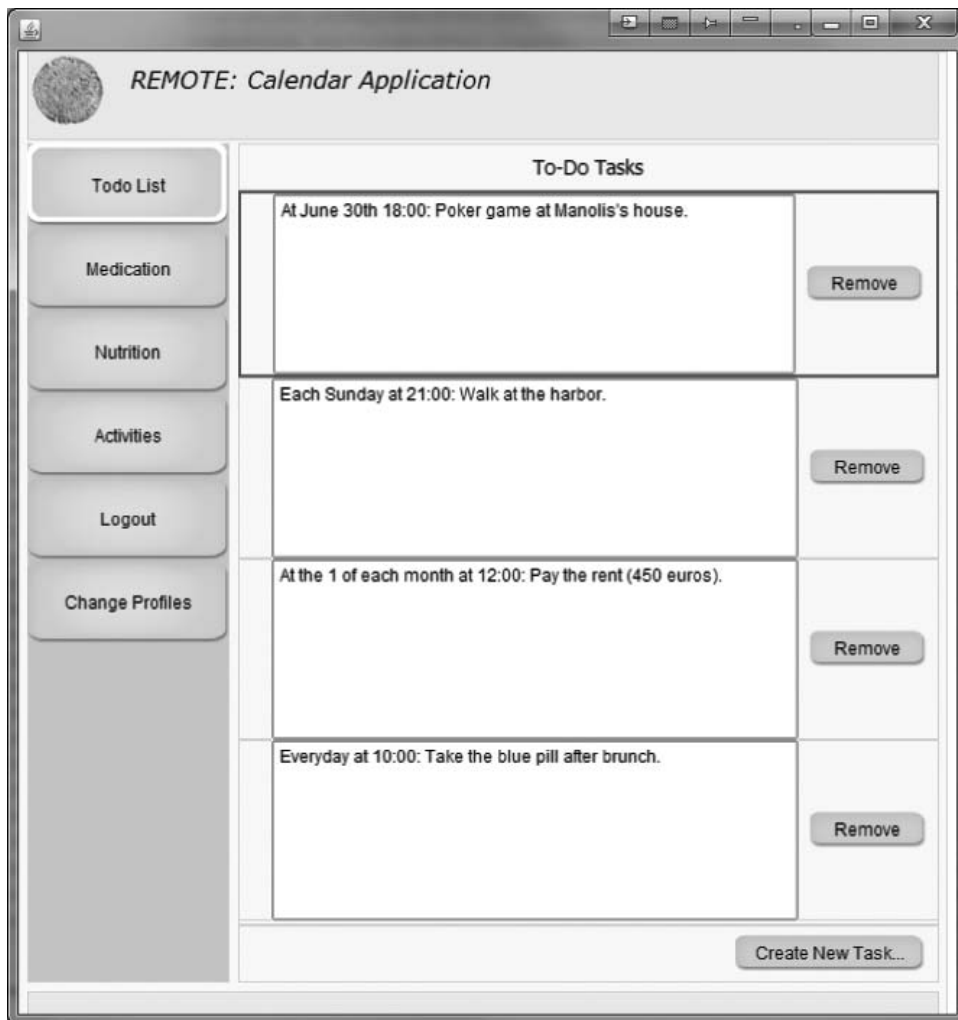
**Figure 13**   To-do list in the REMOTE calendar application.

12 services in three main domains addressing the quality of life of the elderly, namely Independent Living Applications, Autonomous Mobility, and Smart Workplaces Applications.[*] These applications are intended to be available through three different technological platforms, namely tablet PC, PDA, and mobile phone. One such application is a five-card poker game for older users which can be adapted to three different age and visual acuity profiles as well as to different levels of expertise in poker playing. In this context, the JMorph library has been distributed to a pool of universities, research institutions, and companies that are in charge of developing the applications.

Another application currently under development is the REMOTE calendar for older users,[†] offering functionalities such as to-do list, medication reminder, nutrition suggestions, daily activities schedule, and other notifications. Figure 13 presents a preliminary prototype of the calendar to-do list developed using JMorph in the NetBeans IDE.

## 6   CONCLUSIONS

Recent progress in the field of universal access and design for all, that is, access by anyone, anywhere, and

anytime to interactive products and services in the information society, has highlighted a shift of perspective and reinterpretation of HCI design, from current artifact-oriented practices toward a deeper and multidisciplinary understanding of the diverse factors shaping interaction with technology, such as users' characteristics and requirements and contexts of use, and has proposed solutions for methods, techniques, and codes of practice that enable to proactively take into account and appropriately address diversity in the design of interactive artifacts.

As a consequence, user interface design methodologies, techniques, and tools acquire increased importance in the context of universal access and strive toward approaches that support design for diversity based on the consideration of the several dimensions of diversity that emerge from the broad range of user characteristics, the changing nature of human activities, the variety of contexts of use, the increasing availability and diversification of information, the variety of knowledge sources and services, and the proliferation of diverse technological platforms that occur in the information society. Two main dimensions of such a perspective are its user-oriented focus, targeted toward capturing and collecting the requirements of a diversity of users in a diversity of usage context, and its adoption of intelligent interface adaptation as a technological basis, viewing design as the organization and structure of an entire design space of alternatives to cater for diverse requirements. In a universal access perspective, adaptation needs to be "designed into" the system rather than decided upon and implemented a posteriori.

Unified user interface design has been proposed in recent years as a method to support the design of user interfaces which automatically adapt to factors that impact on their accessibility and usability, such as the abilities and characteristics of different user groups, but also factors related to the context of use and the access technological platforms. Despite progress, however, the practice of designing for diversity remains difficult, due to the intrinsic complexity of the task and the current limited expertise of designers and practitioners. To overcome such a difficulty, tool support is required for supporting and facilitating adaptation design.

This chapter has discussed a series of tools and components developed over a period of more than a decade to support and facilitate the conduct of user interface adaptation design. These include:

- A language for the specification of adaptation decision making (DMSL, see Section 4.1)
- An interactive design environment for user interface adaptation (MENTOR, see Section 4.2)
- A toolkit supporting the development of adaptable Web-based user interfaces for the .NET platform (EAGER, see Section 4.3.1)
- A toolkit of platform-adaptable interaction widgets, implemented in Java, supporting the development of applications for PCs and mobile devices (JMorph, see Section 4.3.2)

- A prototyping solution for adaptable user interfaces within the NetBeans IDE (see Section 4.4)

Such tools are claimed to have a significant role in widening and improving the practice of design for all and ensuring a more effective transition from the design to the implementation phase. They have been used in practice in a series of case studies involving different types of applications for different purposes, contexts, and interaction platforms. These extensive case studies have demonstrated the technical feasibility of the overall adaptation-based approach to universal access. Additionally, these developments have provided hands-on experience toward improving the usefulness and effectiveness of the developed tools in different phases of the user interface development life cycle, in particular design. During these developments, it has progressively become clear that user interface adaptation can be adopted in practice as a result of reducing the gap with mainstream design practices. Ultimately, this amounts to providing transparent solutions which do not require specific adaptation knowledge and support prototyping. Therefore, more recent solutions have gone into the direction of providing ready-to-use widget toolkits that integrate all the required adaptation knowledge and logic as well as supporting the view of alternative designs in mainstream development environments. Obviously, however, such solutions, while achieving the objective of simplifying the design of adaptation as far as alternative widget instances are concerned, still require specialized knowledge and mastery of user interface adaptation mechanisms for designing dialogue adaptation at a syntactic or semantic level as well as for creating new or modifying existing adaptable widgets.

The tools discussed in this chapter have also proved their usefulness for educational purposes. In particular, DMSL, MENTOR, and, more recently, the JMorph library with its accompanying prototyping solution have been used in the context of an advanced HCI course at the University of Crete, with the objective of introducing postgraduate students to the basics of developing self-adapting user interfaces.

As the information society further develops, the issue of efficiently designing user interfaces capable of automatic adaptation behavior becomes even more prominent in the context of the next anticipated technological generation, that of ambient intelligence environments (see Chapter 49 xx). Ambient intelligence provides a vision of the information society where humans are surrounded by intelligent intuitive interfaces that are embedded in all kinds of objects and an environment that is capable of recognizing and responding to the presence of different individuals in a seamless, unobtrusive, and often invisible manner. Clearly, ambient intelligence environments are intrinsically based on adaptation, and user and context awareness, as well as adaptation decision making, become fundamental. Therefore, current research efforts are targeted toward providing tools and facilities to support user interface adaptation design in ambient intelligence environments.

## ACKNOWLEDGMENTS

## REFERENCES

Antona, M., Savidis, A., and Stephanidis, C. (2006), "A Process–Oriented Interactive Design Environment for Automatic User Interface Adaptation," International Journal of Human Computer Interaction, Vol. 20, No. 2, pp. 79–116.

Bekiaris, E., and Bonfiglio, S. (2009), "The OASIS Concept," in *Universal Access in Human-Computer Interaction, Proceedings of the 13th International Conference on Human-Computer Interaction* (*HCI International 2009*), Vol. 6, C. Stephanidis, Ed., Lecture Notes in Computer Science, Springer, Berlin and Heidelberg.

Emiliani, P. L. (2009), "Perspectives on Accessibility: From Assistive Technologies to Universal Access and Design for All," in *The Universal Access Handbook*, C. Stephanidis, Ed., Taylor & Francis, Boca Raton, FL, pp. 2-1–2-18.

Kemppainen, E., Kemp, J. D., and Yamada, H. (2009), "Policy and Legislation as a Framework of Accessibility," in *The Universal Access Handbook*, C. Stephanidis, Ed., Taylor & Francis, Boca Raton, FL, pp. 53-1–53-16).

Leonidis, A., Antona, M., and Stephanidis, C. (in press), *Rapid Prototyping of Adaptable User Interfaces*, iFirst http://www.tandfonline.com/doi/full/10.1080/10447318. 2011.581891

Leuteritz, J.-P., Widlroither, H., Mourouzis, A., Panou, M., Antona, M., and Leonidis, S. (2009), "Development of Open Platform Based Adaptive HCI Concepts for Elderly Users," in *Universal Access in Human-Computer Interaction—Intelligent and Ubiquitous Interaction Environments, Proceedings of the 13th International Conference on Human-Computer Interaction* (*HCI International 2009*), Vol. 6, C. Stephanidis, Ed., Lecture Notes in Computer Science, Vol. 5615, Springer, Berlin and Heidelberg, pp. 684–693.

Partarakis, N., Doulgeraki, C., Antona, M. and Stephanidis, C. (2010a), "Designing Web-Based Services," in *Introduction to Service Engineering*, G. Salvendy and W. Karwowski, Eds., Wiley, Hoboken, NJ, pp. 447–487.

Partarakis, N., Doulgeraki, C., Antona, M. and Stephanidis, C. (2010b), "The Development of Web-based Services," in *Introduction to Service Engineering*, G. Salvendy and W. Karwowski, Eds., Wiley, Hoboken, NJ, pp. 502–532.

Pernice, K., and Nielsen, J. (2001), ''Beyond ALT Text: Making the Web Easy to Use for Users with Disabilities," Nielsen Norman Group Report, available: http://www.nngroup.com/reports/accessibility.

Rehabilitation Act Amendments (1998), Section 508, available: http://www.section508.gov/.

Savidis, A., and Stephanidis, C. (2009a), "Unified Design for User Interface Adaptation," in *The Universal Access Handbook*, C. Stephanidis, Ed., Taylor & Francis, Boca Raton, FL, pp. 16-1–16-17.

Savidis, A., and Stephanidis, C. (2009b), "A Unified Software Architecture for User Interface Adaptation," in *The Universal Access Handbook*, C. Stephanidis, Ed., Taylor & Francis, Boca Raton, FL, pp. 21-1–21-17.

Savidis, A., and Stephanidis, C. (2009c), "Software Requirements for Inclusive User Interfaces," in *The Universal Access Handbook*, C. Stephanidis, Ed., Taylor & Francis, Boca Raton, FL, pp. 18-1–18-25.

Savidis, A., Stephanidis, C., and Akoumianakis, D. (1997), "Unifying Toolkit Programming Layers: A Multi-Purpose Toolkit Integration Module," in *Proceedings of the Fourth International Eurographics Workshop on Design, Specification and Verification of Interactive Systems* (*DSV-IS*), Springer, Berlin, pp. 177–192.

Savidis, A., Antona, M., and Stephanidis, C. (2005), "A Decision-Making Specification Language for Verifiable User-Interface Adaptation Logic," *International Journal of Software Engineering and Knowledge Engineering*, Vol. 15, No. 6, pp. 1063–1094.

Shneiderman, B., (2000, May), "Pushing Human-Computer Interaction Research to Empower Every Citizen: Universal Usability," *Communication of the ACM*, Vol. 43, No. 5, pp. 85–91.

Stephanidis, C. (2001a), "User Interfaces for All: New Perspectives into Human-Computer Interaction," in *User Interfaces for All: Concepts, Methods, and Tools*, C. Stephanidis, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 3–17.

Stephanidis, C. (2001b), "Adaptive Techniques for Universal Access," *User Modelling and User Adapted Interaction International Journal*, 10th Anniversary Issue, Vol. 11, Nos. 1/2, pp. 159–179.

Stephanidis C. (Ed.), Salvendy, G., Akoumianakis, D., Bevan, N., Brewer, J., Emiliani, P. L., Galetsas, A., Haataja, S., Iakovidis, I., Jacko, J., Jenkins, P., Karshmer, A., Korn, P., Marcus, A., Murphy, H., Stary, C., Vanderheiden, G., Weber, G., and Ziegler, J. (1998), "Toward an Information Society for All: An International R&D Agenda," *International Journal of Human-Computer Interaction*, Vol. 10, No. 2, pp. 107–134.

Stephanidis, C. (Ed.), Salvendy, G., Akoumianakis, D., Arnold, A., Bevan, N., Dardailler, D., Emiliani, P.L., Iakovidis, I., Jenkins, P., Karshmer, A., Korn, P., Marcus, A., Murphy, H., Oppermann, C., Stary, C., Tamura, H., Tscheligi, M., Ueda, H., Weber, G., and Ziegler, J. (1999), "Toward an Information Society

for All: HCI Challenges and R&D Recommendations," *International Journal of Human-Computer Interaction*, Vol. 11, No. 1, pp. 1–28.

Stephanidis, C., Paramythis, A., and Savidis, A. (2011 in press), "Developing Adaptive Interfaces for the Web," in *Handbook of Human Factors in Web Design*, 2nd ed., R. Proctor and K. Vu, Eds., CRC Press, Boca Raton, FL, Chapter 14.

Sun Microsystems (2010a), "The Synth Architecture," available: http://java.sun.com/docs/books/tutorial/uiswing/lookandfeel/synth.html.

Sun Microsystems (2010b), "JavaBeans Concept," available: http://java.sun.com/docs/books/tutorial/javabeans/whatis/index.html.

Vanderheiden, G., Chisholm, W., and Ewers, N. (1996, May), "Design of HTML Pages to Increase Their Accessibility to Users with Disabilities, Strategies for Today and Tomorrow," Technical Report, Trace R&D Centre, University of Wisconsin—Madison, Madison, WI.

Winograd, T. (2001), "From Programming Environments to Environments for Designing," in *User Interfaces for All: Concepts, Methods, and Tools*, C. Stephanidis, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 165–181.

World Wide Web Consortium (W3C) (1999, May), "Web Content Accessibility Guidelines 1.0," W3C Recommendation, available: http://www.w3.org/TR/WCAG10/.

World Wide Web Consortium (W3C) (2008, December), "Web Content Accessibility Guidelines 2.0," W3C Recommendation, available: http://www.w3.org/TR/WCAG20/.

Zimmermann, G., Vanderheiden, G., and Gilman, A. (2002), "Universal Remote Console—Prototyping for the Alternate Interface Access Standard," in *Universal Access: Theoretical Perspectives, Practice and Experience, Proceedings of the 7th ERCIM UI4ALL Workshop*, N. Carbonell and C. Stephanidis, Eds., Springer, Berlin and Heidelberg, pp. 524–531.

# SELECTED APPLICATIONS IN HUMAN FACTORS AND ERGONOMICS

# CHAPTER 55

# HUMAN FACTORS AND ERGONOMICS STANDARDS

**David Rodrick**
**University of Michigan—Dearborn**
**Dearborn, Michigan**

**Waldemar Karwowski**
**University of Central Florida**
**Orlando, Florida**

**Bohdana Sherehiy**
**Independent Consultant**
**Washington, District of Columbia**

## 1 INTRODUCTION

According to the International Organization for Standardization/International Electrochemical Commission (ISO/IEC) guide 2 (2004), a standard is a document established by consensus and approved by a recognized body that provides, for common and repeated use, rules, guidelines, or characteristics for activities or their results aimed at the achievement of the optimum degree of order in a given context (ISO/IEC, 2004). Additionally, the guide states that standards should be based on the consolidated results of science, technology, and experience and be aimed at the promotion of optimum community benefits. Geographically the standardization process can be distinguished into three main levels: national, regional, and international (see Figure 1). At the highest and broadest level of applicability are the international standards. The basis for worldwide standardization in all areas is provided primarily by three organizations: the ISO, IEC, and International Telecommunications Union (ITU). Standards related to human factors and ergonomics are developed by the ISO.

Following the international level, standards are being developed regionally. For example, in Europe, there are three standardization bodies: the European Committee for Standardization (CEN), European Committee for Electrotechnical Standardization (CENELEC), and European Telecommunications Standards Institute (ETSI). Their mission is to develop and achieve a coherent set of voluntary standards as a basis for a single
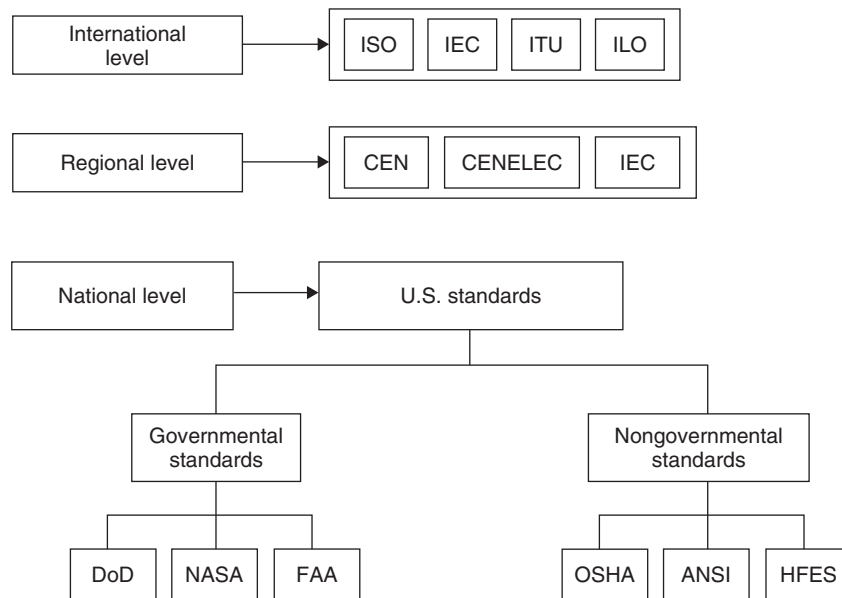
**Figure 1** Hierarchy of standards levels.

European market/European economic area (Wetting, 2002). At the national level almost every nation has its own national body for standards development. Examples of the national standardization organizations are the American National Standards Institute (ANSI), British Standards Institution (BSI), Deutsches Institut fur Normung (DIN), and Association Francaise de Normalisation (AFNOR). Standards can also be prepared by technical societies, labor organizations, consumer organizations, trade associations, and governmental agencies.

International, regional, and national standards are distinguished by documented standards development procedures. These procedures have been designed to ensure that all interested parties that can be affected by a particular standard will have an opportunity to represent their interest and participate in the standards development process. For example, ISO standards are developed by technical committees which consist of experts from the industrial, technical, and business sectors that are in need of standards. Many ISO national members apply public review procedures in order to consult draft standards with the interested parties, including representatives of government agencies, industrial and commercial organizations, professional and consumer associations, and the general public. The ISO national bodies are expected to take into account any feedback they receive and to present a consensus position to appropriate technical committees.

Standards are necessary to provide quality control and to support legislation and regulations to ensure an equal-opportunity and fairly operating international market. The main purpose of standardization is to achieve uniformity and interchangeability. Standardization limits the diversity of sizes, shapes, or component designs and prevents the generation of unneeded variation of products which do not provide unique services. Standardization is also the means by which society gathers and disseminates technical information (Spivak and Brenner, 2001). Harmonization of standards reduces trade barriers; promotes safety; allows interoperability of products, systems, and services; and promotes common technical understanding (Wetting, 2002).

The need for standardization from the human factors and ergonomics viewpoints can be illustrated by many "horror stories" following World War II. The war required that pilots fly on different types of aircraft which had no standard control arrangement in the cockpits (McDaniel, 1996). Many planes crashed because the pilots used wrong controls based on erroneously applied behavioral patterns. The human factors solutions included standardization of a single arrangement for engine controls and development of distinct shapes for the control handles (McDaniel, 1996). In this chapter we provide an overview of the human factors and ergonomics (HFE) standardization efforts around the world. Standards that cover any area of human factors and ergonomics at the international, regional, and national levels have been listed in this chapter. These areas include physical and cognitive ergonomics, human–computer interaction, human–system integration, and occupational health and safety. Unlike other fields, standards and the standardization process in human factors and ergonomics are relatively new. The pioneering landmark of publication on human factors and ergonomics standards by Karwowski (2006) comprehensively reviewed selected international and national standards and guidelines. The handbook was intended to disseminate knowledge about those standards and guidelines to professionals from a great variety of interrelated fields.

## 2 ISO STANDARDS FOR ERGONOMICS

The ISO was created in 1947 to coordinate the development of international standards. The ISO is a worldwide federation of national standards bodies from 163 countries. The mission of the ISO is to promote the development of standardization and related activities in the world in order to facilitate the international exchange of goods and services and to develop cooperation in the spheres of intellectual, scientific, technological, and economic activity (ISO/IEC, 2004). The International Electrotechnical Commission (IEC) is a nonprofit international standards organization that in collaboration with ISO prepares and publishes international standards related to electrical, electronics, and related technologies. In general, The ISO standards are developed based on three principles: (1) consensus—the views of all interests are taken into account: manufacturers, vendors and users, consumer groups, testing laboratories, governments, engineering professions, and research organizations; (2) industrywide—global solutions to satisfy global industries and consumers; and (3) voluntary—international standardization is market driven and therefore based on voluntary involvement of all interests in the marketplace. A standardization effort in the ISO goes through three phases:

1. An industry sector expresses the need for a standard by communicating to a national member body, which in turn proposes the new work item to the ISO. In this phase, definitions of the technical scope of the future standard are established following which the need for the standard is recognized and formally agreed upon by working groups comprising technical experts from the countries interested in the subject matter.

2. In the second phase, the consensus-building phase, the interested countries negotiate the detailed specifications within the standard.

3. The final phase comprises the formal approval of the resulting draft international standard (the acceptance criteria stipulate approval by two thirds of the ISO members that have participated actively in the standards development process and approval by 75% of all members that vote), following which the agreed-upon text is published as an ISO international standard.

In general, the standardization process of the ISO undergoes several stages from proposal for a new standard (stage code 00.00) to withdrawal of the standard (stage code 95.99). The distinct various stages, substages, and decision substages are presented in Table 1.

In 1974, the ISO formed technical committee TC 159 to develop standards in the field of ergonomics. The scope of ISO TC 159 activity has been described as standardization in the field of ergonomics, including terminology, methodology, and human factors data. According to the agreed-upon scope, ISO TC 159 (through standardization and coordination of related activities) promotes the adaptation of working and living conditions to human anatomical, psychological, and physiological characteristics in relation to the physical, sociological, and technological environment. Among the main objectives of such efforts are safety, health, well-being, and effectiveness (Parsons, 1995c). It should be noted that because of historical and organizational factors, many standards in the field of ergonomics are not developed by ISO TC 159.

At present, the ISO TC 159 organizational structure is administered by the DIN. The ergonomics standardization group consists of five subcommittees: SC 1, SC 3, SC 4, and SC 5. Through these subcommittees and various working groups (WG), TC 159 has published 108 standards. The subject areas of subcommittees and their organizational structure are presented in the Table 2.

### 2.1 Ergonomics Guiding Principles

The standards concerned with the ergonomics basic principles are elaborated by the TC 159/SC I subcommittee. The list of published standards and standards in development for the ergonomics guiding principles are provided in Table 3. ISO 6385:2004 is a basic standard that states the objectives of the ergonomics system design and provides definitions of basic terms and concepts in ergonomics (HFE). This standard establishes ergonomics principles of the work system design as basic guidelines. Such guidelines should be applied for the design of optimal working conditions with regard to human well-being, safety, and health, with consideration of technological and economic efficiency (Parsons, 1995a). Cullen (2007) examined the effectiveness of ISO 6385 while integrating between the designers and end users. In another study Andreas et al. (2009) utilized ISO 6385 in designing a verification and validation method (CRIOP) for an industrial setting.

The ISO 10075 standard dealing with mental workload is comprised of three parts. The first part presents terminology and main concepts. Part 2 covers guidelines on the design of work systems, including task, equipment, workspace, and work conditions with reference to the mental workload. Part 3 provides guidelines on measurement and assessment of mental workload. The third part specifies the requirements for the measurement instruments to be met at different levels of precision in measuring mental workload. In these standards it was stated that any human activity, even those that are considered primarily as physical activities, includes a mental workload (Nachreiner, 1995). Therefore, the described standards on mental workload are relevant to all kinds of work design. All of the abovementioned standards published standards in the review stage. Recent research efforts encompass examining the effectiveness of the standard (Schutte et al., 2007, 2009; Helvi et al., 2009) and the use of the standard to define and measure mental strain in noisy environments (Sandrock et al., 2009). ISO/DIS 268000 is currently a draft international standard (DIS).

### 2.2 Anthropometry and Biomechanics

The standards related to anthropometry and biomechanics are developed by the TC 159/SC 3 subcommittee. Currently, this subcommittee consists of two working

**Table 1 Stages of ISO Standardization Process**[a]

| Stage | Substage 00 Registration | Substage 20 State of main action | Substage 60 Completion of main action | Substage 90 Decision Substages — 92 Repeat an Earlier phase | 93 Repeat Current phase | 98 Abandon | 99 Proceed |
|---|---|---|---|---|---|---|---|
| 00 Preliminary stage | 00.00 Proposal for new project received | 00.20 Proposal for new project under review | 00.60 Close of review | | | 00.98 Proposal for new project abandoned | 00.99 Approval to ballot proposal for new project |
| 10 Proposal stage | 10.00 Proposal for new project registered | 10.20 New project ballot initiated | 10.60 Close of voting | 10.92 Proposal returned to submitter for further definition | | 10.98 New project rejected | 10.99 New project approved |
| 20 Preparatory stage | 20.00 New project registered in TC/SC work programme | 20.20 Working draft (wd) study initiated | 20.60 Close of comment period | | | 20.98 Project deleted | 20.99 WD approved for registration as CD |
| 30 Committee stage | 30.00 Committee draft (CD) registered | 30.20 CD study/ballot initiated | 30.60 Close of voting/comment period | 30.92 CD referred back to Working Group | | 30.98 Project deleted | 30.99 CD approved for registration as DIS |
| 40 Enquiry stage | 40.00 DIS registered | 40.20 DIS ballot initiated: 5 *months* | 40.60 Close of voting | 40.92 Full report circulated: DIS referred back to TC or SC | 40.93 Full report circulated: decision for new DIS ballot | 40.98 Project deleted | 40.99 Full report circulated: DIS approved for registration as FDIS |

| Stage | Substage | | | | 90 Decision Substages | | |
|---|---|---|---|---|---|---|---|
| 50 Approval stage | 50.00 FDIS registered for formal approval | 50.20 FDIS ballot initiated: 2 *months*.Proof sent to secretariat | 50.60 Close of voting.Proof returned by secretariat | 50.92 FDIS referred back to TC or SC | | 50.98 Project deleted | 50.99 FDIS approved for publication |
| 60 Publication stage | 60.00 International standard under publication | | 60.60 International standard published | | | | |
| 90 Review stage | | 90.20 International Standard under periodical review | 90.60 Close of review | 90.92 International Standard to be revised | 90.93 International Standard confirmed | | 90.99 Withdrawal of International Standard proposed by TC or SC |
| 95 Withdrawal Stage | | 95.20 Withdrawal ballot initiated | 95.60 Close of voting | 95.92 Decision not to withdraw International Standard | | | 95.99 Withdrawal of International Standard |

[a]DIS ⊃ Draft International Standard, FDIS ⊃ Final Draft International Standard.

**Table 2  Organizational Structure of ISO TC 159**

ISO/TC 159 "Ergonomics"

  ISO/TC 159/AG "AGAD; Advisory Group for Accessible Design"

  ISO/TC 159/CAG "Chairman Advisory Group"

  ISO/TC 159/SC 01 "General ergonomics principles"

    ISO/TC 159/SC 01/WG 01 "Principles of ergonomics and ergonomic design"

    ISO/TC 159/SC 01/WG 02 "Ergonomic principles related to mental work"

  ISO/TC 159/SC 03 "Anthropometry and biomechanics"

    ISO/TC 159/SC 03/WG 01 "Anthropometry"

    ISO/TC 159/SC 03/WG 04 "Human physical strength; manual handling and force limits"

  ISO/TC 159/SC 04 "Ergonomics of human-system interaction"

    ISO/TC 159/SC 04/CAG "Chairman Advisory Group"

    ISO/TC 159/SC 04/WG 01 "Fundamentals of controls and signalling methods"

    ISO/TC 159/SC 04/WG 02 "Visual display requirements"

    ISO/TC 159/SC 04/WG 03 "Controls, workplace and environmental requirements"

    ISO/TC 159/SC 04/WG 05 "Software ergonomics and human-computer dialogues"

    ISO/TC 159/SC 04/WG 06 "Human-centred design processes for interactive systems"

    ISO/TC 159/SC 04/WG 08 "Ergonomic design of control centres"

    ISO/TC 159/SC 04/WG 09 "Tactile and haptic"

    ISO/TC 159/SC 04/WG 10 "Accessible design for consumer products"

    ISO/TC 159/SC 04/WG 11 "Ease of operation of everyday products"

    ISO/TC 159/SC 04/WG 112 "Joint TC 159/SC 4 - JTC 1 SC 7 WG; Common industry formats for usability reports"

    ISO/TC 159/SC 04/WG 12 "Image safety"

  ISO/TC 159/SC 05 "Ergonomics of the physical environment"

    ISO/TC 159/SC 05/WG 01 "Thermal environments"

    ISO/TC 159/SC 05/WG 04 "Integrated environments"

    ISO/TC 159/SC 05/WG 05 "Physical environments for people with special requirements"

    ISO/TC 159/SC 05/WG 06 "Perceived air quality"

    ISO/TC 159/SC 05/WG 07 "Perception of air quality"

  ISO/TC 159/WG 02 "Ergonomics for people with special requirements"

---

**Table 3  ISO Standards for Ergonomic Guiding Principles**

| Reference Number | Title |
|---|---|
| ISO 6385:2004 | Ergonomic principles in the design of work systems |
| ISO 10075:1991 | Ergonomic principles related to mental workload: General terms and definitions |
| ISO 10075-2:1996 | Ergonomic principles related to mental workload, Part 2: Design principles |
| ISO 10075-3:2004 | Ergonomic principles related to mental workload, Part 3: Principles and requirements concerning methods for measuring and assessing mental workload |
| ISO/DIS 26800 | Ergonomics — General approach, principles and concepts |

provided in the ISO 7250 standards (parts 1 and 2). In addition to the lists of the basic anthropometric measurements, part 2 contains body measurements from ISO populations for comparison purposes. Recent research utilized ISO 7250 to obtain anthropometric data for workstation design (Deros et al., 2009), parametric human body modeling (Mustafa and Nadia, 2007), and estimation of anthropometric proportions (Gerd et al., 2009).

The three-part standards for the safety of machinery (ISO 15534) provide guidelines for determining the dimensions required for openings for access for machinery. The first part of this standard (ISO 15534-1:2000) presents principles for determining the dimensions for opening for whole-body access to machinery; the second part (ISO 15534-2:2000) specifies dimensions for the access openings. The third part of the safety of machinery standards (ISO 155343:2000) provides the requirements for the human body measurements (anthropometric data) that are needed for the calculation of access opening dimensions for machinery specified in the two previous parts of this standard (Parsons, 1995c). The anthropometric data are based on static measurements on nude people and representative of the European population of men and women.

ISO 14738:2002 describes principles for deriving dimensions from anthropometric measurements and applying them to the design of workstations at nonmobile machinery. This standard also specifies the body space requirements for equipment during normal operation in sitting and standing positions. ISO 15535:2006 specifies general requirements for anthropometric databases and their associated reports that contain measurements taken in accordance with ISO 7250. This standard presents such information as characteristics of the user population, sampling methods, and measurement items and statistics to make international comparison possible among various population segments.

groups: anthropometry (WG 1) and human physical strength; manual handling and force limits (WG 4). The list of the published standards and standards in development for anthropometry and biomechanics are presented in Table 4. The description of anthropometric measurements, which can be used as a basis for definition and comparison of population groups, are

**Table 4  Published ISO Standards and Standards under Development for Anthropometry and Biomechanics**

| Reference Number | Title |
|---|---|
| ISO 7250-1:2008 | Basic human body measurements for technological design, Part 1: Body measurement definitions and landmark |
| ISO/TR 7250-2:2010 | Basic human body measurements for technological design, Part 2: Statistical summaries of body measurements from individual ISO populations |
| ISO 11226:2000 | Ergonomics: Evaluation of static working postures |
| ISO 11226:2000/Cor 1:2006 | Corrigendum |
| ISO 11228-1:2003 | Ergonomics: Manual handling, Part 1: Lifting and carrying |
| ISO 11228-2:2007 | Ergonomics :Manual handling, Part 2: Pushing and pulling |
| ISO 11228-3:2007 | Ergonomics: Manual handling, Part 3: Handling of low loads at high frequency |
| ISO/NP TR 12295 | Ergonomics — Application document for ISO standards on manual handling (ISO 11228-1, ISO 11228-2 and ISO 11228-3) and working postures (ISO 11226) |
| ISO/NP TR 12296 | Ergonomics — Manual handling of people in the healthcare sector |
| ISO 14738:2002 | Safety of machinery: Anthropometric requirements for the design of workstations at machinery |
| ISO 14738:2002/Cor 1:2003 | Corrigendum |
| ISO 14738:2002/Cor 2:2005 | Corrigendum |
| ISO 15534-1:2000 | Ergonomic design for the safety of machinery, Part 1: Principles for determining the dimensions required for openings for whole-body access into machinery |
| ISO 15534-2:2000 | Ergonomic design for the safety of machinery, Part 2: Principles for determining the dimensions required for access openings |
| ISO 15534-3:2000 | Ergonomic design for the safety of machinery, Part 3: Anthropometric data |
| ISO 15535:2006 | General requirements for establishing anthropometric databases |
| ISO/TS 20646-1:2004 | Ergonomic procedures for the improvement of local muscular workloads, Part 1: Guidelines for reducing local muscular workloads |
| ISO 15536-1:2005 | Ergonomics: Computer manikins and body templates, Part 1: General requirements |
| ISO 15536-2:2007 | Ergonomics: Computer manikins and body templates, Part 2: Verification of functions and validation of dimensions for computer manikin systems |
| ISO 15537:2004 | Principles for selecting and using test persons for testing anthropometric aspects of industrial products and designs |
| ISO 20685:2010 | Three-dimensional scanning methodologies for internationally compatible anthropometric databases |

ISO 11228-1:2003 describes limits for manual lifting and carrying with consideration, respectively, of the intensity, frequency, and duration of the task. The limits recommended can be used in the assessment of several task variables and the health risk evaluation of the working population (Dickinson, 1995). This standard does not include holding of objects (without walking), pushing or pulling of objects, lifting with one hand, manual handling while seated, and lifting by two or more people. Holding, pushing, and pulling objects are included in parts 2 and 3 of ISO 11228, which are currently at the review stage. In conjunction with the ISO 11228 series, an application document (ISO/NP TR 12296) is under development. ISO/TS 20646-1:2004 present guidelines for application of various ergonomics standards related to local muscular workload (LMWL) and specify activities to reduce LMWL in workplaces. As part of development of new standards, in 2010, the ISO published a new standard, ISO 20685:2010, that addresses protocols for the use of three-dimensional (3D) surface-scanning systems in the acquisition of human body shape data and measurements defined in ISO 7250-1 that can be extracted from 3D scans.

## 2.3  Ergonomics of Human–System Interaction

The TC 159/SC 4 subcommittee develops the standards related to ergonomics of human–system interaction. The subcommittees are divided into 11 working groups, each dealing with a specific topic (see Table 2).

### 2.3.1  Controls and Signaling Methods

ISO 9355, *Ergonomic Requirements for the Design of Displays and Control Actuators*, provides guidelines for the design of displays and control actuators on work equipment, especially machines (see Table 5). A list of all parts of ISO 9355 is presented in Table 5. Part 1 describes general principles of human interactions with displays and controls. The other two parts provide recommendations on the selection, design, and location of information displays (part 2) and control actuators (part 3). Part 4 covers general principles for the location and arrangement of displays and actuators. No changes were made to these standards in the last five years. Currently these standards are in the review stages.

**Table 5  ISO Standards for Controls and Signaling Methods**

| Reference Number | Title |
| --- | --- |
| ISO 9355-1:1999 | Ergonomic requirements for the design of displays and control actuators, Part 1: Human interactions with displays and control actuators |
| ISO 9355-2:1999 | Ergonomic requirements for the design of displays and control actuators, Part 2: Displays |
| ISO 9355-3:2006 | Ergonomic requirements for the design of signals and control actuators, Part 3: Control actuators |

### 2.3.2  Visual Display Requirements

The multipart standard ISO 9241, *Ergonomics of Human System Interaction,* is believed to be the most important and known standard for ergonomic design (Stewart, 1995; Eibl, 2005). This standard presents general guidance and specific principles that need to be considered in the design of equipment, software, and tasks for office work with visual display terminals (VDTs). All parts of ISO 9241 are presented in Table 6. Major revisions were made to the parts of the standard. Along with its original parts, the standard includes the following series of standards:

- 100 series: Software ergonomics
- 200 series: Human–system interaction processes
- 300 series: Displays and display–related hardware
- 400 series: Physical input devices—ergonomics principles
- 500 series: Workplace ergonomics
- 600 series: Environment ergonomics
- 700 series: Application domains—Control rooms
- 900 series: Tactile and haptic interactions

Part 1 of ISO 9241, which described the basic underlining principles of the user performance approach, has been withdrawn along with parts 3 (visual display requirements), 7 (requirements for display with reflections), 8 (color displays), and 10 (dialogue principles). Part 2 describes how task requirements may be identified and specified in organizations and how task requirements can be incorporated into the system design and implementation process. Parts 4, 5, 6, and 9 provide assistance in the procurement and specification of the hardware and environmental components. Part 4 provides criteria for the keyboard and part 9 for no-keyboard input devices. Parts 5 and 6 establish ergonomic principles for the appropriate design and procurement of workstation, workstation equipment, and work environment for office work with VDTs (Eibl, 2005). Those two parts include such issues as technical

**Table 6  ISO 9241: Ergonomic Requirements for Office Work with VDTs**

| Reference Number | Title |
| --- | --- |
| ISO 9241-2:1992 | Part 2: Guidance on task requirements |
| ISO 9241-4:1998 | Part 4: Keyboard requirements |
| ISO 9241-5:1998 | Part 5: Workstation layout and postural requirements |
| ISO 9241-6:1999 | Part 6: Guidance on the work environment |
| ISO 9241-9:2000 | Part 9: Requirements for non-keyboard input devices |
| ISO 9241-11:199 | Part 11: Guidance on usability |
| ISO 9241-12:1998 | Part 12: Presentation of information |
| ISO 9241-13:1998 | Part 13: User guidance |
| ISO 9241-14:1997 | Part 14: Menu dialogues |
| ISO 9241-15:1997 | Part 15: Command dialogues |
| ISO 9241-16:1999 | Part 16: Direct manipulation dialogues |
| ISO 9241-17:1998 | Part 17: Form filling dialogues |
| ISO 9241-20:2008 | Part 20: Accessibility guidelines for information/communication technology (ICT) equipment and services |
| ISO/TR 9241-100:2010 | Part 100: Introduction to standards related to software ergonomics |
| ISO 9241-110:2006 | Part 110: Dialogue principles |
| ISO 9241-151:2008 | Part 151: Guidance on World Wide Web user interfaces |
| ISO 9241-171:2008 | Part 171: Guidance on software accessibility |
| ISO 9241-210:2010 | Part 210: Human-centered design for interactive systems |
| ISO/NP 9241-230 | Ergonomics of human-system interaction — Part 230: Human-centred design and evaluation methods |
| ISO 9241-300:2008 | Part 300: Introduction to electronic visual display requirements |
| ISO 9241-302:2008 | Part 302: Terminology for electronic visual displays |
| ISO 9241-303:2008 | Part 303: Requirements for electronic visual displays |
| ISO 9241-304:2008 | Part 304: User performance test methods for electronic visual displays |
| ISO 9241-305:2008 | Part 305: Optical laboratory test methods for electronic visual displays |
| ISO 9241-307:2008 | Part 307: Analysis and compliance test methods for electronic visual displays |
| ISO/TR 9241-308:2008 | Part 308: Surface-conduction electron-emitter displays (SED) |
| ISO/TR 9241-309:2008 | Part 309: Organic light-emitting diode (OLED) displays |

**Table 6** (*Continued*)

| Reference Number | Title |
|---|---|
| ISO/TR 9241-310:2010 | Part 310: Visibility, aesthetics and ergonomics of pixel defects |
| ISO/NP 9241-391 | Part 391: Requirements, analysis and compliance test methods for the reduction of photosensitive seizures |
| ISO 9241-400:2007 | Part 400: Principles and requirements for physical input devices |
| ISO 9241-410:2008 | Part 410: Design criteria for physical input devices |
| ISO 9241-410:2008/DAmd 1 | Amendment |
| ISO/DTS 9241-411 | Part 411: Evaluation methods for the design of physical input devices |
| ISO/DIS 9241-420 | Ergonomics of human-system interaction, Part 420: Selection of physical input devices |
| ISO 9241-920:2009 | Part 920: Guidance on tactile and haptic interactions |

| Under Development | Title |
|---|---|
| ISO/AWI TR 9241-1 | Ergonomics of human-system interaction, Part 1: Introduction to the ISO 9241 series |
| ISO 9241-129 | Ergonomics of human-system interaction, Part 129: Guidance on software individualization |
| ISO/DIS 9241-143 | Ergonomics of human-system interaction, Part 143: Form-based dialogues |
| ISO/CD 9241-154 | Ergonomics of human-system interaction, Part 154: Design guidance for interactive voice response (IVR) applications |
| ISO/NP 9241-230 | Ergonomics of human-system interaction, Part 230: Human-centred design and evaluation methods |
| ISO/NP 9241-391 | Ergonomics of human system interaction, Part 391: Requirements, analysis and compliance test methods for the reduction of photosensitive seizures |
| ISO/DTS 9241-411 | Ergonomics of human-system interaction, Part 411: Evaluation methods for the design of physical input devices |
| ISO/FDIS 9241-420 | Ergonomics of human-system interaction, Part 420: Selection of physical input devices |
| ISO/FDIS 9241-910 | Ergonomics of human-system interaction, Part 910: Framework for tactile and haptic interaction |

design of furniture and equipment for the workplace, space organization and workplace layout, physical characteristics of office work environment: lighting, noise, and vibrations. Part 10 has been replaced by part 110 (dialogue principles). Part 11 defines usability and specifies the usability evaluation in terms of the user performance and satisfaction measures (Dzida, 1995). Part 12 provides ergonomic recommendations for information presentation on the text-based displays and graphical user interfaces. Part 13 presents recommendations for different types of user guidance attributes of software interfaces, such as feedback, status, help, and error handling. Parts 14–17 deal with particular kinds of dialogue styles: menus, commands, direct manipulation, and form filling. In 2008, a new standard was included in the 9241 series, part 20, which deals with accessibility guidelines for information/communication technology (ICT) equipment and services.

The ISO 13406 standard provides recommendations additional to those of ISO 9241 in respect to visual displays based on flat panels. Two parts of this standard cover image quality requirements for the ergonomic design and evaluation of flat-panel displays. ISO 14915 provides additional recommendations to ISO 9241 concerning multimedia presentations. These standards are being developed by WGs 2, 3, 5, 6, and 9.

There is significant research effort examining the ISO 9241 standards to investigate their effectiveness and design and evaluation of products using an appropriate standard(s). Very recently, for example, ISO 9241-9 has been widely used to design and evaluate various pointing devices (input controllers) mainly due to the explosion of the gaming industry (Natapov et al., 2009), haptic feedback (Teather et al., 2010), gesture interface (Silva et al., 2003), and interface design (Ludger and Daniel, 2009).

### 2.3.3 Software Ergonomics

ISO 14915, *Software Ergonomics for Multimedia User Interfaces*, specifies recommendations and principles for the design of interactive multimedia user interfaces that integrate various media, such as static text, graphics, and images, and dynamic media such as audio, animation, and video. This standard focuses on issues related to integration of different media; hardware issues and multimodal input are not considered. The standard consist of three parts (see Table 7), which address general design principles and framework (part 1), multimedia navigation and control (Part 2), and media selection and combination (part 3). The committee draft ISO/CD 23973 considers ergonomics design principles for World Wide Web user interfaces. The standards are developed by WG 5. The effectiveness of ISO 14915 has been examined in the design of user interface (Luis et al., 2008; Bernsen and Dybkjær, 2009; Sgro et al., 2009). The standard has also been used to design multimedia user interface (Sutcliffe, 2009), evaluation of websites (Tobar et al., 2008), and PDA interface design (Blum and Khakzar, 2007).

**Table 7  ISO Standards for Software Ergonomics**

| Reference Number | Title |
|---|---|
| ISO 14915-1:2002 | Software ergonomics for multimedia user interfaces, part 1: Design principles and framework |
| ISO 14915-2:2003 | Software ergonomics for multimedia user interfaces, part 2: Multimedia navigation and control |
| ISO 14915-3:2002 | Software ergonomics for multimedia user interfaces, part 3: Media selection and combination |

### 2.3.4  Ergonomic Design of Control Centers

ISO 11064, *Ergonomic Design of Control Centers,* specifies requirements and presents principles for the ergonomic design of control centers. The list of all parts of ISO 11064 is provided in Table 8. The seven parts of this standard are concerned with the following issues: principles for the design of control centers, principles of control suite arrangements, control room and workstation layout and dimensions, displays and controls, environmental requirements, evaluation of control rooms, and ergonomic requirements for specific applications. WG 8 is responsible for development of these standards. ISO/CD 11064-4, Part 4: Layout and Dimensions of Workstations, is currently under development. In recent studies the standard has been used to design control rooms (Jamil et al., 2007; Isaac et al., 2008).

### 2.3.5  Human–System Interaction

The guidelines on the human-centered design process throughout the life cycle of computer-based interactive systems are described in ISO ISO/TR 18529:2000.

**Table 8  ISO 11064: Ergonomic Design of Control Centers**

| Reference Number | Title |
|---|---|
| ISO 11064-1:2000 | Part 1: Principles for the design of control centers |
| ISO 11064-2:2000 | Part 2: Principles for the arrangement of control suites |
| ISO 11064-3:1999 | Part 3: Control room layout |
| ISO 11064-4:2004 | Part 4: Layout and dimensions of workstations |
| ISO 11064-5:2008 | Part 5: Displays and controls |
| ISO 11064-6:2005 | Part 6: Environmental requirements for control centers |
| ISO 11064-7:2006 | Part 7: Principles for the evaluation of control centers |
| ISO/CD 11064-4 | Ergonomic design of control centres, Part 4: Layout and dimensions of workstations |

Usability methods supporting human-centered design are described in ISO/TR 16982:2002. Further standards concerned with human–system interaction address such issues as development and design of icons (ISO 11581), design of typical controls for multimedia functions (ISO 18035), icons for typical WWW browsers (ISO 18036), and definitions and metrics concerning software quality (ISO 9126). Table 9 shows the list of published ISO standards and standards in development for human–system interaction.

### 2.3.6  Ease of Operation and Accessibility

Recently TC 159 introduced standards related to ease of operation of everyday products, divided into four parts. Part 1 deals with design requirements, parts 2, 3, and 4 deal with test methods, and ISO/NP TS 20282-3 is currently under development. These standards are developed under WG 11. In 2010, through WG 10, the ISO introduced a new draft standard, ISO/FDIS 24503, Accessible Design, that deals with tactile dots and bars on consumer products. These standards are given in Table 10.

## 2.4  Ergonomics of the Physical Environment

The ISO TC159 SC5 document contains an international standard in the area of the ergonomics of the physical environment. The subcommittee is divided into seven working groups (see Table 2).

### 2.4.1  Ergonomics of the Thermal Environment

The standards on the ergonomics of thermal environments are concerned with heat stress, cold stress, and thermal comfort as well as with the thermal properties of clothing and metabolic heat production due to activity (Olesen, 1995). Physiological measures, skin reaction to contact with hot, moderate, and cold surfaces, and thermal comfort requirements for people with special requirements are also considered. The list of all standards and standards in development on thermal environment ergonomics are presented in Tables 11 and 12, respectively.

**Table 9  Published ISO Standards for Human-System Interaction**

| Reference Number | Title |
|---|---|
| ISO 1503:2008 | Spatial orientation and direction of movement: Ergonomic requirements |
| ISO/TR 16982:2002 | Ergonomics of human-system interaction: Usability methods supporting human-centered design |
| ISO/TS 18152:2010 | Ergonomics of human-system interaction: Specification for the process assessment of human-system issues |
| ISO/TR 18529:2000 | Ergonomics of human-system interaction: Human-centered lifecycle process descriptions |

**Table 10  Ease of Operation of Everyday Products and Accessibility**

| Reference Number | Title |
| --- | --- |
| *Published* | |
| ISO 20282-1:2006 | Ease of operation of everyday products, Part 1: Design requirements for context of use and user characteristics |
| ISO/TS 20282-2:2006 | Ease of operation of everyday products, Part 2: Test method for walk-up-and-use products |
| ISO/PAS 20282-3:2007 | Ease of operation of everyday products, Part 3: Test method for consumer products |
| ISO/PAS 20282-4:2007 | Ease of operation of everyday products, Part 4: Test method for the installation of consumer products |
| *Under Development* | |
| ISO/NP TS 20282-3 | Ease of operation of everyday products, Part 3: Test method for consumer products |
| ISO/FDIS 24503 | Ergonomics : Accessible design: Tactile dots and bars on consumer products |

The main thermal comfort standard. ISO 7730, provides a method for predicting the thermal sensation and the degree of discomfort, which can also be used to specify acceptable environmental conditions for comfort. This method is based on the predicted mean vote (PMV) and predicted percentage of dissatisfied (PPD) thermal comfort indices (Olesen and Parsons, 2002). It also provides methods for the assessment of local discomfort caused by draughts, asymmetric radiation, and temperature gradients. Other thermal environment standards address such issues as thermal comfort for people with special requirements (ISO/TS 14415:2005), responses on contact with surfaces at moderate temperature (ISO 13732-2:2001), and thermal comfort in vehicles (ISO 14505, parts 1–3). Standards concerned with thermal comfort assessment specify measuring instruments (ISO 7726:1998), methods for estimation of metabolic heat production (ISO 8996:2004), estimation of clothing properties (ISO 9920:2007), and subjective assessment methods (ISO 10551:1995). ISO 11399:1995 provides information needed for the correct and effective application of international standards concerned with the ergonomics of the thermal environment. The standards that are under development deal with assessment of physical quantities, perceived indoor air quality, and mathematical models for human physiological responses.

### 2.4.2  Communication in Noisy Environments

The standard for communication in noisy environments includes warnings, danger signals, and speech. The list

**Table 11  Published ISO Standards for Ergonomics of the Thermal Environment**

| Reference Number | Title |
| --- | --- |
| ISO 7243:1989 | Hot environments: Estimation of the heat stress on working man, based on the WBGT-index (wet bulb globe temperature) |
| ISO 7726:1998 | Ergonomics of the thermal environment: Instruments for measuring physical quantities |
| ISO 7730:2005 | Ergonomics of the thermal environment: Analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices and local thermal comfort criteria |
| ISO 7933:2004 | Ergonomics of the thermal environment: Analytical determination and interpretation of heat stress using calculation of the predicted heat strain |
| ISO 8996:2004 | Ergonomics of the thermal environment: Determination of metabolic rate |
| ISO 9886:2004 | Ergonomics: Evaluation of thermal strain by physiological measurements |
| ISO 9920:2007 | Ergonomics of the thermal environment: Estimation of thermal insulation and water vapor resistance of a clothing ensemble |
| ISO 10551:1995 | Ergonomics of the thermal environment: Assessment of the influence of the thermal environment using subjective judgment scales |
| ISO 11079:2007 | Ergonomics of the thermal environment: Determination and interpretation of cold stress when using required clothing insulation (IREQ) and local cooling effects |
| ISO 11399:1995 | Ergonomics of the thermal environment: Principles and application of relevant international standards |
| ISO 12894:2001 | Ergonomics of the thermal environment: Medical supervision of individuals exposed to extreme hot or cold environments |
| ISO 13731:2001 | Ergonomics of the thermal environment: Vocabulary and symbols |
| ISO 13732-1:2006 | Ergonomics of the thermal environment: Methods for the assessment of human responses to contact with surfaces, Part 1: Hot surfaces |

*(continued overleaf)*

**Table 11** (*Continued*)

| Reference Number | Title |
|---|---|
| ISO/TS 13732-2:2001 | Ergonomics of the thermal environment: Methods for the assessment of human responses to contact with surfaces, Part 2: Human contact with surfaces at moderate temperature |
| ISO 13732-3:2005 | Ergonomics of the thermal environment: Methods for the assessment of human responses to contact with surfaces, Part 3: Cold surfaces |
| ISO/TS 14415:2005 | Ergonomics of the thermal environment: Application of International Standards to people with special requirements |
| ISO/TS 14505-1:2007 | Ergonomics of the thermal environment: Evaluation of thermal environment in vehicles, Part 1: Principles and methods for assessment of thermal stress |
| ISO 14505-2:2006 | Ergonomics of the thermal environment: Evaluation of thermal environment in vehicles, Part 2: Determination of equivalent temperature |
| ISO 14505-3:2006 | Ergonomics of the thermal environment: Thermal environment in vehicles, Part 3: Evaluation of thermal comfort using human subjects |
| ISO 15265:2004 | Ergonomics of the thermal environment: Risk assessment strategy for the prevention of stress or discomfort in thermal working conditions |
| ISO 15743:2008 | Ergonomics of the thermal environment: Cold workplaces: Risk assessment and management |
| ISO 24500:2010 | Ergonomics: Accessible design: Auditory signals for consumer products |

**Table 12 ISO Standards under Development for Ergonomics of the Thermal Environment**

| Reference Number | Title |
|---|---|
| ISO/AWI 7726 | Ergonomics of the thermal environment: Instruments for measuring physical quantities |
| ISO/NP 16077 | Ergonomic of the physical environment: A method for assessing perceived indoor air quality using human subject panels |
| ISO/NP 16418 | Ergonomics of the thermal environment: Mathematical model for predicting and evaluating the dynamic human physiological responses to the thermal environments |
| ISO/AWI 16594 | Guide for working practices for moderate thermal environments |
| ISO/DIS 28802 | Ergonomics of the physical environment: Assessment by means of an environmental survey involving physical measurement of the environment and subjective responses of people |
| ISO/DIS 28803 | Ergonomics of the physical environment: Application of international standards to people with special requirements |

of related standards is provided in Table 13. The ISO 7731:2003 document specifies the requirements and test methods for auditory danger signals and gives guidelines for the design of the signals in the public and in workplaces. This document also provides definitions to guide in the use of the standards concerned with noisy environments. Criteria for the perception of the visual danger signals are provided in ISO 11428:1996. This international standard specifies the safety and ergonomic requirements and the corresponding physical measurements.

ISO 11429:1996 specifies a system of danger and information signals in reference to various degrees of urgency. This standard applies to all danger signals that have to be clearly perceived and differentiated, from extreme urgency to "all clear." Guidance on delectability is provided in terms of luminance, Illuminance, and contrast, considering both surface and point sources. ISO 9921:2003 describes a method for prediction of the effectiveness of speech communication in the presence of noise generated by machinery as well as in any other noisy environment. The following parameters are taken into account in this standard: the ambient noise at the speaker's position, the ambient noise at the listener's position, the distance between the communication partners, and a variety of physical and personal conditions (Parsons, 1995b). ISO/TR 19358:2002 deals with the testing and assessment of speech-related products and services. The standards that are under development are ISO/AWI 16613 (sound pressure levels for products and PA systems), ISO/FDIS 24501 (consumer products), and ISO/FDIS 24502 (specification of age-related luminance contrast for colored light), where AWI refers to approved new work item and FDIS refers to final draft international standard.

### 2.4.3 Lighting of Workplaces

ISO 8995 (Part 1: 2002 and Part 3: 2006) was developed by the ISO 159 SC5 WG 2 "Lighting" group in collaboration with the International Commission on Illumination (CIE). This standard describes the

**Table 13 ISO Standards for Danger Signals and Communication in Noisy Environments**

| Reference Number | Title |
|---|---|
| *Published* | |
| ISO 7731:2003 | Ergonomics: Danger signals for public and work areas — Auditory danger signals |
| ISO 11428:1996 | Ergonomics: Visual danger signals — General requirements, design and testing |
| ISO 11429:1996 | Ergonomics: System of auditory and visual danger and information signals |
| ISO 9921:2003 | Ergonomics: Assessment of speech communication |
| ISO/TR 19358:2002 | Ergonomics: Construction and application of tests for speech technology |
| *Under Development* | |
| ISO/AWI 16613 | Ergonomics: Accessible design — Sound pressure levels of spoken announcements for products and public address systems |
| ISO/FDIS 24501 | Ergonomics: Accessible design — Sound pressure levels of auditory signals for consumer products |
| ISO/FDIS 24502 | Ergonomics: Accessible design — Specification of age-related luminance contrast for coloured light |

principles of the visual ergonomics, identifies factors that influence visual performance, and presents criteria for the achievement of an acceptable visual environment (Parsons, 1995b).

## 3 CEN STANDARDS FOR ERGONOMICS

In Europe, there are three standardization organizations: CEN, CENELEC, and ETSI. Their aim is development and achievement of a coherent set of voluntary standards that can provide a basis for a single European market/European economic area without internal frontiers for goods and services inside Europe. Their work is carried out in conjunction with worldwide bodies and the national standards bodies in Europe (Wetting, 2002). Members of the European Union (EU) and the European Fair Trade Association (EFTA) have agreed to implement CEN standards in their national system and to withdraw conflicting national standards.

In 1987, the CEN established CEN/TC 122, "Ergonomics," which is responsible for development of the European ergonomic standards (Dul et al., 1996). The scope of CEN/TC 122 is standardization in the field of ergonomics, in order to meet the requirements for ergonomic and efficient products under the conditions

of free trade so that enhanced health, safety, and well-being of the consumers and users as well as the overall performance are ensured. These ergonomics standards are aimed to be applied in work systems as well as in private use, for new technologies, and changes in the population (CEN, 2008). The organizational structure of the CEN/TC 122 is presented in Table 14.

The ISO and CEN have signed a formal agreement, *Agreement on Technical Cooperation between ISO and CEN* (the Vienna Agreement), that established close cooperation between these standardization bodies. The ISO and CEN decided to harmonize the development of their standards and to cooperate regarding exchange of information and standards drafting. According to this agreement, the ISO standards are adopted by the CEN, and vice versa. Table 15 presents published CEN ergonomics standards. Most of the ergonomic standards published by CEN/TC 122 are adoption, or adaptation, of ISO standards. In the last five years, the CEN ergonomic standards in development have been published as formal standards.

### 3.1 Other International Standards Related to Ergonomics

For historical and organizational factors, many ISO and CEN standards in the field of ergonomics have not been developed by the technical committees ISO TC 159 and CEN and TC 122. Some ergonomics areas covered by other ISO and CEN technical committees are presented in Table 16. The lists of published ISO standards related to the ergonomics area, but developed by groups other than the TC 159 committee, are provided in Table 17.

**Table 14 Organizational Structure of CEN/TC 122**

| Working Group | Title |
|---|---|
| CEN/TC 122/WG 1 | Anthropometry |
| CEN/TC 122/WG 2 | Ergonomic design principles |
| CEN/TC 122/WG 3 | Surface temperatures |
| CEN/TC 122/WG 4 | Biomechanics |
| CEN/TC 122/WG 5 | Ergonomics of human–computer interaction |
| CEN/TC 122/WG 6 | Signals and controls |
| CEN/TC 122/WG 8 | Danger signals and speech communication in noisy environments |
| CEN/TC 122/WG 9 | Ergonomics of personal protective equipment (PPE) |
| CEN/TC 122/WG 10 | Ergonomic design principles for the operability of mobile machinery |
| CEN/TC 122/WG 11 | Ergonomics of the thermal environment |
| CEN/TC 122/WG 12 | Integrating ergonomic principles for machinery design |

**Table 15 Published CEN Standards for Ergonomics**

| CEN Reference | Title | ISO Standard |
|---|---|---|
| *Ergonomics Principles* | | |
| EN ISO 10075-1:2000 | Ergonomic principles related to mental workload, Part 1: General terms and definitions | ISO 10075:1991 |
| EN ISO 10075-2:2000 | Ergonomic principles related to mental workload, Part 2: Design principles | ISO 10075-2:1996 |
| EN ISO 10075-3:2004 | Ergonomic principles related to mental workload, Part 3: Principles and requirements concerning methods for measuring and assessing mental workload | ISO 10075-3:2004 |
| EN ISO 6385:2004 | Ergonomic principles in the design of work systems | ISO 6385:2004 |
| EN 13921:2007 | Personal protective equipment — Ergonomic principles | |
| EN ISO 15537:2004 | Principles for selecting and using test persons for testing anthropometric aspects of industrial products and designs | ISO 15537:2004 |
| *Anthropometrics and Biomechanics* | | |
| EN 1005-1:2001+A1:2008 | Safety of machinery: Human physical performance, Part 1: Terms and definitions | |
| EN 1005-2:2003+A1:2008 | Safety of machinery: Human physical performance, Part 2: Manual handling of machinery and component parts of machinery | |
| EN 1005-3:2002+A1:2008 | Safety of machinery: Human physical performance, Part 3: Recommended force limits for machinery operation | |
| EN 1005-4:2005+A1:2008 | Safety of machinery — Human physical performance, Part 4: Evaluation of working postures and movements in relation to machinery | |
| EN 1005-5:2007 | Safety of machinery — Human physical performance, Part 5: Risk assessment for repetitive handling at high frequency | |
| EN 13861:2002 | Safety of machinery: Guidance for the application of ergonomics standards in the design of machinery | |
| EN 547-1:1996+A1:2008 | Safety of machinery: Human body measurements, Part 1: Principles for determining the dimensions required for openings for whole body access into machinery | |
| EN 547-2:1996+A1:2008 | Safety of machinery: Human body measurements, Part 2: Principles for determining the dimensions required for access openings | |
| EN 547-3:1996+A1:2008 | Safety of machinery: Human body measurements, Part 3: Anthropometric data | |
| EN 614-1:2006+A1:2009 | Safety of machinery: Ergonomic design principles, Part 1: Terminology and general principles | |
| EN 614-2:2000+A1:2008 | Safety of machinery: Ergonomic design principles, Part 2: Interactions between the design of machinery and work tasks | |
| EN ISO 7250-1:2010 | Basic human body measurements for technological design | ISO 7250:2008 |
| EN ISO 14738:2008 | Safety of machinery: Anthropometric requirements for the design of workstations at machinery | ISO 14738:2002, including Cor 1:2003 and Cor 2:2005 |
| EN ISO 15535:2006 | General requirements for establishing anthropometric databases | ISO 15535:2006 |
| EN ISO 20685:2010 | 3-D scanning methodologies for internationally compatible anthropometric databases | ISO 20685:2010 |
| *Ergonomics Design of Control Centers* | | |
| EN ISO 11064-1:2000 | Ergonomic design of control centers, Part 1: Principles for the design of control centers | ISO 11064-1:2000 |
| EN ISO 11064-2:2000 | Ergonomic design of control centers, Part 2: Principles for the arrangement of control suites | ISO 11064-2:2000 |
| EN ISO 11064-3:1999 | Ergonomic design of control centers, Part 3: Control room layout | ISO 11064-3:1999 |
| EN ISO 11064-3:1999/AC:2002 | Ergonomic design of control centers, Part 3: Control room layout | ISO 11064-3:1999/Cor.1:2002 |

**Table 15  (*Continued*)**

| CEN Reference | Title | ISO Standard |
|---|---|---|
| EN ISO 11064-4:2004 | Ergonomic design of control centres, Part 4: Layout and dimensions of workstations | ISO 11064-4:2004 |
| EN ISO 11064-5:2008 | Ergonomic design of control centres, Part 5: Displays and controls | ISO 11064-5:2008 |
| EN ISO 11064-6:2005 | Ergonomic design of control centres, Part 6: Environmental requirements for control centres | ISO 11064-6:2005 |
| EN ISO 11064-7:2006 | Ergonomic design of control centres, Part 7: Principles for the evaluation of control centres | ISO 11064-7:2006 |
| | ***Human–System Interaction*** | |
| EN ISO 13731:2001 | Ergonomics of the thermal environment: Vocabulary and symbols | ISO 13731:2001 |
| EN ISO 14915-1:2002 | Software ergonomics for multimedia user interfaces, Part 1: Design principles and framework | ISO 14915-1:2002 |
| EN ISO 14915-2:2003 | Software ergonomics for multimedia user interfaces, Part 2: Multimedia navigation and control | ISO 14915-2:2003 |
| EN ISO 14915-3:2002 | Software ergonomics for multimedia user interfaces, Part 3: Media selection and combination | ISO 14915-3:2002 |
| EN ISO 15536-1:2008 | Ergonomics — Computer manikins and body templates, Part 1: General requirements | ISO 15536-1:2005 |
| EN ISO 15536-2:2007 | Ergonomics — Computer manikins and body templates, Part 2: Verification of functions and validation of dimensions for computer manikin systems | ISO 15536-2:2007 |
| EN ISO 9921:2003 | Ergonomics: Assessment of speech communication | ISO 9921:2003 |
| | ***Danger Signals*** | |
| EN 842:1996 +A1:2008 | Safety of machinery: Visual danger signals —  General requirements, design and testing | |
| EN 981:1996 +A1:2008 | Safety of machinery: System of auditory and visual danger and information signals | |
| | ***Thermal Environments*** | |
| EN 27243:1993 | Hot environments: estimation of the heat stress on working man, based on the WBGT-index (wet bulb globe temperature) | ISO 7243:1989 |
| EN ISO 10551:2001 | Ergonomics of the thermal environment: Assessment of the influence of the thermal environment using subjective judgement scales | ISO 10551:1995 |
| EN ISO 11399:2000 | Ergonomics of the thermal environment: Principles and application of relevant international standards | ISO 11399:1995 |
| EN ISO 12894:2001 | Ergonomics of the thermal environment: Medical supervision of individuals exposed to extreme hot or cold environments | ISO 12894:2001 |
| EN ISO 13732-1:2008 | Ergonomics of the thermal environment — Methods for the assessment of human responses to contact with surfaces, Part 1: Hot surfaces | ISO 13732-1:2006 |
| EN ISO 13732-3:2008 | Ergonomics of the thermal environment — Methods for the assessment of human responses to contact with surfaces, Part 3: Cold surfaces | ISO 13732-3:2005 |
| EN ISO 14505-2:2006 | Ergonomics of the thermal environment — Evaluation of thermal environments in vehicles, Part 2: Determination of equivalent temperature | ISO 14505-2:2006 |
| EN ISO 14505-2:2006/AC:2009 | Ergonomics of the thermal environment — Evaluation of thermal environments in vehicles, Part 2: Determination of equivalent temperature | ISO 14505-2:2006/Cor 1:2007 |
| EN ISO 14505-3:2006 | Ergonomics of the thermal environment — Evaluation of the thermal environment in vehicles, Part 3: Evaluation of thermal comfort using human subjects | ISO 14505-3:2006 |
| EN ISO 15265:2004 | Ergonomics of the thermal environment — Risk assessment strategy for the prevention of stress or discomfort in thermal working conditions | ISO 15265:2004 |

**Table 15** (*Continued*)

| CEN Reference | Title | ISO Standard |
|---|---|---|
| EN ISO 15743:2008 | Ergonomics of the thermal environment — Cold workplaces — Risk assessment and management | ISO 15743:2008 |
| EN ISO 7726:2001 | Ergonomics of the thermal environment: Instruments for measuring physical quantities | ISO 7726:1998 |
| EN ISO 7730:2005 | Ergonomics of the thermal environment — Analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices and local thermal comfort criteria | ISO 7730:2005 |
| EN ISO 7933:2004 | Ergonomics of the thermal environment — Analytical determination and interpretation of heat stress using calculation of the predicted heat strain | ISO 7933:2004 |
| EN ISO 8996:2004 | Ergonomics of the thermal environment — Determination of metabolic rate | ISO 8996:2004 |
| EN ISO 9886:2004 | Ergonomics: Evaluation of thermal strain by physiological measurements | ISO 9886:2004 |
| EN ISO 9920:2009 | Ergonomics of the thermal environment: Estimation of the thermal insulation and evaporative resistance of a clothing ensemble | ISO 9920:2007, Corrected version 2008-11-01 |
| EN ISO 11079:2007 | Ergonomics of the thermal environment — Determination and interpretation of cold stress when using required clothing insulation (IREQ) and local cooling effects | ISO 11079:2007 |
| | *Displays and Control Actuators* | |
| EN 894-1:1997 +A1:2008 | Safety of machinery: Ergonomics requirements for the design of displays and control actuators, Part 1: General principles for human interactions with displays and control actuators | |
| EN 894-2:1997+A1:2008 | Safety of machinery: Ergonomics requirements for the design of displays and control actuators, Part 2: Displays | |
| EN 894-3:2000+A1:2008 | Safety of machinery: Ergonomics requirements for the design of displays and control actuators, Part 3: Control actuators | |
| EN 894-4:2010 | Safety of machinery — Ergonomics requirements for the design of displays and control actuators, Part 4: Location and arrangement of displays and control actuators | |

## 4 ILO GUIDELINES FOR OCCUPATIONAL SAFETY AND HEALTH MANAGEMENT SYSTEMS

The popularity and success of a systematic and standardized approach to the management systems introduced by the ISO led to the view that this type of approach can also improve the management of occupational safety and health. Following this idea, the International Labour Organization (ILO) developed voluntary guidelines on OSH management systems which reflect ILO values and ensure protection of workers' safety and health (ILO-OSH; ILO, 2001).

The ILO was founded at the Versailles Congress in 1919 and became a specialized agency of the United Nations (UN) in 1946. The aims of the ILO are to promote rights at work, encourage decent employment opportunities, enhance social protection, and strengthen dialogue in handling work-related issues through its four principal strategic objectives: (1) to promote and realize standards and fundamental principles and rights at work, (2) to create greater opportunities for women and men to secure decent employment, (3) to enhance the coverage and effectiveness of social protection for all, and (4) to strengthen tripartisan and social dialogue

(ILO, 2010). The ILO represents the interests of three parties treated equally: employers, employee organizations, and government agencies.

The ILO (2001) guidelines provide recommendations concerning design and implementation of occupational safety and health management systems (OSHMS) that allow for integration of OSH with the general enterprise management system. The ILO guidelines state that these recommendations are addressed to all who are responsible for the occupational safety and health management. These guidelines are nonmandatory and are not intended to replace national laws and regulations. The ILO (2001) document distinguished two levels of guideline application: national and organizational. At the national level ILO-OSH (ILO, 2001) provides recommendations for the establishment of a national framework for OSHMS. The guidelines suggest that this process should be supported by the provision of the relevant national laws and regulations.

Establishment of a national framework for OSHMS included the following actions (ILO, 2001): (1) nomination of competent institution(s) for OSHMS, (2) formulation of a coherent national policy, and (3) development

**Table 16  Ergonomic Areas Covered in Standards Developed by Other ISO and CEN Technical Committees**

| Topic | Technical Committee | |
| --- | --- | --- |
| | ISO | CEN |
| Safety of machines | TC 199 | TC 114 |
| Vibration and shock | TC 108 | TC 211 |
| Noise and acoustics | TC 43 | TC 211 |
| Lighting | | TC 169 |
| Respiratory protective devices | | TC 79 |
| Eye protection | | TC 85 |
| Head protection | | TC 158 |
| Hearing protection | | TC 159 |
| Protection against falls | TC 94 | TC 160 |
| Foot and leg protection | | TC 161 |
| Protective clothing | | TC 162 |
| Radiation protection | TC 85 | |
| Air quality | TC 146 | |
| Assessment and workplace exposure | | TC 137 |
| Office machines | TC 95 | |
| Information procession | TC 97 | |
| Road vehicles | TC 22 | |
| Safety color and signs | TC 80 | |
| Graphical symbols | TC 145 | |

*Source:* Dul et al. (1996).

of national and tailored guidelines. The process of establishment of a national framework for OSHMS and its components is presented in Figure 2.

At the organizational level ILO (2001), guidelines establish employer responsibilities regarding occupational safety and health management and emphasize the importance of compliance with national laws and regulations. ILO (2001) suggests that OSH management system elements be integrated into overall organizational policy and management strategies actions (ILO, 2001).
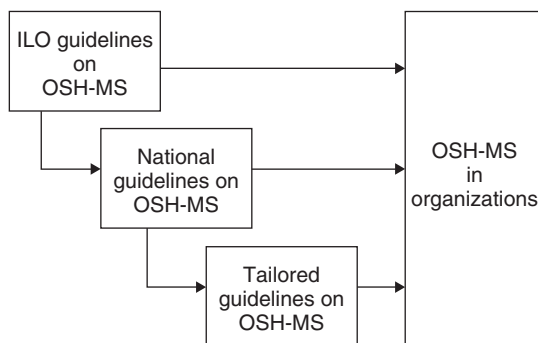


**Figure 2**  Establishment of a national framework for the OSHMS. (From ILO, 2001.)

The OSH management systems in the organization consist of five main sections: policy, organizing, planning and implementation, evaluation, and action for improvement. These elements correspond to the Demming cycle of plan–do–check–act, internationally accepted as the basis for the systems approach to management. The OSHMS main sections and their elements are listed in Table 18.

ILO (2001) guidelines require establishment by the employer of the OSH policy in consultation with workers and their representatives and define the content of such policy. The ILO-OSH (ILO, 2001) guidelines also indicate the importance of OSH policy integration and compatibility with other management systems in the organization. These guidelines emphasize the necessity of worker participation in the OSH management system in the organization. Therefore, workers should be consulted regarding OSH activities and should be encouraged to participate in OSHMS, including a safety and health committee. The organizing section of the guidelines underlines the need for allocation of responsibility and accountability for the implementation and performance of the OSH management system to the senior management. This section also includes requirements related to competence and training in the OSH field and defines the necessary documentation and communications activities. The planning and implementation section includes the elements of initial review, system planning, development and implementation, OSH objectives, and hazard prevention. The initial review identifies the actual states of the organization with regard to the OSH and creates the baseline for OSH policy implementation. The evaluation section consists of performance monitoring and measurement, investigation of work-related diseases and incidents, audit, and management review. The guidelines require carrying out internal audits of the OSHMS according to the policies established. Action for improvement includes the elements of preventive and corrective action and continual improvement. The final section underlines the need for continual improvement of OSH performance through the development of policies, systems, and techniques to prevent and control work-related injuries and diseases.

## 5  U.S. STANDARDS FOR HUMAN FACTORS AND ERGONOMICS

### 5.1  U.S. Government Standards

Among the HFE U.S. government standards, two documents are usually mentioned as basic: a military standard providing human engineering design criteria (MIL-STD-1472) and a human–system integration standard (NASA-STD-300) (Chapanis. 1996; McDaniel, 1996). In addition, there are more specific standards that have been developed by such departments as the Department of Defense (DOD), Department of Transportation (DOT), Department of Energy (DOE), and U.S. Nuclear Regulatory Commission (NRC). Additionally, a large number of handbooks that contain more detailed and descriptive information concerning human factor and

**Table 17  HFE Standards Published by Other Than TC 159 ISO Technical Committees**

| Reference Number | Title |
|---|---|
| *CIE. International Commission on Illumination* | |
| ISO 8995-1:2002 (CIE S 008/E:2001) | Lighting of work places, Part 1: Indoor |
| ISO 8995-3:2006 (CIE S 016/E:2005) | Lighting of work places, Part 3: Lighting requirements for safety and security of outdoor work places |
| *JTC 1/SC 6: Telecommunications and Information Exchange between Systems* | |
| ISO/IEC 10021-1:2003 | Information technology: Message Handling Systems (MHS), Part 1: System and service overview |
| ISO/IEC 10021-2:2003 | Information technology: Message Handling Systems (MHS): Overall architecture |
| ISO/IEC TR 10021-11:1999 | Information technology: Message Handling Systems (MHS): MHS Routing — Guide for messaging systems managers |
| ISO/IEC 10021-7:2003 | Information technology: Message Handling Systems (MHS): Interpersonal messaging system |
| ISO/IEC 10021-4:2003 | Information technology: Message Handling Systems (MHS): Message transfer system: Abstract service definition and procedures |
| ISO/IEC 10021-5:1999 | Information technology: Message Handling Systems (MHS): Message store: Abstract service definition |
| ISO/IEC 10021-10:1999 | Information technology: Message Handling Systems (MHS): MHS routing |
| ISO/IEC 10021-9:1999 | Information technology: Message Handling Systems (MHS): Electronic Data Interchange Messaging System |
| ISO/IEC 10021-8:1999 | Information technology: Message Handling Systems (MHS), Part 8: Electronic Data Interchange Messaging Service |
| ISO/IEC 10021-6:2003 | Information technology: Message Handling Systems (MHS): Protocol specifications |
| *JTC 1/SC 7: Software and System Engineering* | |
| ISO/IEC 9126-1:2001 | Software engineering: Product quality, Part 1: Quality model |
| ISO/IEC TR 9126-2:2003 | Software engineering: Product quality, Part 2: External metrics |
| ISO/IEC TR 9126-3:2003 | Software engineering: Product quality, Part 3: Internal metrics |
| ISO/IEC TR 9126-4:2004 | Software engineering: Product quality, Part 4: Quality in use metrics |
| ISO/IEC 12207:2008 | Systems and software engineering: Software life cycle processes |
| ISO/IEC 14598-1:1999 | Information technology: Software product evaluation, Part 1: General overview |
| ISO/IEC 14598-2:2000 | Software engineering: Product evaluation, Part 2: Planning and management |
| ISO/IEC 14598-3:2000 | Software engineering: Product evaluation, Part 3: Process for developers |
| ISO/IEC 14598-4:1999 | Software engineering: Product evaluation, Part 4: Process for acquirers |
| ISO/IEC 14598-5:1998 | Information technology: Software product evaluation, Part 5: Process for evaluators |
| ISO/IEC 14598-6:2001 | Software engineering: Product evaluation, Part 6: Documentation of evaluation modules |
| ISO/IEC 15288:2008 | Systems and software engineering: System life cycle processes |
| ISO/IEC 15504-1:2004 | Information technology: Process assessment, Part 1: Concepts and vocabulary |
| ISO/IEC 15504-2:2003 | Information technology: Process assessment, Part 2: Performing an assessment |
| ISO/IEC 15504-3:2004 | Information technology: Process assessment, Part 3: Guidance on performing an assessment |
| ISO/IEC 15504-5:2006 | Information technology: Process Assessment, Part 5: An exemplar process assessment model |
| ISO/IEC TR 15504-6:2008 | Information technology: Process assessment, Part 6: An exemplar system life cycle process assessment model |
| ISO/IEC TR 15504-7:2008 | Information technology: Process assessment, Part 7: Assessment of organizational maturity |
| ISO/IEC 15910:1999 | Information technology: Software user documentation process |
| ISO/IEC 18019:2004 | Software and system engineering: Guidelines for the design and preparation of user documentation for application software |
| ISO/IEC TR 19760:2003 | Systems engineering: A guide for the application of ISO/IEC 15288 (System life cycle processes) |
| ISO/IEC 20926:2009 | Software and systems engineering: Software measurement: IFPUG functional size measurement method 2009 |

**Table 17 (*Continued*)**

| Reference Number | Title |
|---|---|
| ISO/IEC 20968:2002 | Software engineering: Mk II function point analysis — Counting practices manual |
| *JTC 1/SC 22: Programming Languages, Their Environments, and System Software Interfaces* | |
| ISO/IEC TR 11017:1998 | Information technology: Framework for internationalization |
| ISO/IEC TR 15942:2000 | Information technology: Programming languages — Guide for the use of the Ada programming language in high integrity systems |
| *JTC 1 /SC 27: IT Security Techniques* | |
| ISO/IEC 21827:2008 | Information technology : Security techniques: Systems Security Engineering–Capability Maturity Model® (SSE-CMM®) |
| *JTC 1/SC 35: User Interfaces* | |
| ISO/IEC 11581-1:2000 | Information technology: User system interfaces and symbols — Icon symbols and functions, Part 1: Icons — general |
| ISO/IEC 11581-2:2000 | Information technology: User system interfaces and symbols — Icon symbols and functions, Part 2: Object icons |
| ISO/IEC 11581-3:2000 | Information technology: User system interfaces and symbols — Icon symbols and functions, Part 3: Pointer icons |
| ISO/IEC 11581-5:2004 | Information technology: User system interfaces and symbols — Icon symbols and functions, Part 5: Tool icons |
| ISO/IEC 11581-6:1999 | Information technology: User system interfaces and symbols — Icon symbols and functions, Part 6: Action icons |
| ISO/IEC 15411:1999 | Information technology: Segmented keyboard layouts |
| ISO/IEC 18035:2003 | Information technology: Icon symbols and functions for controlling multimedia software applications |
| *TC 8/SC 5: Ships' Bridge Layout* | |
| ISO 8468:2007 | Ships and marine technology: Ship's bridge layout and associated equipment: Requirements and guidelines |
| *TC 8/SC 6: Navigation* | |
| ISO 16273:2003 | Ships and marine technology: Night vision equipment for high-speed craf — Operational and performance requirements, methods of testing and required test results |
| *TC 20: Aircraft and Space Vehicles* | |
| ISO/TR 10201:2001 | Aerospace: Standards for electronic instruments and systems |
| *TC 20/SC 1: Aerospace Electrical Requirements* | |
| ISO 6858:1982 | Aircraft: Ground support electrical supplies — General requirements |
| *TC 20/SC 14: Space Systems and Operations* | |
| ISO 16091:2002 | Space systems: Integrated logistic support |
| ISO 17399:2003 | Space systems: Man-systems integration |
| *TC 21/SC 3: Fire Detection and Alarm Systems* | |
| ISO 12239:2003 | Fire detection and fire alarm systems: Smoke alarms |
| *TC 22/SC 3: Electrical and Electronic Equipment* | |
| ISO 11748-1:2001 | Road vehicles: Technical documentation of electrical and electronic systems, Part 1: Content of exchanged documents |
| ISO 11748-2:2001 | Road vehicles: Technical documentation of electrical and electronic systems, Part 2: Documentation agreement |
| ISO 11748-3:2002 | Road vehicles: Technical documentation of electrical and electronic systems, Part 3: Application example |
| ISO/TR 15497:2000 | Road vehicles: Development guidelines for vehicle based software |
| *TC 22/SC 13: Ergonomics Applicable to Road Vehicles* | |
| ISO 2575:2010 | Road vehicles: Symbols for controls, indicators and tell-tales |
| ISO 3958:1996 | Passenger cars: Driver hand-control reach |
| ISO 4040:2009 | Road vehicles: Location of hand controls, indicators and tell-tales in motor vehicles |
| ISO 6549:1999 | Road vehicles: Procedure for H- and R-point determination |
| ISO/TR 9511:1991 | Road vehicles: Driver hand-control reach — In-vehicle checking procedure |

*(continued overleaf)*

**Table 17** (*Continued*)

| Reference Number | Title |
|---|---|
| ISO/TS 12104:2003 | Road vehicles: Gearshift patterns — Manual transmissions with power-assisted gear change and automatic transmissions with manual-gearshift mode |
| ISO 12214:2010 | Road vehicles: Direction-of-motion stereotypes for automotive hand controls |
| ISO 15005:2002 | Road vehicles: Ergonomic aspects of transport information and control systems — Dialogue management principles and compliance procedures |
| ISO 15007-1:2002 | Road vehicles: Measurement of driver visual behaviour with respect to transport information and control systems, Part 1: Definitions and parameters |
| ISO/TS 15007-2:2001 | Road vehicles: Measurement of driver visual behaviour with respect to transport information and control systems, Part 2: Equipment and procedures |
| ISO 15008:2009 | Road vehicles: Ergonomic aspects of transport information and control systems — Specifications and test procedures for in-vehicle visual presentation |
| ISO/TS 16951:2004 | Road vehicles: Ergonomic aspects of transport information and control systems (TICS) — Procedures for determining priority of on-board messages presented to drivers |
| ISO 17287:2003 | Road vehicles: Ergonomic aspects of transport information and control systems — Procedure for assessing suitability for use while driving |
| | *TC 22/SC 17: Visibility* |
| ISO 7397-1:1993 | Passenger cars: Verification of driver's direct field of view, Part 1: Vehicle positioning for static measurement |
| ISO 7397-2:1993 | Passenger cars: Verification of driver's direct field of view, Part 2: Test method |
| | *TC 23/SC 3: Safety and Comfort of the Operator* |
| ISO 4254-1:2008 | Agricultural machinery: Safety, Part 1: General requirements |
| ISO 4254-5:2008 | Agricultural machinery: Safety, Part 5: Power-driven soil-working machines |
| ISO 4254-6:2009 | Agricultural machinery: Safety, Part 6: Sprayers and liquid fertilizer distributors |
| ISO 4254-7:2008 | Agricultural machinery: Safety, Part 7: Combine harvesters, forage harvesters and cotton harvesters |
| ISO 4254-8:2009 | Agricultural machinery, Safety, Part 8: Solid fertilizer distributors |
| ISO 4254-9:2008 | Agricultural machinery, Safety, Part 9: Seed drills |
| ISO 4254-10:2009 | Agricultural machinery, Safety, Part 10: Rotary tedders and rakes |
| ISO/TS 15077:2008 | Tractors and self-propelled machinery for agriculture: Operator controls — Actuating forces, displacement, location and method of operation |
| | *TC 23/SC 4: Tractors* |
| ISO 4253:1993 | Agricultural tractors: Operator's seating accommodation — Dimensions |
| ISO 5721:1989 | Tractors for agriculture: Operator's field of vision |
| | *TC 23/SC 7: Equipment for Harvesting and Conservation* |
| ISO 8210:1989 | Equipment for harvesting: Combine harvesters — Test procedure |
| | *TC 23/SC 14: Operator Controls, Operator Symbols and Other Displays, Operator Manuals* |
| ISO 3767-1:1998 | Tractors, machinery for agriculture and forestry, powered lawn and garden equipment: Symbols for operator controls and other displays, Part 1: Common symbols |
| ISO 3767-2:2008 | Tractors, machinery for agriculture and forestry, powered lawn and garden equipment: Symbols for operator controls and other displays, Part 2: Symbols for agricultural tractors and machinery |
| ISO 3767-3:1995 | Tractors, machinery for agriculture and forestry, powered lawn and garden equipment: Symbols for operator controls and other displays, Part 3: Symbols for powered lawn and garden equipment |
| ISO 3767-4:1993 | Tractors, machinery for agriculture and forestry, powered lawn and garden equipment: Symbols for operator controls and other displays, Part 4: Symbols for forestry machinery |
| ISO 3767-5:1992 | Tractors, machinery for agriculture and forestry, powered lawn and garden equipment: Symbols for operator controls and other displays, Part 5: Symbols for manual portable forestry machinery |
| | *TC 23/SC 15: Machinery for Forestry* |
| ISO 11850:2003 | Machinery for forestry: Self-propelled machinery — Safety requirements |

**Table 17 (*Continued*)**

| Reference Number | Title |
|---|---|
| *TC 23/SC 17: Manually Portable Forest Machinery* | |
| ISO 8334:2007 | Forestry machinery: Portable chain-saws — Determination of balance and maximum holding moment |
| ISO 11680-1:2000 | Machinery for forestry: Safety requirements and testing for pole-mounted powered pruners, Part 1: Units fitted with an integral combustion engine |
| ISO 11680-2:2000 | Machinery for forestry: Safety requirements and testing for pole-mounted powered pruners, Part 2: Units for use with a back-pack power source |
| ISO 11681-1:2004 | Machinery for forestry: Portable chain-saw safety requirements and testing, Part 1: Chain-saws for forest service |
| ISO 11681-2:2006 | Machinery for forestry: Portable chain-saw safety requirements and testing, Part 2: Chain-saws for tree service |
| ISO 11806:1997 | Agricultural and forestry machinery: Portable hand-held combustion engine driven brush cutters and grass trimmers — Safety |
| ISO 14740:1998 | Forest machinery: Backpack power units for brush-cutters, grass-trimmers, pole-cutters and similar appliances — Safety requirements and testing |
| *TC 23/SC 18: Irrigation and Drainage Equipment and Systems* | |
| ISO/TR 8059:1986 | Irrigation equipment: Automatic irrigation systems — Hydraulic control |
| *TC38: Textiles* | |
| ISO 15831:2004 | Clothing: Physiological effect — Measurement of thermal insulation by means of a thermal manikin |
| *TC 43/SC 1: Noise* | |
| ISO 11690-1:1996 | Acoustics: Recommended practice for the design of low-noise workplaces containing machinery, Part 1: Noise control strategies |
| ISO 11690-2:1996 | Acoustics: Recommended practice for the design of low-noise workplaces containing machinery, Part 2: Noise control measures |
| ISO/TR 11690-3:1997 | Acoustics: Recommended practice for the design of low-noise workplaces containing machinery, Part 3: Sound propagation and noise prediction in workrooms |
| ISO 15667:2000 | Acoustics: Guidelines for noise control by enclosures and cabins |
| *TC 46: Information and Documentation* | |
| ISO 7220:1996 | Information and documentation: Presentation of catalogues of standards |
| *TC 59/SC 3: Functional/User Requirements and Performance in Building Construction* | |
| ISO 6242-1:1992 | Building construction: Expression of users' requirements, Part 1: Thermal requirements |
| ISO 6242-2:1992 | Building construction: Expression of users' requirements, Part 2: Air purity requirements |
| ISO 6242-3:1992 | Building construction: Expression of users' requirements, Part 3: Acoustical requirements |
| *TC 67: Materials, Equipment, and Offshore Structures for Petroleum, Petrochemical, and Natural Gas Industries* | |
| ISO 13879:1999 | Petroleum and natural gas industries: Content and drafting of a functional specification |
| ISO 13880:1999 | Petroleum and natural gas industries: Content and drafting of a technical specification |
| *TC 67/SC 6: Processing Equipment and Systems* | |
| ISO 13702:1999 | Petroleum and natural gas industries: Control and mitigation of fires and explosions on offshore production installations — Requirements and guidelines |
| ISO 15544:2000 | Petroleum and natural gas industries: Offshore production installation — Requirements and guidelines for emergency response |
| ISO 17776:2000 | Petroleum and natural gas industries: Offshore production installations — Guidelines on tools and techniques for hazard identification and risk assessment |
| *TC 69: Applications of Statistical Methods* | |
| ISO 10725:2000 | Acceptance sampling plans and procedures for the inspection of bulk materials |
| *TC 72/SC 5: Industrial Laundry and Dry-Cleaning Machinery and Accessories* | |
| ISO 8230-1:2008 | Safety requirements for dry-cleaning machines, Part 1: Common safety requirements |
| ISO 8230-2:2008 | Safety requirements for dry-cleaning machines, Part 2: Machines using perchloroethylene |
| ISO 8230-3:2008 | Safety requirements for dry-cleaning machines, Part 3: Machines using combustible solvents |
| ISO 10472-1:1997 | Safety requirements for industrial laundry machinery, Part 1: Common requirements |

**Table 17** (*Continued*)

| Reference Number | Title |
|---|---|
| ISO 10472-2:1997 | Safety requirements for industrial laundry machinery, Part 2: Washing machines and washer-extractors |
| ISO 10472-3:1997 | Safety requirements for industrial laundry machinery, Part 3: Washing tunnel lines including component machines |
| ISO 10472-4:1997 | Safety requirements for industrial laundry machinery, Part 4: Air dryers |
| ISO 10472-5:1997 | Safety requirements for industrial laundry machinery, Part 5: Flatwork ironers, feeders and folders |
| ISO 10472-6:1997 | Safety requirements for industrial laundry machinery, Part 6: Ironing and fusing presses |
| *TC 72/SC 8: Safety Requirements for Textile Machinery* | |
| ISO 11111-1:2009 | Textile machinery: Safety requirements, Part 1: Common requirements |
| ISO 11111-2:2005 | Textile machinery: Safety requirements, Part 2: Spinning preparatory and spinning machines |
| ISO 11111-3:2005 | Textile machinery: Safety requirements, Part 3: Nonwoven machinery |
| ISO 11111-4:2005 | Textile machinery: Safety requirements, Part 4: Yarn processing, cordage and rope manufacturing machinery |
| ISO 11111-5:2005 | Textile machinery: Safety requirements, Part 5: Preparatory machinery to weaving and knitting |
| ISO 11111-6:2005 | Textile machinery: Safety requirements, Part 6: Fabric manufacturing machinery |
| ISO 11111-7:2005 | Textile machinery: Safety requirements, Part 7: Dyeing and finishing machinery |
| *TC 85/SC 2: Radiation Protection* | |
| ISO 17874-1:2010 | Remote handling devices for radioactive materials, Part 1: General requirements |
| ISO 17874-2:2004 | Remote-handling devices for radioactive materials, Part 2: Mechanical master-slave manipulators |
| ISO 17874-4:2006 | Remote handling devices for radioactive materials, Part 4: Power manipulators |
| ISO 17874-5:2007 | Remote handling devices for radioactive materials, Part 5: Remote handling tongs |
| *TC 92/SC 3: Fire Threat to People and Environment* | |
| ISO/TS 13571:2007 | Life-threatening components of fire: Guidelines for the estimation of time available for escape using fire data |
| ISO 10068:1998 | Mechanical vibration and shock: Free, mechanical impedance of the human hand−arm system at the driving point |
| ISO 13090-1:1998 | Mechanical vibration and shock: Guidance on safety aspects of tests and experiments with people, Part 1: Exposure to whole-body mechanical vibration and repeated shock |
| *TC 92/SC 4: Fire Safety Engineering* | |
| ISO/TR 13387-1:1999 | Fire safety engineering, Part 1: Application of fire performance concepts to design objectives |
| ISO/TR 13387-2:1999 | Fire safety engineering, Part 2: Design fire scenarios and design fires |
| ISO/TR 13387-3:1999 | Fire safety engineering, Part 3: Assessment and verification of mathematical fire models |
| ISO/TR 13387-4:1999 | Fire safety engineering, Part 4: Initiation and development of fire and generation of fire effluents |
| ISO/TR 13387-5:1999 | Fire safety engineering, Part 5: Movement of fire effluents |
| ISO/TR 13387-6:1999 | Fire safety engineering, Part 6: Structural response and fire spread beyond the enclosure of origin |
| ISO/TR 13387-7:1999 | Fire safety engineering, Part 7: Detection, activation and suppression |
| ISO/TR 13387-8:1999 | Fire safety engineering, Part 8: Life safety: Occupant behaviour, location and condition |
| *TC 94/SC 4: Personal Equipment for Protection Against Falls* | |
| ISO 10333-1:2000 | Personal fall-arrest systems, Part 1: Full-body harnesses |
| ISO 10333-2:2000 | Personal fall-arrest systems, Part 2: Lanyards and energy absorbers |
| ISO 10333-3:2000 | Personal fall-arrest systems, Part 3: Self-retracting lifelines |
| ISO 10333-4:2002 | Personal fall-arrest systems, Part 4: Vertical rails and vertical lifelines incorporating a sliding-type fall arrester |
| ISO 10333-5:2001 | Personal fall-arrest systems, Part 5: Connectors with self-closing and self-locking gates |
| ISO 10333-6:2004 | Personal fall-arrest systems, Part 6: System performance tests |
| *TC 94/SC 13: Protective Clothing* | |
| ISO 11393-1:1998 | Protective clothing for users of hand-held chain-saws, Part 1: Test rig driven by a flywheel for testing resistance to cutting by a chain-saw |

**Table 17** (*Continued*)

| Reference Number | Title |
|---|---|
| ISO 11393-2:1999 | Protective clothing for users of hand-held chain-saws, Part 2: Test methods and performance requirements for leg protectors |
| ISO 11393-3:1999 | Protective clothing for users of hand-held chain-saws, Part 3: Test methods for footwear |
| ISO 11393-4:2003 | Protective clothing for users of hand-held chain-saws, Part 4: Test methods and performance requirements for protective gloves |
| ISO 11393-5:2001 | Protective clothing for users of hand-held chain-saws, Part 5: Test methods and performance requirements for protective gaiters |
| ISO 11393-6:2007 | Protective clothing for users of hand-held chain-saws, Part 6: Test methods and performance requirements for upper body protectors |
| ISO 13688:1998 | Protective clothing: General requirements |
| ISO 16603:2004 | Clothing for protection against contact with blood and body fluids: Determination of the resistance of protective clothing materials to penetration by blood and body fluids — Test method using synthetic blood |
| ISO 16604:2004 | Clothing for protection against contact with blood and body fluids: Determination of resistance of protective clothing materials to penetration by blood-borne pathogens — Test method using Phi-X 174 bacteriophage |
| | ***TC 101: Continuous Mechanical Handling Equipment*** |
| ISO/TR 5045:1979 | Continuous mechanical handling equipment: Safety code for belt conveyors — Examples for guarding of nip points |
| | ***TC 108/SC 2: Measurement and Evaluation of Mechanical Vibration and Shock as Applied to Machines, Vehicles, and Structures*** |
| ISO 14964:2000 | Mechanical vibration and shock: Vibration of stationary structures — Specific requirements for quality management in measurement and evaluation of vibration |
| | ***TC 108/SC 4: Human Exposure to Mechanical Vibration and Shock*** |
| ISO 2631-1:1997 | Mechanical vibration and shock: Evaluation of human exposure to whole-body vibration, Part 1: General requirements |
| ISO 2631-2:2003 | Mechanical vibration and shock: Evaluation of human exposure to whole-body vibration, Part 2: Vibration in buildings (1 Hz to 80 Hz) |
| ISO 2631-4:2001 | Mechanical vibration and shock: Evaluation of human exposure to whole-body vibration, Part 4: Guidelines for the evaluation of the effects of vibration and rotational motion on passenger and crew comfort in fixed-guide way transport systems |
| ISO 2631-5:2004 | Mechanical vibration and shock: Evaluation of human exposure to whole-body vibration, Part 5: Method for evaluation of vibration containing multiple shocks |
| ISO 5349-1:2001 | Mechanical vibration: Measurement and evaluation of human exposure to hand-transmitted vibration, Part 1: General requirements |
| ISO 5349-2:2001 | Mechanical vibration: Measurement and evaluation of human exposure to hand-transmitted vibration, Part 2: Practical guidance for measurement at the workplace |
| | ***TC 108/SC 4: Human Exposure to Mechanical Vibration and Shock*** |
| ISO 5982:2001 | Mechanical vibration and shock: Range of idealized values to characterize seated-body biodynamic response under vertical vibration |
| ISO 6897:1984 | Guidelines for the evaluation of the response of occupants of fixed structures, especially buildings and off-shore structures, to low-frequency horizontal motion (0,063 to 1 Hz) |
| ISO 8727:1997 | Mechanical vibration and shock: Human exposure — Biodynamic coordinate systems |
| ISO 9996:1996 | Mechanical vibration and shock: Disturbance to human activity and performance — Classification |
| ISO 13091-1:2001 | Mechanical vibration: Vibrotactile perception thresholds for the assessment of nerve dysfunction, Part 1: Methods of measurement at the fingertips |
| ISO 13091-2:2003 | Mechanical vibration: Vibrotactile perception thresholds for the assessment of nerve dysfunction, Part 2: Analysis and interpretation of measurements at the fingertips |
| | ***TC 121/SC 3: Lung Ventilators and Related Equipment*** |
| ISO 8185:2007 | Respiratory tract humidifiers for medical use — Particular requirements for respiratory humidification systems |
| IEC 60601-1-8:2006 | Medical electrical equipment, Part 1-8: General requirements for safety — Collateral standard: General requirements, tests and guidance for alarm systems in medical electrical equipment and medical electrical systems |

**Table 17** (*Continued*)

| Reference Number | Title |
|---|---|
| IEC 60601-1-10:2007 | Medical electrical equipment, Part 1-10: General requirements for basic safety and essential performance — Collateral standard: Requirements for the development of physiologic closed-loop controllers |
| IEC 60601-1-11:2010 | Medical electrical equipment, Part 1-11: General requirements for basic safety and essential performance — Collateral standard: Requirements for medical electrical equipment and medical electrical systems used in the home healthcare environment |
| IEC 60601-2-12:2001 | Medical electrical equipment, Part 2-12: Particular requirements for the safety of lung ventilators — Critical care ventilators |
| IEC 60601-2-13:2003 | Medical electrical equipment, Part 2-13: Particular requirements for the safety and essential performance of an aesthetic systems |
| IEC 60601-2-52:2009 | Medical electrical equipment, Part 2-52: Particular requirements for the basic safety and essential performance of medical beds |
| | *TC 127/SC 1: Test Methods Relating to Machine Performance* |
| ISO 8813:1992 | Earth-moving machinery: Lift capacity of pipe layers and wheeled tractors or loaders equipped with side boom |
| | *TC 127/SC 2: Safety Requirements and Human Factors* |
| ISO 2860:1992 | Earth-moving machinery: Minimum access dimensions |
| ISO 2867:2006 | Earth-moving machinery: Access systems |
| ISO 3164:1995 | Earth-moving machinery: Laboratory evaluations of protective structures — Specifications for deflection-limiting volume |
| ISO 3411:2007 | Earth-moving machinery: Physical dimensions of operators and minimum operator space envelope |
| ISO 3449:2005 | Earth-moving machinery: Falling-object protective structures — Laboratory tests and performance requirements |
| ISO 3450:1996 | Earth-moving machinery: Braking systems of rubber-tyred machines — Systems and performance requirements and test procedures |
| ISO 3457:2003 | Earth-moving machinery: Guards — Definitions and requirements |
| ISO 3471:2008 | Earth-moving machinery: Roll-over protective structures — Laboratory tests and performance requirements |
| ISO 5006:2006 | Earth-moving machinery: Operator's field of view: Test method and performance criteria |
| ISO 5010:2007 | Earth-moving machinery: Rubber-tyred machines — Steering requirements |
| ISO 5353:1995 | Earth-moving machinery, and tractors and machinery for agriculture and forestry: Seat index point |
| ISO 6682:1986 | Earth-moving machinery: Zones of comfort and reach for controls |
| ISO 7096:2000 | Earth-moving machinery: Laboratory evaluation of operator seat vibration |
| ISO 8643:1997 | Earth-moving machinery: Hydraulic excavator and backhoe loader boom-lowering control device — Requirements and tests |
| ISO 9244:2008 | Earth-moving machinery: Machine safety labels — General principles |
| ISO/TR 9953:1996 | Earth-moving machinery: Warning devices for slow-moving machines — Ultrasonic and other systems |
| ISO 10262:1998 | Earth-moving machinery: Hydraulic excavators — Laboratory tests and performance requirements for operator protective guards |
| ISO 10567:2007 | Earth-moving machinery: Hydraulic excavators — Lift capacity |
| ISO 10570:2004 | Earth-moving machinery: Articulated frame lock — Performance requirements |
| ISO 10263-1:2009 | Earth-moving machinery: Operator enclosure environment, Part 1: Terms and definitions |
| ISO 10263-2:2009 | Earth-moving machinery: Operator enclosure environment, Part 2: Air filter element test method |
| ISO 10263-3:2009 | Earth-moving machinery: Operator enclosure environment, Part 3: Pressurization test method |
| ISO 10263-4:2009 | Earth-moving machinery: Operator enclosure environment, Part 4: Heating, ventilating and air conditioning (HVAC) test method and performance |
| ISO 10263-5:2009 | Earth-moving machinery: Operator enclosure environment, Part 5: Windscreen defrosting system test method |
| ISO 10263-6:2009 | Earth-moving machinery: Operator enclosure environment, Part 6: Determination of effect of solar heating |

**Table 17 (*Continued*)**

| Reference Number | Title |
|---|---|
| ISO 10533:1993 | Earth-moving machinery: Lift-arm support devices |
| ISO 10968:2004 | Earth-moving machinery: Operator's controls |
| ISO 11112:1995 | Earth-moving machinery: Operator's seat — Dimensions and requirements |
| ISO 12117:1997 | Earth-moving machinery: Tip-over protection structure (TOPS) for compact excavators — Laboratory tests and performance requirements |
| ISO 12117-2:2008 | Earth-moving machinery: Laboratory tests and performance requirements for protective structures of excavators, Part 2: Roll-over protective structures (ROPS) for excavators of over 6 t |
| ISO 12508:1994 | Earth-moving machinery: Operator station and maintenance areas — Bluntness of edges |
| ISO 13333:1994 | Earth-moving machinery: Dumper body support and operator's cab tilt support devices |
| ISO 13459:1997 | Earth-moving machinery: Dumpers — Trainer seat/enclosure |
| ISO 17063:2003 | Earth-moving machinery: Braking systems of pedestrian-controlled machines — Performance requirements and test procedures |
| *TC 131/SC 9: Installations and Systems* | |
| ISO 4413:1998 | Hydraulic fluid power: General rules relating to systems |
| ISO 4414:1998 | Pneumatic fluid power: General rules relating to systems |
| *TC 136: Furniture* | |
| ISO 5970:1979 | Furniture: Chairs and tables for educational institutions — Functional sizes |
| *TC 163/SC 2: Calculation Methods* | |
| ISO 13790:2008 | Energy performance of buildings: Calculation of energy use for space heating and cooling |
| *TC 171/SC 2: Application Issues* | |
| ISO/TR 14105:2001 | Electronic imaging: Human and organizational issues for successful electronic image management (EIM) implementation |
| *TC 172/SC 9: Electro-Optical Systems* | |
| ISO 11553-1:2005 | Safety of machinery: Laser processing machines — Safety requirements, Part 1: General safety requirements |
| ISO 11553-2:2007 | Safety of machinery: Laser processing machines — Safety requirements, Part 2: Safety requirements for hand-held laser processing devices |
| *TC 173: Assistive Products for Persons with Disability* | |
| ISO 11199-1:1999 | Walking aids manipulated by both arms: Requirements and test methods, Part 1: Walking frames |
| ISO 11199-2:2005 | Walking aids manipulated by both arms: Requirements and test methods, Part 2: Rollators |
| ISO 11199-3:2005 | Walking aids manipulated by both arms: Requirements and test methods, Part 3: Walking tables |
| ISO 11334-1:2007 | Walking aids manipulated by one arm: Requirements and test methods, Part 1: Elbow crutches |
| ISO 11334-4:1999 | Walking aids manipulated by one arm: Requirements and test methods, Part 4: Walking sticks with three or more legs |
| *TC 173/SC 3: Aids for Ostomy and Incontinence* | |
| ISO 15621:1999 | Urine-absorbing aids: General guidance on evaluation |
| *TC 173/SC 6: Hoists for Transfer of Persons* | |
| ISO 10535:2006 | Hoists for the transfer of disabled persons: Requirements and test methods |
| *TC 176/SC 1: Concepts and Terminology* | |
| ISO 9000:2005 | Quality management systems: Fundamentals and vocabulary |
| *TC 176/SC 2: Quality Systems* | |
| ISO 9004:2009 | Managing for the sustained success of an organization — A quality management approach |
| *TC 178: Lifts, Escalators, and Moving Walks* | |
| ISO 14798:2009 | Lifts (elevators), escalators and moving walks: Risk assessment and reduction methodology |
| *TC 184/SC 4: Industrial Data* | |
| ISO 10303-214:2010 | Industrial automation systems and integration: Product data representation and exchange, Part 214: Application protocol — Core data for automotive mechanical design processes |

**Table 17  (Continued)**

| Reference Number | Title |
|---|---|
| **TC 184/SC 5: Architecture, Communications and Integration Frameworks** | |
| ISO 15704:2000 | Industrial automation systems: Requirements for enterprise-reference architectures and methodologies |
| ISO 16100-1:2009 | Industrial automation systems and integration: Manufacturing software capability profiling for interoperability, Part 1: Framework |
| ISO 16100-2:2003 | Industrial automation systems and integration: Manufacturing software capability profiling for interoperability, Part 2: Profiling methodology |
| ISO 16100-3:2005 | Industrial automation systems and integration: Manufacturing software capability profiling for interoperability, Part 3: Interface services, protocols and capability templates |
| ISO 16100-4:2006 | Industrial automation systems and integration: Manufacturing software capability profiling for interoperability, Part 4: Conformance test methods, criteria and reports |
| ISO 16100-5:2009 | Industrial automation systems and integration: Manufacturing software capability profiling for interoperability, Part 5: Methodology for profile matching using multiple capability class structures |
| **TC 188: Small Craft** | |
| ISO 15027-1:2002 | Immersion suits, Part 1: Constant wear suits, requirements including safety |
| ISO 15027-2:2002 | Immersion suits, Part 2: Abandonment suits, requirements including safety |
| ISO 15027-3:2002 | Immersion suits, Part 3: Test methods |
| **TC 199: Safety of Machinery** | |
| ISO 11161:2007 | Safety of machinery: Integrated manufacturing systems — Basic requirements |
| ISO 12100:2010 | Safety of machinery: General principles for design — Risk assessment and risk reduction |
| ISO 13849-1:2006 | Safety of machinery: Safety-related parts of control systems, Part 1: General principles for design |
| ISO 13849-2:2003 | Safety of machinery: Safety-related parts of control systems, Part 2: Validation |
| ISO 13851:2002 | Safety of machinery: Two-hand control devices — Functional aspects and design principles |
| ISO 13856-2:2005 | Safety of machinery: Pressure-sensitive protective devices, Part 2: General principles for the design and testing of pressure-sensitive edges and pressure-sensitive bars |
| ISO 13856-3:2006 | Safety of machinery: Pressure-sensitive protective devices, Part 3: General principles for the design and testing of pressure-sensitive bumpers, plates, wires and similar devices |
| ISO 13856-1:2001 | Safety of machinery: Pressure-sensitive protective devices, Part 1: General principles for design and testing of pressure-sensitive mats and pressure-sensitive floors |
| ISO/TR 14121-2:2007 | Safety of machinery: Risk assessment, Part 2: Practical guidance and examples of methods |
| ISO 14123-1:1998 | Safety of machinery: Reduction of risks to health from hazardous substances emitted by machinery, Part 1: Principles and specifications for machinery manufacturers |
| ISO 14123-2:1998 | Safety of machinery: Reduction of risks to health from hazardous substances emitted by machinery, Part 2: Methodology leading to verification procedures |
| ISO/TR 18569:2004 | Safety of machinery: Guidelines for the understanding and use of safety of machinery standards |
| **TC 204: Intelligent Transport Systems** | |
| ISO 15623:2002 | Transport information and control systems: Forward vehicle collision warning systems — Performance requirements and test procedures |
| **TC 210: Quality Management and Corresponding General Aspects for Medical Devices** | |
| ISO 14969:2004 | Medical devices: Quality management systems, Guidance on the application of ISO 13485: 2003 |
| ISO 14971:2007 | Medical devices: Application of risk management to medical devices |
| **TC 212: Clinical Laboratory Testing and in Vitro Diagnostic Test Systems** | |
| ISO 15190:2003 | Medical laboratories: Requirements for safety |
| ISO 15197:2003 | In vitro diagnostic test systems: Requirements for blood-glucose monitoring systems for self-testing in managing diabetes mellitus |
| **TMB: Technical Management Board** | |
| IWA 1:2005 | Quality management systems: Guidelines for process improvements in health service organizations |
| ISO/IEC Guide 50:2002 | Safety aspects: Guidelines for child safety |
| ISO/IEC Guide 71:2001 | Guidelines for standards developers to address the needs of older persons and persons with disabilities |
| **CASCO: Committee on Conformity Assessment** | |
| ISO/IEC 17025:2005 | General requirements for the competence of testing and calibration laboratories |

**Table 18 ILO OSHMS Main Sections and Their Elements**

| Section | Elements |
|---|---|
| Policy | 3.1. Occupational safety and health policy |
| | 3.2. Worker participation |
| Organizing | 3.3. Responsibility and accountability |
| | 3.4. Competence and training |
| | 3.5. OSH management system documentation |
| | 3.5. Communication |
| Planning and implementation | 3.6. Initial review |
| | 3.7. System planning and implementation |
| | 3.8. Occupational safety and health objectives |
| | 3.9. Hazard prevention |
| Evaluation | 3.10. Performance monitoring and measurement |
| | 3.11. Investigation of work-related incidents and their impact on OSH performance |
| | 3.12. Audit |
| | 3.13. Management review |
| Action for improvement | 3.15. Preventive and corrective action |
| | 3.16. Continual improvement |

*Source:* ILO (2001).

ergonomics guidelines, preferred practices, methodology, and reference data that may be needed during the design of equipment and systems have also been developed. The handbooks provide assistance in the use and application of relevant government standards during the design process.

### 5.1.1 Military Standards

The set of consensus military standards was developed by human factors engineers from the U.S. military's three services (Army, Navy, and Air Force), industry, and technical societies (McDaniel, 1996). As a result of standardization reform in the late 1990s, most of the single-service standards were canceled and were integrated into a few DOD standards and handbooks. However, the distinction between two main categories of human factors military standards—general (MIL-STD-1472 and related handbooks) and aircraft QSSG 2010 and related handbooks)—remains unchanged, which reflects the criticality of aircraft design. The list of the main military standards and handbooks are presented in Table 19.

The basic human engineering principles, design criteria, and practices required for integration of humans with systems and facilities are established in MIL-STD-1472F, *Human Engineering Design Criteria for Military*

*Systems, Equipment and Facilities.* This standard document can be applied to the design of all systems, subsystems, equipment, and facilities, not only military but commercial as well. MIL-STD-1472F includes requirements for displays, controls, control–display integration, anthropometry, ground workspace design, environment, design for maintainability, design of equipment for remote handling, small systems and equipment, operational and maintenance ground/shipboard vehicles, hazards and safety, aerospace vehicle compartment design requirements, and human–computer interface. MILSTD-1472 also includes nongovernmental standards ANSI/Human Factors Society (HFS) 100 on VDT workstations. After standardization reform the design data and information part of MIL-STD-1472F were removed and inserted into MIL-HDBK-759.

Another important military standard document is MIL-HDBK-46855, *Human Engineering Requirements for Military Systems Equipment and Facilities.* This handbook presents human engineering program tasks, procedures, and preferred practices. MIL-HDBK-46855 covers such topics as analysis functions, including human performance parameters, equipment capabilities, and task environments design; test and evaluation: workload analysis; dynamic simulation; and data requirements. This handbook also adopted materials from DOD-HDBK-763, *Human Engineering Procedures Guide,* concerned with human engineering methods and tools, which remained stable over time. The newest rapidly evolving automated human engineering tools are not described in MIL-HDBK-46855 but can be found at Directory of Design Support Methods (DSSM) at http://www.dtic.mil/dticasd/ddsm/index.html.

Other military standards cover such topics as standard practice for conducting system safety (MIL-STD-882D); acoustical noise limits, testing requirements, and measurement techniques (MIL-STD-1474D); physical characteristics of symbols for army system displays (MIL-STD-1477C); and symbology requirements for aircraft displays (MILSTD-1787C). The definitions for all human factors standard documents are provided in MIL-HDBK-1908B, *Department of Defense Handbook: Definitions of Human Factors Terms.*

### 5.1.2 Other Government Standards

The lists of other government standards are provided in Table 20. National Aeronautics and Space Administration (NASA) STD-3000 provides generic requirements for space facilities and related equipment important for proper human–system integration. This document is integrated with the website, which also offers video images from space missions that illustrate human factors design issues. This standard document is not limited to any specific NASA, military, or commercial program and can be applied to almost any type of equipment. NASA-STD-3000 consists of two volumes: Volume 1, *Man–Systems Integration Standards,* presents all of the design standards and requirements, and Volume II, *Appendices,* contains the background information related to standards. NASA-STD-3000 covers the following areas of human factors: anthropometry

**Table 19  Military Standards and Handbooks for Human Factors and Ergonomics**

| Document Number | Title | Date | Source |
|---|---|---|---|
| **Standards** | | | |
| MIL-STD-882D | Standard practice for system safety | 2000 | https://assist.daps.dla.mil/quicksearch/basic_profile. cfm?ident_number=36027 |
| MIL-STD-1472F | Human engineering | 2003 | https://assist.daps.dla.mil/quicksearch/basic_profile. cfm?ident_number=36903 |
| MIL-STD-1474D | Noise limits | 1997 | https://assist.daps.dla.mil/quicksearch/basic_profile. cfm?ident_number=36905 |
| MIL-STD-1477C | Symbols for army systems displays | 2009 | https://assist.daps.dla.mil/quicksearch/basic_profile. cfm?ident_number=69268 |
| MIL-STD-1787C | Aircraft display symbology | 2001 | Controlled distribution document |
| **Handbooks** | | | |
| DOD-HDBK-743A | Anthropometry of U.S. military personnel | 1991 | https://assist.daps.dla.mil/quicksearch/basic_profile. cfm?ident_number=54083 |
| MIL-HDBK-759C | Human engineering design guidelines | 1998 | https://assist.daps.dla.mil/quicksearch/basic_profile. cfm?ident_number=54086 |
| MIL-HDBK-767 | Design guidance for interior noise reduction in light-armored tracked vehicles | 2004 | https://assist.daps.dla.mil/quicksearch/basic_profile. cfm?ident_number=112000 |
| MIL-HDBK-1473A | Color and marking of army materiel | 1997 | https://assist.daps.dla.mil/quicksearch/basic_profile. cfm?ident_number=200029 |
| MIL-HDBK-1908B | Definitions of human factors terms | 2004 | https://assist.daps.dla.mil/quicksearch/basic_profile. cfm?ident_number=121264 |
| MIL-HDBK-46855 | Human engineering requirements for military systems equipment and facilities | 2004 | https://assist.daps.dla.mil/quicksearch/basic_profile. cfm?ident_number=201925 |

and biomechanics, human performance capabilities, natural and induced environments, health management, workstations, activity centers, hardware and equipment, design for maintainability, and facility management.

Standards of the Federal Aviation Administration (FAA) are concerned with the following topics: human factors design criteria oriented to the FAA mission and systems (HF-STD-001); design and evaluation of air traffic control systems (DOT-VNTSC-FAA-95-3); elements of the human engineering program (FAA-HF-001); evaluation of human factors criteria conformance of equipment that interface with the operator (FAA-HF-002) and with the maintainer (FAA-HF-003).

In their standard DOE-HDBK-1140-2001, the DOE provides the system maintainability design criteria for DOE systems, equipment, and facilities. The Federal Highway Administration (FHA) establishes standards concerning the development and operation of traffic management centers (FHWA-JPO-99-042). The FHA also describes human factors guidelines and recommendations for design of advanced traveler information systems (ATISs), commercial vehicle operations (CVOs), and accommodation of older drivers and pedestrians. The NRC provides guidelines of HFE conformance evaluation of the interface design of nuclear power plant systems (NUREG—0700 and NUREG—0711). FED-STD-795, which has been developed for use in federal

and federally funded facilities, establishes standards for facility accessibility by physically handicapped persons.

## 5.2  OSHA Standards

Development of occupational safety and health standards in the United States is mandated by the general duty clause, Section 5(a)(1), of the *Occupational Safety and Health Act of 1970,* which states: "Each employer shall furnish to each of his employees, employment and a place of employment which is free from recognized hazards that are causing or are likely to cause death or serious harm to his employees." In general, penalties related to deficient and unsafe working conditions have been issued under this general duty clause. The general duty clause has also been supplemented by the *Americans with Disabilites Act* (ADA, Public Law 101-336, 1990). The disabilities act has an important bearing on ergonomics design of workplaces. The ADA prohibits disability-based discrimination in hiring practices and requires that all employers make reasonable accommodations to working conditions to allow qualified disabled workers to perform their job functions.

In 1990, the Occupational Safety and Health Administration (OSHA) issued a set of voluntary guidelines entitled *Ergonomics Program Management Guidelines for Meatpacking Plants* (OSHA 3123), which have been used successfully by many types of industries, including those from outside the food production business.

**Table 20 U.S. Government Human Factors/Ergonomics Standards**

| Document Number | Title | Date | Source |
|---|---|---|---|
| *National Aeronautics and Space Administration* | | | |
| NASA-STD-3000B | Man–systems integration standards | 1995 | http://msis.jsc.nasa.gov |
| *Department of Transportation, Federal Aviation Administration* | | | |
| HF-STD-001 | Human factors design standard | 2003 | http://www.hf.faa.gov/docs/508/docs/wjhtc/hfds.zip |
| DOT-VNTSC-FAA-95-3 | Human factors in the design and evaluation of air traffic control systems | 1995 | http://www.hf.faa.gov/docs/volpehndk.zip |
| FAA-HF-001 | Human engineering program plan | 1999 | http://www.hf.faa.gov/docs/did_001.htm |
| FAA-HF-002 | Human engineering design approach document — Operator | 1999 | http://www.hf.faa.gov/docs/did_002.htm |
| FAA-HF-003 | Human engineering design approach document — Maintainer | 1999 | http://www.hf.faa.gov/docs/did_003.htm |
| FAA-HF-004 | Critical task analysis report | 2000 | http://hfetag.dtic.mil/docs-hfs/faa-hf004 critical_ taskanalysis_report.doc |
| FAA-HF-005 | Human engineering simulation concept | 2000 | http://hfetag.dtic.mil/docs-hfs/faa-hf-005_human-engineering_simulation.doc |
| *Department of Transportation, Federal Highway Administration* | | | |
| FHWA-JPO-99-042 | Preliminary human factors guidelines for traffic management centers | 1999 | http://www.fhwa.dot.gov/publications/research/safety/99042/index.cfm |
| FHWA-RD-98-057 | Human factors design guidelines for advanced traveler information systems (ATIS) and commercial vehicle operations (CVO) | 1998 | http://www.fhwa.dot.gov/tfhrc/safety/pubs/atis/ |
| FHWA-RD-01-051 | Guidelines and recommendations to accommodate older drivers and pedestrians | 2001 | http://www.fhwa.dot.gov/publications/research/safety/humanfac/01051/refsrecom.cfm |
| FHWA-RD-01-103 | Highway design handbook for older drivers and pedestrians | 2001 | http://www.fhwa.dot.gov/publications/research/safety/humanfac/01103/ |
| *Department of Energy* | | | |
| DOE-HDBK-1140-2001 | Human factors/ergonomics handbook for the design for ease of maintenance | 2001 | http://www.hss.energy.gov/nuclearsafety/ns/techstds/standard/hdbk1140/hdbk1140.html |
| *Multiple Departments* | | | |
| FED-STD-795 | Uniform federal accessibility standards | 1988 | http://www.assistdocs.com/search/document_details.cfm?ident_number=53835&StartRow=86851&PaginatorPageNumber=1738&status_all=ON&search_method=BASIC%20US%20Department%20of%20Defense%20specification%203443 |

In 2000, the U.S. government proposed the *Ergonomics Program Rule (Federal Register,* November 14, 2000, Vol. 65, No. 220). The main elements of the standard included (1) training in basic ergonomics awareness, (2) providing medical management of work-related musculoskeletal disorders, (3) implementing a quick fix or going to a full program, and (4) implementing a full ergonomic program when indicated, including such elements as management leadership, employee participation, job hazard analysis, hazard reduction and control, training, and program evaluation. However, the regulation was repealed in March 2001.

Recently, OSHA has developed a four-pronged comprehensive approach to ergonomics designed to address musculoskeletal disorders (MSDs) in the workplace. The four segments of the OSHA's strategy were stated as follows:

1. *Guidelines:* to develop industry- or task-specific guidelines for industries based on current incidence rates and available information about effective and feasible solutions

2. *Enforcement:* to conduct inspections for ergonomic hazards and issue citations under the general duty clause and to issue ergonomic hazard alert letters where appropriate

3. *Outreach and assistance:* to provide assistance to businesses, particularly small businesses, and

help them proactively address ergonomic issues in the workplace

4. *National advisory committee:* to charter an advisory committee that will be authorized to, among other things, identify gaps in research to the application of ergonomics and ergonomic principles in the workplace

Recently, OSHA has also published three voluntary guidelines to assist employers of the specific type of industries in recognizing and controlling hazards: (1) *Nursing Home Guideline* (2003), (2) *Draft Guideline for Poultry Processing* (2003), and (3) *Guideline for the Retail Grocery Industry* (2004). The objective of this protocol is to establish a fair and transparent process for developing industry- and task-specific guidelines that will assist employers and employees in recognizing and controlling potential ergonomic hazards. By using this protocol, each set of guidelines will address a particular industry or task. It is intended that the industry- and task-specific guidelines will generally be presented in three major parts:

1. Program management recommendations for management practices addressing ergonomic hazards in the industry or task

2. Work site analysis recommendations for work site/workstation analysis techniques geared to the specific operations that are present in the industry or task

3. Hazard control recommendations that contain descriptions of specific jobs and that detail the hazards associated with the operation, possible approaches to controlling the hazard, and the effectiveness of each control approach

Since there are many different types of work-related hazards and injuries, and controls vary from industry to industry and task to task, OSHA expects that the scope and content of the guidelines will vary. Thus, OSHA enforces a general industry-related regulation [29 Code of Federal Regualtions (CFR) 1910] and two industry specific regulations: (1) construction (29 CFR 1926) and (2) maritime (29 CFR 1915, 1917, and 1918). Through the directorate of enforcement programs, OSHA ensures the implementation of these regulations.

### 5.3   Other Standards for Occupational Safety and Health

In 2000, the National Safety Council (NSC), acting on behalf of the Accredited Standards Committee (ASC), has issued a draft document (known as Z-365) entitled *Management of Work-Related Musculoskeletal Disorders* (*MSD*). The draft defines the following areas of importance to preventing work-related injuries: (1) management responsibility, (2) employee involvement, (3) training, (4) surveillance, (5) evaluation and management of work-related MSD cases, (6) job analysis and design, and (7) follow-up.

Independent of the efforts noted above, in 2001 another ANSI committee, *ASC Z-10, Occupational*

*Health Safety Systems,* was formed under the auspices of the American Industrial Hygiene Association (AIHA). The main objective of ASC Z-10 is to develop a standard of management principles and systems for improving the occupational safety and health in companies.

### 5.4   ANSI Standards

The following HFE-relevant standards have been developed by the ANSI.

#### 5.4.1   Human Factors Engineering of VDTs

ANSI/HFS 100-1988 presents ergonomics principles related to visual display terminals. The standard has been updated as ANSI/Human Factors and Ergonomics Society (HFES) 100-2007 and is available at http://www. hfes.org/publications/ProductDetail.aspx?ProductId=69.

#### 5.4.2   Human Factors Engineering of Computer Workstations

As stated in the trial use, the BSR/HFES 100 (currently ANSI/HFES 100-2007) Human Factors Engineering of Computer Workstations (HFES 100) is *a specification of the recommended human factors and ergonomic principles related to the design of the computer workstation, and is intended for fixed, office-type computer workstations for individuals who are moderate to intensive computer users* (Albin, 2006). This standard is organized into four major chapters: (1) installed systems, (2) input devices, (3) visual displays, and (4) furniture. The installed systems chapter specifies how to arrange all the workstation system components to match the capabilities of the intended user. The input devices chapter focuses on the design of input devices (including the issues of physical size, operation force, handedness, etc). The visual displays chapter discusses the human factors in the design of monochrome and color CRT and flat-panel displays. The furniture chapter provides design specifications for workstation components, including chairs and desks. The major topics described in each of these chapters are listed in Table 21.

#### 5.4.3   Ergonomic Requirements for Software User Interfaces

The HFES/HCI 200 Committee, which operates under the auspices of the HFES Technical Standards Committee, has been working on development of a proposed U.S. national standard for software user interfaces. This standard, updated in 2005, provides requirements and recommendations for software interfaces, with a primary focus on business and personal computing applications. The standard is related to the ISO 9241 series of user interface standards. The topics described in each section of the HFES 200 standard are listed in the Table 22.

#### 5.4.4   Ergonomic Guidelines for the Design, Installation, and Use of Machine Tools

ANSI B 11, *Technical Report: Ergonomic Guidelines for the Design, Installation and Use of Machine Tools,* is a consensual ergonomic guidelines developed by the Machine Tool Safety Standards Committee (B 11) of the

**Table 21  Main Chapters and Topics of the Human Factors Engineering of Computer Workstations: ANSI/HFES 100-2007 Standard**

| Chapter | Topics |
| --- | --- |
| Installed systems | Hardware components, noise, thermal comfort, and lighting |
| Input devices | Keyboards, mouse and puck devices, trackballs, joysticks, styluses and light pens, tablets and overlays, touch-sensitive panels |
| Visual displays | Monochrome and color cathode ray tube (CRT) and flat-panel displays (viewing characteristics, contrast, legibility, etc.) |
| Furniture | Specifications for workstation components (chairs, desks, etc.); postures (reference postures, reclined sitting, upright sitting, declined sitting and standing); anthropometry |

**Table 22  Topics Addressed in the Ergonomic Requirements for Software User Interfaces: HFES 200-2005**

| Chapter | Topics |
| --- | --- |
| Accessibility | Keyboard input; multiple keystrokes |
| | Customization; repeat rates; acceptance delays |
| | Pointer alternative; accelerators; remapping; navigation |
| | *Display fonts:* size, legibility, styles, colors |
| | *Audio output:* volume and frequencies, customization, content and alerts, graphics |
| | *Color:* palettes, background–foreground, customization, coding |
| | *Errors and persistence:* online documentation and help |
| | *Customization:* cursor, button presses, click interval, pointer speed, chording |
| | *Window appearance and behavior:* navigation and location, window focus, titles |
| | *Input focus:* navigation, behavior, order, location |
| Color | Color *selection:* chromostereopsis, blending and depth effects, use of blue and red, identification and contrast |
| | *Color assignments:* conventions, uniqueness and reuse, naming, cultural assignments |
| | *General use consideration:* number of colors, highlighting, positioning and separation |
| | *Special uses:* warnings, coding, state indications, pointers, area identification |
| Voice and telephony | *Speech recognition (input):* commands, vocabularies, prompts, consistency, feedback, error handling, dictation |
| | *Speech output:* vocabularies message format, speech characteristics, dialogue techniques, physical properties, alerting tones, stereophonic presentation |
| | *Nonspeech auditory output:* consistency, tone format, critical messages, frequency, amplitude |
| | Interactive voice response |
| Technical sections | Presentation of information, user guidance, menu dialogues, command dialogues, direct manipulation, dialogue boxes, and form-filling dialogue windows |

ANSI. The subcommittee responsible for the preparation of these guidelines consisted of representatives from manufacturing, higher education, safety, design, and ergonomics. The document specifies ergonomic guidelines to assist in the design, installation, and use of individual and integrated machine tools and auxiliary components in manufacturing systems.

The guidelines document underlines the importance of three basic ideas for achievement of effective and safe design, installation, and use of machine tools: (1) communication among all persons involved with the machine tools (users, installers, manufacturers, and designers), (2) dissemination of knowledge concerning ergonomics concepts and principles among all persons, and (3) the ability to apply ergonomics concepts and principles effectively to machine tools and auxiliary components. The guidelines document states that the provision of worker safety, work efficiency, and optimization of the entire production system requires consideration of the following ergonomics issues:

- The variation in employee physiological and psychological characteristics such as strength and capacity
- Incorporation of ergonomics concepts and principles into all new project, tool, machine, and work processes at the beginning of the process
- The goal that routine tasks that are to be done precisely, rapidly, and continuously, especially tasks in hazardous environments, should be performed by machines
- The goal that tasks that require judgment and integration of information (i.e., the tasks that humans do best) should be assigned to workers
- The knowledge that a system that does not consider human limits such as information handling,

perception, reach, clearance, posture, or strength exertion can predispose to accident or injury

The documents also recommend matching the design of the tool or process with the physical characteristics and capabilities of workers, to ensure accommodation, compatibility, operability, and maintainability of the machine tools and/or auxiliary components. A complete list of standards included in B11 is shown in Table 23.

**Table 23 ANSI B11 Parts**

| Standard | Title |
| --- | --- |
| ANSI B11-2008 | General safety requirements common to ANSI B11 machines |
| ANSI B11.1 | Mechanical power presses |
| ANSI B11.2 | Hydraulic power presses |
| ANSI B11.3 | Mechanical power press brakes |
| ANSI B11.4 | Shears |
| ANSI B11.5 | Ironworkers |
| ANSI B11.6 | Lathes |
| ANSI B11.7 | Cold headers and cold formers |
| ANSI B11.8 | Drilling, milling, and boring machines |
| ANSI B11.9 | Grinding machines |
| ANSI B11.10 | Metal sawing machines |
| ANSI B11.11 | Gear and spline cutting machines |
| ANSI B11.12 | Roll forming and roll bending machines |
| ANSI B11.13 | Automatic screw/bar and chucking machines |
| ANSI B11.14 | Withdrawn (now see B11.18) |
| ANSI B11.15 | Pipe, tube, and shape bending machines |
| ANSI B11.16 | Metal powder compacting presses |
| ANSI B11.17 | Horizontal hydraulic extrusion presses |
| ANSI B11.18 | Coil processing systems |
| ANSI B11.19 | Performance criteria for safeguarding |
| ANSI B11.20 | Integrated manufacturing systems |
| ANSI B11.21 | Machine tools using lasers for processing materials |
| ANSI B11.22 | Turning centers and CNC turning machines |
| ANSI B11.23 | Machining centers and CNC milling, drilling, and boring machines |
| ANSI B11.24 | Transfer machines |
| ANSI B11.TR1-2004 | Ergonomic guidelines |
| ANSI B11.TR2-1997 | Mist control considerations |
| ANSI B11.TR3-2000 | Risk assessment and risk reduction |
| ANSI B11.TR4-2004 | Selection of programmable electronic systems (PES/PLC) for machine tools |
| ANSI B11.TR5-2006 | Sound level measurement guidelines |
| ANSI B11.TR6-20XX (not an approved document; in development) | Control reliable circuits and servo drive technology |
| ANSI B11.TR7-2007 | Design for safety and lean manufacturing |

## 5.5 State-Mandated Occupational Safety and Health Standards

Section 18 of the Occupational Safety and Health Act (1970) encourages states to develop and operate their own job safety and health programs. In general, states with OSHA-approved occupational safety and health programs may follow OSHA's approach to ergonomics: to adopt ergonomic standards, include ergonomics in standards establishing safety and health program requirements, and utilize the general duty authority for enforcement purposes (Seabrook. 2001; Stuart-Buttle, 2005). To date, 22 stes and jurisdictions are operating complete state plans that cover both the private sector and state and local government employees and five states, that is, Connecticut, Illinois, New Jersey, New York, and the Virgin Islands cover public employees only. Of the states and jurisdictions, seven (California, Michigan, New Mexico, Oregon, Puerto Rico, Vermont, and Washington) have the operational status agreement, that is, the state is independently capable of enforcing its own standard. Out of those seven, only Oregon has received the final approval of OSHA in 2005. However, this final approval does not include temporary labor camps in agriculture, general industry, construction, and logging. Nine states that previously had final approval of an independent state program withdrew their programs and currently use OSHA funded on-site consultation programs. Details of state-mandated occupational safety and health programs can be found at http://www.osha.gov/dcsp/osp/index.html.

## 5.6 Other Standardization Efforts

The American Conference of Governmental Industrial Hygienists (ACGIH) (www.aegih.org) established threshold limit values (TLVs) for the following physical categories of work: (1) hand–arm and whole-body vibration, (2) thermal stress, (3) hand activity level ("monotask" jobs, performed for 4 h or more), and (4) lifting tasks (load limits based on lift frequency, task duration, horizontal distance, and height at the start of the lift). Other organizations that develop HFE-related standards include the American Society of Mechanical Engineers (ASME), American Society for Testing and Materials (ASTM), Institute of Electrical and Electronics Engineers (IEEE), Society of Automotive Engineers (SAE), and National Institute of Standards and Technology (NIST, www.nist.gov).

A notable use of standards is the SAE HFE standards widely used in the US automotive industry. Table 24 shows the various standards and guidelines that are in use currently by the various automotive manufacturing companies in the United States.

## 6 ISO 9000-2005: QUALITY MANAGEMENT STANDARDS

Quality standards can also play an important role in assuring safety and health at the workplace. The ISO stipulates that if a quality management system is implemented appropriately utilizing the eight quality management principles (see below) and in accordance with

**Table 24  SAE Technical Automotive Standards for Human Factors and Ergonomics**

| SAE Standard | Title | Issuing Committee | Date Published |
|---|---|---|---|
| J1012 | Operator enclosure pressurization system test procedure | Hftc6, Operator Accommodation | 2008-07-17 |
| J1050 | Describing and measuring the driver's field of view | Driver Vision Standards Committee | 2009-02-13 |
| J1052 | Motor vehicle driver and passenger head position | Human Accommodation and Design Devices Standards Committee | 2010-09-30 |
| J1083 | Unauthorized starting or movement of machines | Optc1, Personnel Protection (General) | 2002-12-18 |
| J1138 | Design criteria — Driver hand controls location for passenger cars, multipurpose passenger vehicles, and trucks (10 000 GVW and under) | Controls and Displays Standards Committee | 2009-08-25 |
| J1139 | Direction-of-motion stereotypes for automotive hand controls | Controls and Displays Standards Committee | 2010-03-01 |
| J1257 | Rating chart for cantilevered boom cranes | Cranes and Lifting Devices Committee | 2002-07-11 |
| J128 | Occupant restraint system evaluation — Passenger cars and light-duty trucks | Restraint Systems Standards Steering Committee | 2008-11-25 |
| J1281 | Operator sound pressure level exposure measurement procedure for powered recreational craft | Marine Technical Steering Committee | 2009-08-21 |
| J1289 | Mobile crane stability ratings | Cranes and Lifting Devices Committee | 2002-07-11 |
| J1305 | Two-block warning and limit systems in lifting crane service | Cranes and Lifting Devices Committee | 2007-11-16 |
| J1308 | Fan guard for off-road machines | Optc1, Personnel Protection (General) | 2008-08-05 |
| J1460/1 | Human mechanical impact response characteristics — Dynamic response of the human abdomen | Human Biomechanics and Simulations Standards Steering Committee | 2000-11-28 |
| J1460/2 | Human mechanical impact response characteristics — Response of the human neck to inertial loading by the head for automotive seated postures | Human Biomechanics and Simulations Standards Steering Committee | 2008-06-17 |
| J1516 | Accommodation tool reference point | Truck and Bus Human Factors Committee | 2009-11-06 |
| J1517 | Driver selected seat position | Truck and Bus Human Factors Committee | 2009-11-06 |
| J1521 | Truck driver shin-knee position for clutch and accelerator | Truck and Bus Human Factors Committee | 2009-02-10 |
| J1522 | Truck driver stomach position | Truck and Bus Human Factors Committee | 2009-02-10 |
| J153 | Operator precautions | Hftc6, Operator Accommodation | 2005-07-11 |
| J1533 | Operator enclosure air filter element test procedure | Hftc6, Operator Accommodation | 2005-02-24 |
| J1559 | Determination of effect of solar heating | Hftc6, Operator Accommodation | 2003-01-09 |
| J1574/1 | Measurement of vehicle and suspension parameters for directional control studies | Vehicle Dynamics Standards Committee | 2005-05-09 |
| J1606 | Headlamp design guidelines for mature drivers | Road Illumination Devices Standards Committee | 1997-10-01 |
| J1663 | Truth-in-labeling standard for navigation map databases | Motor Vehicle Council | 2003-10-02 |
| J1725 | Structural modification for personally licensed vehicles to meet the transportation needs of persons with disabilities | Adaptive Devices Standards Committee | 2010-01-04 |
| J1750 | Describing and evaluating the truck driver's viewing environment | Truck and Bus Human Factors Committee | 2010-10-22 |
| J182 | Motor vehicle fiducial marks and three-dimensional reference system | Human Accommodation and Design Devices Standards Committee | 2009-07-23 |
| J1903 | Automotive adaptive driver controls, manual | Adaptive Devices Standards Committee | 2010-01-04 |

**Table 24** (*Continued*)

| SAE Standard | Title | Issuing Committee | Date Published |
|---|---|---|---|
| J1980 | Guidelines for evaluating out-of-position vehicle occupant interactions with deploying frontal airbags | Human Biomechanics and Simulations Standards Steering Committee | 2008-06-17 |
| J209 | Instrument face design and location for construction and industrial equipment (cancelled May 2003) | Hftc2, Machine Displays and Symbols | 2003-05-13 |
| J2092 | Testing of wheelchair lifts for entry to or exit from a personally licensed vehicle | Adaptive Devices Standards Committee | 1999-11-01 |
| J2094 | Vehicle and control modifications for drivers with physical disabilities terminology | Adaptive Devices Standards Committee | 2010-03-15 |
| J2114 | dolly rollover recommended test procedure | Impact and Rollover Test Procedure Standards Committee | 1999-10-01 |
| J2119 | Manual controls for mature drivers | Mature Driver Standards Committee | 1997-10-01 |
| J2217 | Photometric guidelines for instrument panel displays that accommodate older drivers | Mature Driver Standards Committee | 1991-10-01 |
| J2388 | Secondary control modifications | Adaptive Devices Standards Committee | 2009-02-24 |
| J2396 | Definitions and experimental measures related to the specification of driver visual behavior using video based techniques | Safety and Human Factors Steering Committee | 2005-11-29 |
| J2400 | Human factors in forward collision warning systems: operating characteristics and user interface requirements | Safety and Human Factors Steering Committee | 2009-11-17 |
| J2732 | Motor vehicle seat dimensions | Human Accommodation and Design Devices Standards Committee | 2008-06-23 |
| J287 | Driver hand control reach | Human Accommodation and Design Devices Standards Committee | 2007-02-27 |
| J2896 | Motor vehicle seat performance measures | Human Accommodation and Design Devices Standards Committee | 2009-04-01 |
| J375 | Radius-of-load or boom angle indicating systems | Cranes and Lifting Devices Committee | 1994-04-01 |
| J4004 | Positioning the H-point design tool — Seating reference point and seat track length | Human Accommodation and Design Devices Standards Committee | 2008-08-29 |
| J670 | Vehicle dynamics terminology | Vehicle Dynamics Standards Committee | 2008-01-24 |
| J826 | Devices for use in defining and measuring vehicle seating accommodation | Human Accommodation and Design Devices Standards Committee | 2008-11-11 |
| J881 | Lifting crane sheave and drum sizes | Cranes and Lifting Devices Committee | 2003-06-02 |
| J919 | Sound measurement — Off-road work machines — Operator — Singular type | Sltc, Earth Moving Machinery Sound Level | 2009-01-13 |
| J941 | Motor vehicle drivers' eye locations | Driver Vision Standards Committee | 2010-03-16 |
| J98 | Personnel protection for general purpose industrial machines | Optc1, Personnel Protection (General) | 2007-06-05 |
| J983 | Crane and cable excavator basic operating control arrangements | Cranes and Lifting Devices Committee | 1998-10-01 |
| J985 | Vision factors considerations in rearview mirror design | Driver Vision Standards Committee | 2009-02-13 |
| J999 | Crane boom hoist disengaging device | Cranes and Lifting Devices Committee | 1998-07-01 |

ISO 9004, all of an organization's interested parties should benefit. For example, people in the organization will benefit from (1) improved working conditions, (2) increased job satisfaction, (3) improved health and safety, (4) improved morale, and (5) improved stability of employment, and the society at large will benefit from (1) fulfillment of legal and regulatory requirements, (2) improved health and safety, (3) reduced environmental impact, and (4) increased security.

As discussed by Hoyle (2009), the term ISO 9000 refers to a set of quality management standards. ISO 9000 currently includes three quality standards: ISO 9000:2005, ISO 9001:2008, and ISO 9004:2009. ISO 9001:2008 presents requirements; ISO 9000:2005

and ISO 9004:2009 present guidelines. The ISO first published its quality standards in 1987, revised them in 1994 and 2000, and then republished an updated version in 2005. These new standards are referred to as the ISO 9000:2005 standards.

ISO 9000:2005 describes fundamentals of quality management systems, which form the subject of the ISO 9000 family, and defines related terms. It is applicable to the following: (a) organizations seeking advantage through the implementation of a quality management system; (b) organizations seeking confidence from their suppliers that their product requirements will be satisfied; (c) users of the products; (d) those concerned with a mutual understanding of the terminology used in quality management (e.g., suppliers, customers, regulators); (e) those internal or external to the organization who assess the quality management system or audit it for conformity with the requirements of ISO 9001 (e.g., auditors, regulators, certification/registration bodies); (f) those internal or external to the organization who give advice or training on the quality management system appropriate to that organization; and (g) developers of related standards. The standard recognizes that the word *product* applies to services, processed material, and hardware and software intended for, or required by, the customer (Hoyle, 2001).

The ISO 9000:2005 standards apply to all types of organizations, including manufacturing, service, government, and education. The standards are based on eight *quality management principles* (Hoyle, 2006):

- *Principle* 1: customer focus
- *Principle* 2: leadership
- *Principle* 3: involvement of people
- *Principle* 4: process approach
- *Principle* 5: system approach to management
- *Principle* 6: continual improvement
- *Principle* 7: factual approach to decision making
- *Principle* 8: mutually beneficial supplier relationships

As part of ISO 9000, requirements for a quality management system are met by implementing ISO 9001: 2008. ISO 9001:2008 specifies requirements for a quality management system where an organization (1) needs to demonstrate its ability to consistently provide product that meets customer and applicable statutory and regulatory requirements and (2) aims to enhance customer satisfaction through the effective application of the system, including processes for continual improvement of the system and the assurance of conformity to customer and applicable statutory and regulatory requirements. All requirements of ISO 9001:2008 are generic and are intended to be applicable to all organizations, regardless of type, size, and product provided. Where any requirement(s) of ISO 9001:2008 cannot be applied due to the nature of an organization and its product, this can be considered for exclusion. Where exclusions are made,

claims of conformity to ISO 9001:2008 are not acceptable unless these exclusions are limited to requirements within Clause 7, and such exclusions do not affect the organization's ability, or responsibility, to provide product that meets customer and applicable statutory and regulatory requirements.

ISO 9004:2009 can be used to extend the benefits obtained from ISO 9001:2008 to employees, owners, suppliers, and society in general. This standard provides guidance to organizations to support the achievement of sustained success by a quality management approach. It is applicable to any organization, regardless of size, type, and activity.

ISO 9001:2008 and ISO 9004:2009 are harmonized in structure and terminology to assist an organization to move smoothly from one to the other. Both standards apply a process approach. Processes are recognized as consisting of one or more linked activities that require resources and must be managed to achieve predetermined output. The output of one process may form directly the input to the next process, and the final product is often the result of a network or system of processes. The eight quality management principles stated in ISO 9000:2005 and ISO 9004:2009 provide the basis for the performance improvement outlined in ISO 9004:2009. The ISO 9000 standards cluster also includes other 10000 series standards. Table 25 shows a list of the relevant standards and their purposes.

ISO requires that the organization determine what it needs to do to satisfy its customers, establish a system to accomplish its objectives, and measure, review, and continually improve its performance. More specifically, the ISO 9001 and 9004 requirements stipulate that an organization must:

1. Determine the needs and expectations of customers and other interested parties

2. Establish policies, objectives, and a work environment necessary to motivate the organization to satisfy these needs

3. Design, resource, and manage a system of interconnected processes necessary to implement the policy and attain the objectives

4. Measure and analyze the adequacy, efficiency, and effectiveness of each process in fulfilling its purpose and objectives

5. Pursue the continual improvement of the system from an objective evaluation of its performance

The ISO identified several potential benefits of using the quality management standards. These benefits may include the connection of quality management systems to organizational processes, encouragement of a natural progression toward improved organizational performance, and consideration of the needs of all interested parties. Along with the main quality management standards, the ISO also published standards that deal with quality management of specific products (e.g., medical devices), specific industry (e.g., ships and marine,

**Table 25  ISO 9000:2005 Quality Management Standards and Guidelines**

| Standard or Guideline | Purpose |
| --- | --- |
| ISO 9000:2005, Quality management systems: Fundamentals and vocabulary | Establishes a starting point for understanding the standards and defines the fundamental terms and definitions used in the ISO 9000 family to avoid misunderstandings in their use |
| ISO 9001:2008, Quality management systems: Requirements | Requirement standard to be used to assess the organization's ability to meet customer and applicable regulatory requirements and thereby address customer satisfaction: now the only standard in the ISO 9000 family against which third-party certification can be carried |
| ISO 9004:2009, Managing for the sustained success of an organization: A quality management approach | Provides guidance for continual improvement of an organization's quality management system to benefit all parties through sustained customer satisfaction |
| ISO 19011:2002, Guidelines on quality and/or environmental management systems auditing | Provides an organization with guidelines for verifying the system's ability to achieve defined quality objectives (use internally or for auditing suppliers) |
| ISO 10005:2005, Quality management: Guidelines for quality plans | Provides guidelines to assist in the preparation, review, acceptance, and revision of quality plans |
| ISO 10006:2003, Quality management: Guidelines to quality management in projects | Guidelines to help the organization to ensure the quality of both project processes and project products |
| ISO 10007:2003, Quality management: Guidelines for configuration management | Gives an organization guidelines to ensure that a complex product continues to function when components are changed individually |
| ISO 10012:2003, Measurement management systems: Requirements for measurement processes and measuring equipment | Specifies generic requirements and provides guidance for the management of measurement processes and metrological confirmation of measuring equipment used to support and demonstrate compliance with metrological requirements |
| ISO/TR 10013:2001, Guidelines for quality management system documentation | Provides guidelines for the development and maintenance of quality manuals tailored to specific needs |
| ISO 10014:2006, Guidelines for realizing financial and economic benefits | Provides guidance on how to achieve economic benefits from the application of quality management |
| ISO 10015:1999, Quality management: Guidelines for training | Provides guidance on the development, implementation, maintenance, and improvement of strategies and systems for training that affects the quality of products |
| ISO/TS 16949:2009, Quality management systems: Particular requirements for the application of ISO 9001:2008 for automotive production and relevant service part organizations | Provides sector-specific guidance to the application of ISO 9001 in the automotive industry |

petroleum, software engineering, information technology, systems engineering, intelligent transport systems), environment, and customer satisfaction. Those standards are listed in Table 26.

## 7  CONCLUSIONS

Although human factor and ergonomics standards cannot guarantee appropriate workplace design, they can provide clear and well-defined requirements and guidelines and therefore the basis for good ergonomics design. Standards for workstation design and the work environment can ensure the safety and comfort of working people through establishing requirements for optimal working conditions. By providing consistency in the human–system interface and improving ergonomics quality of the interface components, ergonomics

standards can also contribute to the enhanced systems usability and overall system performance. This benefit is based on the general requirement of harmonization across different tools and systems to support user performance and avoid unnecessary human errors.

One of the most important benefits from standardization efforts is a formal recognition of the significance of ergonomics requirements and guidelines for system design on the national and international levels (Harker, 1995). The consensus procedure applied to standards development demands consultation with a wide range of commercial, professional, and industrial organizations. Therefore, the decision to develop standards and a consensus of diverse organizations concerning the need for standards reflects the formal recognition that there are important human factors and ergonomics issues that need to be taken into account during the design and development of workplaces and systems. In the last few

**Table 26 Other Quality Management Standards and Guidelines**

| Standard Reference | Title |
| --- | --- |
| ISO 10001:2007 | Quality management: Customer satisfaction: Guidelines for codes of conduct for organizations |
| ISO 10002:2004 | Quality management: Customer satisfaction: Guidelines for complaints handling in organizations |
| ISO 10003:2007 | Quality management: Customer satisfaction: Guidelines for dispute resolution external to organizations |
| ISO 10019:2005 | Guidelines for the selection of quality management system consultants and use of their services |
| ISO 13485:2003 | Medical devices: Quality management systems: Requirements for regulatory purposes |
| ISO 14001:2004 | Environmental management systems: Requirements with guidance for use |
| ISO 14004:2004 | Environmental management systems: General guidelines on principles, systems and support techniques |
| ISO 14050:2009 | Environmental management: Vocabulary |
| ISO 14063:2006 | Environmental management: Environmental communication: Guidelines and examples |
| ISO 15189:2007 | Medical laboratories: Particular requirements for quality and competence |
| ISO 15225:2010 | Medical devices: Quality management: Medical device nomenclature data structure |
| ISO 19379:2003 | Ships and marine technology: ECS databases: Content, quality, updating and testing |
| ISO 20815:2008 | Petroleum, petrochemical and natural gas industries: Production assurance and reliability management |
| ISO 22006:2009 | Quality management systems: Guidelines for the application of ISO 9001:2008 to crop production |
| ISO 27025:2010 | Space systems: Programme management: Quality assurance requirements |
| ISO 29861:2009 | Document management applications: Quality control for scanning office documents in colour |
| ISO/IEC 17021:2006 | Conformity assessment: Requirements for bodies providing audit and certification of management systems |
| ISO/IEC 19796-3:2009 | Information technology: Learning, education and training: Quality management, assurance and metrics, Part 3: Reference methods and metrics |
| ISO/IEC 23004-4:2007 | Information technology: Multimedia Middleware, Part 4: Resource and quality management |
| ISO/IEC 25001:2007 | Software engineering: Software product Quality Requirements and Evaluation (SQuaRE): Planning and management |
| ISO/IEC 27001:2005 | Information technology: Security techniques: Information security management systems, requirements |
| ISO/IEC 27003:2010 | Information technology: Security techniques: Information security management system implementation guidance |
| ISO/IEC 27004:2009 | Information technology: Security techniques: Information security management, measurement |
| ISO/IEC 27005:2008 | Information technology: Security techniques: Information security risk management |
| ISO/IEC 27006:2007 | Information technology: Security techniques: Requirements for bodies providing audit and certification of information security management systems |
| ISO/IEC 90003:2004 | Software engineering: Guidelines for the application of ISO 9001:2000 to computer software |
| ISO/IEC Guide 53:2005 | Conformity assessment: Guidance on the use of an organization's quality management system in product certification |
| ISO/IEC TR 90005:2008 | Systems engineering: Guidelines for the application of ISO 9001 to system life cycle processes |
| ISO/TR 14969:2004 | Medical devices: Quality management systems: Guidance on the application of ISO 13485: 2003 |
| ISO/TR 21707:2008 | Intelligent transport systems: Integrated transport information, management and control, data quality in ITS systems |
| ISO/TS 10004:2010 | Quality management: Customer satisfaction: Guidelines for monitoring and measuring |
| ISO/TS 14048:2002 | Environmental management: Life cycle assessment, data documentation format |
| ISO/TS 29001:2010 | Petroleum, petrochemical and natural gas industries: Sector-specific quality management systems: Requirements for product and service supply organizations |
| IWA 2:2007 | Quality management systems: Guidelines for the application of ISO 9001:2000 in education |
| IWA 4:2009 | Quality management systems: Guidelines for the application of ISO 9001:2008 in local government |

years, we have seen a greater effort on the part of the ISO and CEN in publishing the underdeveloped standards into full standards. At the same time, the ISO is making every effort to outreach, in particular to developing and underdeveloped countries. Despite these efforts, there is a lack of research endeavor examining or investigating the effectiveness of these standards. There is also the question of whether organizations will feel implementing a large number of standards a burden, specially the small business entrepreneurs.

Nevertheless, standards represent the essence of the best available knowledge and practice extracted from a variety of academic sources, presented in the way that is easy to use by professional designers, and to include this knowledge in the design process. The consensus procedure makes the standards under development known and available to interested parties and the general public. Such a procedure also facilitates dissemination and promotion of human factors and ergonomics knowledge across the world of nonexperts.

## REFERENCES

Albin, T. J. (2006), "Human Factors Engineering of Computer Workstations: HFES Draft Standard for Trial Use," in *Handbook of Human Factors and Ergonomics Standards and Guidelines*, W. Karwowsk, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 361–364.

Andreas, L. A., Stig, O. J., and Torbjørn. S. (2009), "CRIOP: A Human Factors Verification and Validation Methodology That Works in an Industrial Setting," *Lecture Notes in Computer Science*, Vol. 5775, pp. 243–256.

Bernsen, N. O., and Dybkjær, L. (2009), *Multimodal Usability*, Springer-Verlag, London, pp. 233–262.

Blum, R., and Khakzar, K. (2007), "Design Guidelines for PDA User Interface in the Context of Retail Sales Support," *Human-Computer Interaction, Interaction Platforms and Techniques*, Vol. 4551, pp. 226–235.

Chapanis, A. (1996), *Human Factors in Systems Engineering*, Wiley, New York.

Cullen, L. (2007), "Human Factors Integration—Bridging the Gap between System Designers and End-Users: A Case Study," *Safety Science*, Vol. 45, pp. 621–629.

Deros, B. M., Mohamad, D., Ismail, A. R., Soon, O. W., Lee, K. C., and Nordin, M. S. (2009), "Recommended Chair and Work Surfaces Dimensions of VDT Tasks for Malaysian Citizens," *European Journal of Scientific Research*, Vol. 34, No. 2, pp. 156–167.

Dickinson, C. E. (1995), "Proposed Manual Handling International and European Standards," *Applied Ergonomics*, Vol. 26, No. 4, pp. 265–270.

Dul, J., de Vlaming, P. M., and Munnik, M. J. (1996), "A Review of ISO and CEN Standards on Ergonomics," *International Journal of Industrial Ergonomics*, Vol. 17, No. 3, pp. 291–297.

Dzida, W. (1995), "Standards for User-Interfaces," *Computer Standards and Interfaces*, Vol. 17, No. 1, pp. 89–97.

Eibl, M. (2005), "International Standards of Interface Design," in *Handbook of Human Factors and Ergonomics Standards*, W. Karwowski, Ed., Lawrence Erlbaum Associates, Mahwah, NJ.

European Committee for Standardization (CEN) (2008), CEN/TC 122 Business Plan, CEN/TC 122 N838, available: http://www.cen.eu/cen/Sectors/TechnicalCommittees Workshops/CENTechnicalCommittees/Pages/PdfDisplay .aspx, accessed November 11, 2010.

Gerd, K., Kay, B., Pieper, U., and Dieter, L. (2009), "The Ergonomic Relevance of Anthropometrical Proportions: Part I: Body Length Measures," *Journal of Physiological Anthropology*, Vol. 28, No. 4, pp. 173–179.

Harker, S. (1995), "The Development of Ergonomics Standards for Software," *Applied Ergonomics*, Vol. 26, No. 4, pp. 275–279.

Helvi, H. T., Tiina, R., and Jari, K. (2009), "Airborne Enteric Coliphages and Bacteria in Sewage Treatment Plants," *Water Research*, Vol. 43, pp. 2558–2566.

Hoyle, D. (2001), *ISO 9000: Quality Systems Handbook*, Butterworth-Heinemann, Oxford.

Hoyle, D. (2006), *Quality Management Essentials*, Butterworth-Heinemann, Oxford.

Hoyle, D. (2009), *ISO 9000: Quality Systems Handbook*, Elsevier, Oxford.

Human Factors and Ergonomics Society (HFES) (2002), Board of Standards Review/Human Factors and Ergonomics Society 100, *Human Factors Engineering of Computer Workstations: Draft Standard for Trial Use*, HFES, Santa Monica, CA.

International Labour Organization (ILO) (2001), *Guidelines on Occupational Safety and Health Management Systems*, ILO-OSH 2001, ILO, Geneva.

International Labour Organization (ILO) (2010), "ILO Information Leaflet," available: http://www.ilo.org/wcmsp5/groups/public/—dgreports/—dcomm/—webdev/documents/publication/wcms_082361.pdf, accessed November 2, 2010.

Interantional Organization for Standardization/International Electrochemical Commission (ISO/IEC) (2004), *Standardization and Related Activities: General Vocabulary (Guide 2)*, ISO, Geneva.

Isaac, J. A. L. S., Douglas. V. T., Fernando. T. F., and Paulo. V. R. C. (2008), "The Use of a Simulator to Include Human Factors Issues in the Interface Design of a Nuclear Power Plant Control Room," *Journal of Loss Prevention in the Process Industries*, Vol. 21, No. 3, pp. 227–238.

Jamil, S., Golding, A., Floyd, H. L., and Capelli-Schellpfeffer, M. (2007), "Human Factors in Electrical Safety," paper presented at the Petroleum and Chemical Industry Technical Conference, PCIC '07, Institute of Electrical and Electronics Engineers, Calgary, Canada, September 17–19, pp. 1–8.

Karwowski, W., Ed. (2006), *Handbook of Human Factors and Ergonomics Standards and Guidelines*, Lawrence Erlbaum Associates, Mahwah, NJ.

Ludger, S., and Daniel, L. (2009), "Human-Computer Interaction in Aerial Surveillance Tasks," *Industrial Engineering and Ergonomics*, Part 5, pp. 511–521.

Luis, M. T., Pedro, M. L. A., and Elena, L. L. (2008), "WebA: A Tool for the Assistance in Design and Evaluation of Websites," *Journal of Universal Computer Science*, Vol. 14, No. 9, pp. 1496–1512.

McDaniel, J. W. (1996), "The Demise of Military Standards May Affect Ergonomics," *International Journal of Industrial Ergonomics*, Vol. 18, Nos. 5/6, pp. 339–348.

Mustafa, K., and Nadia, M. T. (2007), "Parameterized Human Body Model for Real-Time Applications," paper presented at the International Conference on Cyber Worlds, pp. 160–167.

Nachreiner, F. (1995), "Standards for Ergonomics Principles Relating to the Design of Work Systems and to Mental Workload," *Applied Ergonomics*, Vol. 26, No. 4, pp. 259–263.

Natapov, D., Castellucci, S. J., and MacKenzie, I. S. (2009), "ISO 9241-9 Evaluation of Video Game Controller," in *ACM International Conference Proceeding Series: Proceedings of Graphics Interface*, British Columbia, Canada, pp. 223–230.

Occupational Safety and Health Administration (OSHA) (2004), "Protocol for Developing Industry and Task Specific Ergonomic Guidelines," available: http://www.osha.gov/SLTC/ergonomics/guidelines_protocol.html.

Olesen, B. W. (1995), "International Standards and the Ergonomics of the Thermal Environment," *Applied Ergonomics*, Vol. 26, No. 4, pp. 293–302.

Olesen, B. W., and Parsons, K. C. (2002), "Introduction to Thermal Comfort Standards and to the Proposed New Version of EN ISO 7730," *Energy and Buildings*, Vol. 34, No. 6, pp. 537–548.

Parsons, K. (1995a), "Ergonomics and International Standards," *Applied Ergonomics*, Vol. 26, No. 4, pp. 237–238.

Parsons, K. C. (1995b), "Ergonomics of the Physical Environment: International Ergonomics Standards Concerning Speech Communication, Danger Signals, Lighting, Vibration and Surface Temperatures," *Applied Ergonomics*, Vol. 26, No. 4, pp. 281–292.

Parsons, K. C. (1995c), "Ergonomics and International Standards: Introduction, Brief Review of Standards for Anthropometry and Control Room Design and Useful Information," *Applied Ergonomics*, Vol. 26, No. 4, pp. 239–247.

Sandrock, S., Schutte, M., and Griefahn, B., (2009), "Impairing Effects of Noise in High and Low Noise Sensitive Persons Working on Different Mental Tasks," *International Archives of Occupational and Environmental Health*, Vol. 82, No. 6, pp. 779–785.

Schutte, M., Sandrock, S., and Griefahn, B (2007), "Factorial Validity of the Noise Sensitivity Questionnaire," *Noise & Health: A Quarterly Inter-Disciplinary International Journal*, Vol. 9, No. 37, pp. 96–100.

Schutte, M., Muller, U., Sandrock, S., Griefahn, B., Lavandier, C., and Barbot, B. (2009), "Perceived Quality Features of Aircraft Sounds: An Analysis of the Measurement Characteristics of a Newly Created Semantic Differential," *Applied Acoustics*, Vol. 70, No. 7, pp. 903–914.

Seabrook, K. A. (2001), "International Standards Update: Occupational Safety and Health Management Systems," in *Proceedings of the American Society of Safety Engineers' 2001 Professional Development Conference*, Anaheim, CA.

Sgro, F., Lo Bello, L., and Lipoma, M. (2009), "A Networked Embedded Computing Platform for Physical Activity Assessment," in Proceedings of the 2nd International Conference on Human System Interactions, Catania, Italy, May 21-23, pp. 146–151.

Silva, G. C. De., Lyons, M. J., Kawato, S., and Tetsutani, N. (2003), "Human Factors Evaluation of a Vision-Based Facial Gesture Interface," paper presented at the Computer Vision and Pattern Recognition Workshop, Madison, WI, pp. 52–52.

Spivak, S. M., and Brenner, F. C. (2001), *Standardization Essentials: Principles and Practice*, Marcel Dekker, New York.

Stewart, T. (1995), "Ergonomics Standards Concerning Human-System Interaction: Visual Displays, Controls and Environmental Requirements," *Applied Ergonomics*, Vol. 26, No. 4, pp. 271–274.

Stuart-Buttle, C. (2005), "Overview of International Standards and Guidelines," in *Handbook of Human Factors and Ergonomics Standards and Guidelines*, W. Karwowski, Ed., Lawrence Erlbaum Associates, Mahwah, NJ.

Sutcliffe, A. (2009), "Multimedia User Interface Design," in *Human-Computer Interaction: Design Issues, Solutions, and Applications*, A. Sears and J. A. Jacko, Eds., CRC Press, Boca Raton, FL, pp. 66–83.

Teather, R. J., Natapov, D., and Jenkin. M. (2010), "Evaluating Haptic Feedback in Virtual Environments Using ISO 9241–9," paper presented at the Virtual Reality Conference (VR), Institute of Electrical and Electronics Engineers, Massachusetts, pp. 307–308.

Tobar, L. M., Andres, P. M. L., and Lapena, E. L. (2008), "WebA: A Tool for the Assistance in Design AND Evaluation of Websites," *Journal of Universal Computer Science*, Vol. 14, No. 9, pp. 1496–1512.

Wetting, J. (2002), "New Developments in Standardization in the Past 15 Years: Product versus Process Related Standards," *Safety Science*, Vol. 40, Nos. 1–4, pp. 51–56.

# CHAPTER 56

# OFFICE ERGONOMICS*

**Marvin Dainoff, Wayne Maynard, and Michelle Robertson**
**Liberty Mutual Research Institute of Safety**
**Hopkinton, Massachusetts**

**Johan Hviid Andersen**
**Department of Occupational Medicine**
**Herning Hospital, Herning, Denmark**

---

\* The opinions in this chapter are those of the authors and do not represent the positions of Liberty Mutual Group or Herning Hospitals.

## 1  CONCEPTUAL OVERVIEW

### 1.1  Introduction

The purpose of this chapter is to present an overview of ergonomic issues in the office workplace. However, in determining the scope of this material, it is necessary to consider a fundamental question: What is an office? In the modern electronic workplace, the answer is not straightforward. A traditional (*Compact Edition of the Oxford English Dictionary*, 1971) definition states that an office is "a place for the transaction of private or public business." With the profusion of portable computers and hand-held devices currently available, almost any location can fit that definition: an airport waiting room, a kitchen table, an automobile, even a park bench. Nevertheless, this chapter must focus on evidence-based findings, and the bulk of such research has been conducted in traditional offices. Therefore, we will focus on workplaces whose primary purpose is some aspect of information processing and transformation, where some sort of computer equipment is employed, and whose occupants are expected to remain in place for extended periods of time (i.e., several hours). On the other hand, the general principles discussed here can be usefully applied to more temporary venues (e.g., setting up a temporary workspace with a laptop and modem in a hotel room). Of particular interest is the increasingly frequent case of telework in which individual employees are—either voluntarily or involuntarily—expected to set up their primary workplace at home.

The chapter cannot hope to be comprehensive. A search of the terms "office ergonomics" in Google Scholar resulted in 72,000 hits. Even allowing for overlap, a comprehensive literature review would likely result in all of the allocated space for this chapter being composed of citations. Nevertheless, it is the intent of the authors that at least the major issues will be introduced and discussed with sufficient citations to enable the reader to explore these issues in depth.

The contribution of individual authors can be specifically identified. In Section 2, Johan Andersen presents a systematic review of the epidemiological literature linking carpel tunnel syndrome and upper extremity disorders among computer users. In Sections 3 through 7 Wayne Maynard relies on consensus standards [Business and Institutional Furniture Manufacturers Association (BIFMA), 2002; Human Factors and Ergonomics Society (HFES), 2007], the draft report of the Z365 committee of the National Safety Council (Accredited Standards Committee, 2002), and many years of practical experience to summarize ergonomic recommendations for office design from a practitioner's perspective. In Section 8, Michelle Robertson considers the evidence for effective interventions. Finally, Marvin Dainoff provides a conceptual framework (Sections 1 and 9) and is responsible for overall organization and integration. The authors do not speak with one voice and disagreements will appear. We regard this as a strength of the chapter. As Root-Bernstein (1989) argues, "it is not consensus for which we must strive, but the elaboration of as many adequate descriptions of nature as we can imagine—in short the sort of complementarist view espoused by Bohr" (p. 375).

### 1.2  An Ecological Approach to Ergonomics

This overview is based on previous work (Dainoff and Mark, 2001; Dainoff, 2005, 2008).

Galison (1997) has provided an intriguing study of the development of the field of microphysics, from the nineteenth century cloud chamber to the factorylike laboratories of the present day at places like the European Organization for Nuclear Research (CERN), Stanford, and Berkeley. His particular focus is on the way in which the development of laboratory apparatus transformed the *social/organizational* structure of microphysics from individual investigators working alone or in small groups with total control and understanding of their apparatus to industrial-style organizations requiring collaboration among many professionals. In order for this collaboration to have occurred, Galison invokes the concept of "trading zones" (1997, p.46). Derived from the field of linguistics, trading zones refer to simplified languages (creoles, pidgins) that arise when adjacent cultures require a mutually understandable means of communication in order to transact business. Galison's insight has important implications for the field of ergonomics.

Just in the area of ergonomics of chairs and furniture, the applicable research involves individuals from a number of academic specialties, including biomechanics, epidemiology, economics, industrial engineering, industrial medicine, industrial design, muscle physiology, multivariate statistics, psychology of human performance, psychophysics, organizational design, orthopedics, and optometry. What is required is a trading zone, a conceptual framework within which specialists in one area can communicate with specialists in another.

### 1.3  Foundations

Two books were written at the end of the 1940s, which together should have changed the face of psychology. Both of these books offered naturalistic approaches to basic psychological processes which, prior to that time, had been treated in highly abstract ways, torn out of their functional contexts (Reed and Bril, 1996, p. 242).

The first of these books was J. J. Gibson's *Perception of the Visual World* (Gibson, 1950); the second, which was suppressed for political reasons, was Bernstein's *On Dexterity and Its Development*. The volume was finally published (Bernstein, 1996), and it remained for those who followed and elaborated the Gibsonian position to incorporate Bernstein's insights into the ecological position. See, for example, Turvey et al. (1978).

Gibson's ecological approach to psychology, as outlined in his last book (Gibson, 1979), provides part of the foundational basis for ecological ergonomics. Gibson rejected the prevailing view of psychology as the study of an individual organism, in which environmental context is simplified or ignored. Instead, he argued that the individual and environment are so tightly and reciprocally coupled that they cannot be studied independently of one another. Thus, to understand the simple case of a person walking across a field, a detailed physical analysis of the terrain is required—including characterization of the projected optical flow patterns across the retina

associated with movement in a given direction. Because this approach takes into account the interacting aspects of both person and environment, Gibson called it an "ecological" approach. Gibson's arguments are similar to the *systems*-versus-*psychological* distinction made by Meister (1989) and, in fact, Gibson himself saw similarities between his view and systems theory (Vicente, 1999).

While Gibson argued for the importance of viewing perception within the functional context of behavior in the environment, Bernstein presented a parallel argument for viewing actions within their functional/adaptive contexts. What he called "dexterity" is a capacity of solving motor control problems under dynamically changing parameters (Reed and Bril, 1996). Hence, movement science cannot study movements as abstract patterns without taking into account functional demands of task constraints, environmental constraints, and (changing) constraints within the individuals themselves (Newell, 1996).

## 1.4 Elements of an Ecological Framework for Ergonomics

The core of this approach rests on the parallelism between basic concepts of ergonomics and ecological psychology. Ergonomics can be characterized as the *fit* between the human being and those things (tools, workplaces, environments) with which he or she interacts (Dainoff and Dainoff, 1986). At the same time, ecological psychology provides a theoretical foundation which allows us to relate the physical attributes of people and their environments with behavioral acts required to function in that environment (Gibson, 1979). Therefore, the fact that both ecological psychology and ergonomics focus on the mutual relationship between person and environments leads us to propose an ecological framework for ergonomics.

The starting point for analysis is a single individual (or "actor") interacting with his or her environment. Two conceptual building blocks form the core of this analysis: *affordances* and *perception–action cycles*.

The first component of the ecological framework is the concept of *affordance*. Affordances are attributes of the environment of an individual (or "actor") defined with respect to the action capabilities of that individual (Dainoff and Mark, 2001; Gibson, 1979). Insofar as the fundamental definition of ergonomics can be construed in terms of the fit between individual and environment (Dainoff and Dainoff, 1986; see above), the concept of affordance, as developed within the theoretical framework of ecological psychology, provides a systematic approach to understanding and critical analysis of person–environment complementarity. Thus, components of an ergonomic chair are affordances for alternating between different seated work postures, but only for a particular set of users. The chair is not usable as designed for a two-year-old child who is too small, an extremely obese adult who is too large, or a person with muscle impairment who is unable to adjust the controls. The chair is not functionally usable by the person who does fall in the above categories but does not understand either how to adjust the controls or why such adjustments might be useful.

The concept of affordance is particularly relevant to ergonomic aspects of design, since it requires the designer to explicitly take into account how physical objects relate to the action capabilities of users. Action capabilities, in turn, are determined by certain classes of *constraints*: personal, environmental, and task (Newell, 1996). Personal constraints refer to individual variability, including body dimensions (anthropometry), biodynamics (body strength, mass, flexibility), and relevant psychological factors (perceptual, cognitive, motivational). Environmental constraints include both size and shape of objects and surfaces as well as their physical properties relevant to the action.

The second component of the ecological framework for ergonomics is the *perception–action cycle*. Any integrated behavior pattern (task) can be decomposed into a series of steps in which perceived information about the possibility of action is followed by the action itself, which, in turn, reveals new information about potential actions, and so on. For example, information is extracted from the home page of a website indicating the presence of a clickable button. Hand and finger muscles move the mouse to the location of the button and click. The effect of the click is to act on the environment—a new Web page appears. The new page has additional information, some of which is extracted (perceived) and the cycle continues. (See Figure 1.)

Perception–action cycles are defined by certain classes of constraints, which can be classified into four groups. Task constraints reflect the functional requirements of a task. This includes individual task demands. A second class of constraints consists of surrounding social and organizational factors. Workspace constraints reflect the layout of components within the workspace (e.g., display, input/out device, furniture) as well as relevant environmental constraints (lighting, air quality, temperature and humidity, gravity). Individual constraints reflect both physical (anthropometric and physiological) and psychological (cognitive, motivational, emotional) attributes of the actor.

The action components of perception–action cycles take place within a three-dimensional postural envelope. Within the postural envelope, the operator must reach, lift, and manipulate, while parts of his or her body are or are not supported.

The perception–action cycle is the theoretical conception which links Gibson's concept of affordance with Bernstein's concept of skill as the coordination of multiple degrees of freedom. It is the information about affordances which, when perceived, initiates the perception–action cycle by revealing the capabilities for action within the environment. At the same time, the existence of multiple affordances within the system requires a degree of coordination through selection and timing of appropriate task-relevant actions associated with appropriate affordances. When this coordination can be achieved under changing environmental constraints, the user has achieved what Bernstein called "dexterity" (Bernstein, 1996). *Consequently, we argue that the perception–action cycle should be the fundamental unit of analysis of work systems and, therefore, a basic tool for ergonomics.*
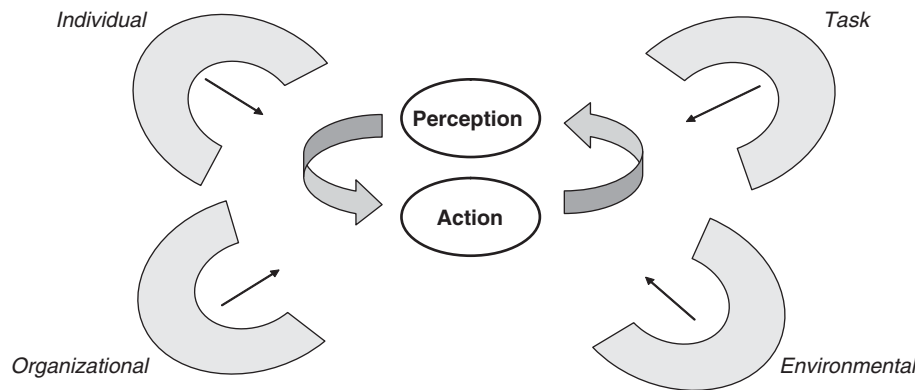
**Figure 1**  Perception–action cycle defined by constraints.

### 1.4.1  Application of the Framework

The preceding conceptual framework can be used as a practical tool, a kind of lens with which the remaining information in this chapter may be viewed.

Let us consider in detail the example of a user seated at a fully adjustable computer workstation while writing an article for a book chapter. The adjustability of the height and angle of the seat, angle of the backrest, height and angle of the keyboard support, height and angle of the monitor, as well as the scalable font size of the word-processing program are all affordances which allow this particular person to achieve a comfortable working posture. Each of these affordances refer to physical properties of the workstation defined in terms of corresponding attributes (action capabilities) of the actor's anthropometry, motor control ability, and understanding of operation and coordination of the control mechanisms. These affordances are defined in explicit detail in Section 4. Thus, in a real sense, the products of design are affordances.

A critical point here, and one that is often misunderstood, is that the physical attributes described above are only considered affordances if they are perceived as such by the actor. That is, a chair may have a height adjustment lever, but if the actor does not know that it exists, knows that it exists but does not know its location, knows its location but not how it operates, or knows how it operates but does not have the physical capability to operate it, the chair *does not* have an affordance for height adjustability.

If, on the other hand, the actor is fully aware of the functionality of each of the components above, they become affordances enabling the individual to adapt to the complexity of the workstation in that he or she can achieve comfortable/efficient working posture. More precisely, this awareness can allow the execution of multiple nested perception–action cycles.

Consider the detailed task requirements involved in creating text in the process of composing a book chapter. This involves periods of rapid keyboard operation interspersed with periods of reflection and pondering. Therefore, one set of perception–action cycles involves the linkage between actions of the fingers on the keys and corresponding reading of the text on the screen. Here the fingers and eyes are playing a primary role whereas the remaining bodily structures (head, neck, arms, shoulders, trunk) are in a more passive supporting role. What we call efficient/comfortable working posture is a set of perception–action cycles which adjust the relative locations of these limbs into orientations which allow the primary keying–reading cycle to be easily performed. [*Note*: The term "comfort" here is a convenient label for a hypothetical underlying physiological principle of efficiency or least effort. (See, e.g., Nubar and Contini, 1961.) ]

In a well-designed office environment, efficient/comfortable posture is afforded by the adjustability mechanisms in that the seat and backrest can be adjusted so that the trunk can be inclined backward with the feet flat on the floor and the lower back supported. The keyboard support is adjusted so that the hands are flat and forearms parallel to the floor. The head is erect and the monitor is within a field of view 30° below the horizontal. The backward-inclined working posture places the eyes over 100 cm from the display screen, but this is compensated for by increasing the font size of the displayed characters. These criteria for comfortable posture are contained in many standards and guidelines [e.g., American National Standards Institute (ANSI)–HFES 100-2007: U.S National Standard for Human Factors Engineering of Computer Workstations (HFES, 2007)] and are reviewed later in Section 4.

It is useful to examine in the detail the perception–action cycle for just one of these adjustments—seat height. It is assumed that the actor is fully informed about the functionality of the particular chair he or she is using which has a height adjustment lever (HAL) below the seat surface on the right side. Table 1 depicts the four stages of the perception–action cycle required to operate this lever and raise the level of the seat to the desired height. This entails locating the lever, pulling it upward while at the same time elevating the buttocks, and allowing the seat surface to move upward to a desired height. In this case, both the HAL and seat surface can be considered affordances.

**Table 1 Perception–Action Cycle for Height Adjustment of Ergonomic Chair**

| Information | Perception | Activation | Action |
|---|---|---|---|
| Height adjustment lever (HAL) right side below seat | Tactile/visual pickup of HAL location | Arm extension plus grasp | Hand moves to HAL |
| HAL is in operating position (graspable) | Tactile pickup that HAL is graspable | Grasp and pull HAL: rotate thighs and buttocks upward | Seat surface moves upward to new position of buttocks |
| Elbow height lower than work surface | Tactile/visual pickup: seat height not yet correct | Continue | |

What is the desired height? In this case, the goal is to raise the trunk so that the elbows are approximately level with the keyboard. At the same time, the legs should be straight and feet flat on the floor. Depending on the individual anthropometry of the actor, this goal may not be achievable without additional perception–action cycles, such as adjusting the height of the keyboard support surface.

A completely different set of perception–action cycles is brought into play if the task requirements change. Assume that a section of the chapter has been finished and printed on paper copy. The actor's preference is to place the paper document on a copy holder adjacent to the display screen for further editing. Hence a new adaptation in working posture to these changed demands is necessary. Because the small font size on the paper document is no longer visible while the seat back is inclined rearward, the seat back must be adjusted to a more upright position.

In Bernstein's terms, a certain degree of dexterity (coordination of multiple degrees of freedom) is required to utilize the adaptive potential of a modern ergonomic office to achieve the desired goal of efficient/comfortable working posture for multiple task demands. These degrees of freedom are manifest as individual furniture adjustment mechanisms and major joints of the body. Depending on the particular equipment supplied, adjustment mechanisms can include seatpan angle, seatpan height, backrest angle, backrest height, backrest tension, work surface height, monitor height, and monitor angle. Major joints include wrist, elbow, shoulder, neck, thigh, knee, and ankle.

While the motor control mechanisms for coordinating postural degrees of freedom is the subject of considerable research activity within the ecological psychology community (see, e.g., Latash and Turvey, 1996), for the purposes of this chapter, it is sufficient to reiterate the point made earlier that the actor must understand both how to adjust the controls and why such adjustments might be useful. Simply put, user training must be an essential component of office ergonomics.

Therefore, to return to where we started, fit (Dainoff and Dainoff, 1986) is achieved when the appropriate constraints and affordances are available so as to allow adaptive perception–action cycles to move the actor into a three-dimensional postural envelope (comfort zone) within which task activities can be carried out with a minimum of physical effort (see, e.g., Newell, 1996).

## 2 EPIDEMIOLOGICAL EVIDENCE FOR CARPAL TUNNEL SYNDROME AND UPPER EXTREMITY MUSCULOSKELETAL DISORDERS AMONG COMPUTER USERS

### 2.1 Introduction

Musculoskeletal disorders of the neck and upper limb (UEMSDs) and carpal tunnel syndrome (CTS) have been linked to keyboard and visual display terminal (VDT) use since the beginning of the 1970s. Early reports on occupational cramps and muscle pain appeared in Australia (Ferguson, 1971) and Japan (Maeda, 1977) after use of electric keyboards or among accounting machine operators. Later, an apparent epidemic occurred in Australia in the mid-1980s, where so-called repetition strain injuries (RSIs) were frequently reported among computer users. The epidemic disappeared, and the background and causes of the epidemic have been discussed ever since. Historically, there have been similar examples of outbreaks of pain and cramps—for example, writer's cramp or telegraphers' cramp (Dembe, 1996)—often coinciding with the introduction of new technology into the society. Arguments on causes to explain the outbreaks have ranged from specific physical exposures at the workplace to cultural beliefs and societal expectations (Lucire, 2003). In Europe and the United States, the first concerns on health effects of VDTs were potential adverse effects on reproduction, which has been refuted by large epidemiological studies.

The majority of UEMSDs are characterized by recurrent episodes of pain and consequent disability, varying in severity and impact. Most of the episodes are self-limiting and subside within days or weeks, while

some end up with long-lasting chronic problems. Risk factors from physical, psychological, and social domains have been identified, but the relative contribution of the various risk factors to the onset and aggravation of UEMSDs is less known. As a result, controversies still exist regarding the degree of work-relatedness of UEMSDs (Silverstein et al., 1996).

The last decades, since the anecdotal stories in the 1970s and the early reviews, have been characterized by a steady increase in the number of published studies on computer work and UEMSDs, The studies generally fall into one of two categories: either experimental studies trying to identify a possible pathophysiology of computer-related disorders or intervention studies and epidemiological studies focusing on the association between workplace risk factors and musculoskeletal outcomes.

The pathophysiological studies have produced a large number of hypothetical injury mechanisms ranging from a systematic overload of low-threshold motor units (the "Cinderella" hypothesis) to intracellular $Ca^{2+}$ accumulation, impaired blood flow, reperfusion injury, blood vessel nociceptor interaction, myofascial force transmission, intramuscular shear forces, trigger points, and aggravated heat shock response. There is a certain degree of overlap between the hypotheses, but in spite of intensive research the empirical data to support a unifying hypothesis or identify a specific injury mechanism has been limited. It is assumed that pain results from muscle tissue damage due to prolonged low-force muscle activity with few breaks and little variation. However, muscle activity measured by electromyography seems to be slightly higher during rest breaks and noncomputer office work than during computer work. In a study by Richter et al. (2009), the division between computer activity and noncomputer activity was based on electronic activity registration with different cut-offs. These observations may be seen as support for few, if any, biologically plausible effects of mouse work on UEMSDs.

The pathophysiological or mechanistic studies are not included in this chapter. Instead, the scope is to summarize the knowledge and synthesize the evidence gained from the large number of risk factor studies, including prospective studies, which have been published since 2000. Although several systematic reviews on computer work and UEMSDs and CTS have been published in recent years in an attempt to provide this kind of information, the conclusions in the reviews are often in discord and the heterogeneity has created a situation of confusion rather than of clarity.

## 2.2 Carpal Tunnel Syndrome

Carpal tunnel syndrome is a compression neuropathy of the median nerve as it passes through the carpal tunnel. It is regarded as the most frequent compression neuropathy. Based on both clinical symptoms and nerve conduction tests (NCTs), overall prevalences of 3.0–5.8% among women and 0.6–2.1% among men have been found in general population samples (Atroshi et al., 1999). CTS is generally believed to be caused by increased pressure in the carpal tunnel. It is widely accepted that exposure to hand–arm vibrations and exposure to a combination of repetitive hand use and hand force may be causally related to CTS. In recent years, with the expanding use of computers, it has been a matter of concern if computer use could be a risk factor for the development of CTS. Three recent reviews of high quality concluded that there is insufficient epidemiological evidence that computer work causes CTS (Palmer et al., 2007; Thomsen et al., 2008; Van Rijn et al., 2009). Table 2 summarizes the aims and main conclusions from the three reviews.

## 2.3 Upper Extremity Musculoskeletal Disorders

Upper extremity musculoskeletal disorders cover a wide range of complaints from the neck, shoulder, elbow, forearm, and wrist/hand. Umbrella terms such as repetition strain injuries (RSIs), occupational cervicobrachial disorders (OCDs), and cumulative trauma disorders (CTDs) have often been used in the literature, but terms like these assume that the proposed mechanisms or exposure must be avoided. It is fairly consistent from the literature that distal arm pain and, to a lesser extent, neck–shoulder pain are associated with intensive use of the keyboard and the mouse, and the conclusions from several reviews (Gerr et al., 2004, 2006; Griffiths et al., 2007; IJmker et al., 2007; Village et al., 2005; Waersted et al., 2010; Wahlström, 2005) are shown in Table 2. The reviews are based on a total of 80 original studies.

The association is much more uncertain when it comes to computer use and more prolonged or chronic pain and clinical entities such as shoulder tendonitis, lateral and medial epicondylitis, forearm disorders, or wrist tendonitis. Researchers recently performed a critical review of the epidemiological evidence for a possible causal relationship between different aspects of computer work, including keyboard and mouse use, and neck and upper extremity musculoskeletal disorders diagnosed with a physical examination (Wahlström, 2005). As can be seen in Table 2, they found limited epidemiological evidence for an association between aspects of computer work and the clinical diagnoses. There is a tendency for recent reviews to be more critical than earlier reviews, even though newer and prospective studies have been included in the reviews. In epidemiological studies it is usually found that more and better studies provide stronger evidence for a causal relation if such exists in the real world. The current level of findings is in contradiction with a causal relation between aspects of computer use and clinical verified UEMSDs. Nevertheless, this should be carefully interpreted. With the very widespread use of computers in professional work life and in leisure activities, even a small increase in risk could have profound importance and, based on our current knowledge, we cannot discount such small risks. Over the last 25 years we have witnessed some big changes in the office environment. The physical work environment has changed and, maybe of more importance, the psychosocial and work organizational circumstances have changed toward increase in office worker flexibility, more precarious work, and incessant organizational changes. Loss of worker autonomy, lack of predictability and meaning, and appreciation by supervisors or peers are probably of more importance in today's office work than biomechanical loads.

**Table 2  Aims and Main Conclusions from Reviews on Risk Factors for CTS and UEMSDs among Computer Users**

| | **Carpal Tunnel Syndrome** |
|---|---|
| Palmer et al., 2007 | *Aim*: To assess occupational risk factors for CTS<br>*Conclusion*: The balance of evidence on keyboard and computer work does not indicate an important association with CTS. |
| Thomsen et al., 2008 | *Aim*: To examine evidence for an association between computer work and CTS<br>*Conclusion*: There is insufficient epidemiological evidence that computer work causes CTS. |
| Van Rijn et al., 2009 | *Aim*: A quantitative assessment of exposure–response relationships between work-related physical and psychosocial factors and the occurrence of CTS in occupational populations<br>*Conclusion*: The contradictory findings for computer use and the development of CTS are in agreement with the conclusion of a recent review (Thomsen et al., 2008). |
| | **Upper Extremity Musculoskeletal Disorders** |
| Gerr et al., 2004 | *Aim*: The epidemiological evidence examining associations between UEMSDs and computer use posture and keyboard use intensity (hours of computer use per day or per week).<br>*Conclusion*: Daily or weekly hours of computer use is more consistently associated with hand and arm MSDs than with neck and shoulder MSDs. |
| Wahlström, 2005 | *Aim*: To give a summary of the knowledge regarding ergonomics, musculoskeletal disorders, and computer work and to present a model that could be used in future research.<br>*Conclusion*: None. It is hypothesized that perceived muscular tension is an early sign of musculoskeletal disorder, which arises as a result of work organizational and psychosocial factors as well as from physical load and individual factors. |
| Village et al., 2005 | *Aim*: To evaluate the evidence supporting a causal relationship between computer work and musculoskeletal symptoms and disorders (MSDs) of the hand, wrist, forearm, and elbow.<br>*Conclusion*: There is consistent evidence of a positive relationship across numerous prospective and cross-sectional studies with increased risk most pronounced beyond 20 h/week of computer use or with increasing years of computer work. The disorders confirmed with physical examinations are wrist tendonitis and tenosynovitis, medial and lateral epicondylitis, and DeQuervain's tenosynovitis. The risk of carpal tunnel syndrome is increased with a use of a computer, especially with mouse use for more than 20 h/week. |
| Gerr et al., 2006 | *Aim*: To explore the epidemiological evidence of associations between upper extremity musculoskeletal symptoms and disorders and keyboard use intensity (hours of computer use per day or per week) and computer use postures<br>*Conclusion*: A somewhat consistent finding is an observed association between hours of computer use and adverse hand/arm MSD outcomes and, to a slightly lesser extent, between hours of computer use and adverse neck/shoulder outcomes.<br>The conclusion also points to severe methodological limitations in the literature. |
| Griffiths et al., 2007 | *Aim*: To draw attention to the potential risks to musculoskeletal health with the computerization of work among professional occupational groups.<br>*Conclusion*: The risk factors for work-related musculoskeletal symptoms with computer work have been extensively researched and are generally well established. |
| IJmker et al., 2007 | *Aim*: To get a more conclusive insight into the relationship between the duration of computer use and the incidence of hand–arm and neck–shoulder symptoms and disorders, a systematic review of longitudinal studies was performed.<br>*Conclusion*: This review showed moderate evidence for an association between the duration of mouse use and the incidence of hand–arm symptoms. Indications for a dose–response were found. In addition, the neck–shoulder region seemed less susceptible to exposure to computer use than the hand–arm region. |
| Waersted et al., 2010 | *Aim*: To examine the evidence between computer work and neck and upper extremity disorders (except carpal tunnel syndrome).<br>*Conclusion*: There is limited epidemiological evidence for an association between aspects of computer work and some of the clinical diagnoses. None of the evidence was considered as moderate or strong and there is a need for more and better documentation. |

## 3   ERGONOMICS PROGRAMS FOR OFFICE ENVIRONMENTS

Ergonomics is the design of jobs to match the capabilities and limitations of workers. Jobs designed ergonomically result in higher productivity and quality and improved workplace safety. To achieve such outcomes in ergonomics requires a managed health and safety process that targets the design of tasks, workstations, tools, equipment, and organizations to reduce risk factors that can contribute to injury and disability. Many believe the most effective ergonomics program in an office environment is one that properly fits an employee at their computer workstation with the chair, keyboard and mouse, monitor, or display at the correct height with proper seated posture. In reality, proper workstation adjustment is but a small component of an office ergonomics process.

The following guidelines, comprising Section 3 of this chapter, are adapted from the 2002 final draft of ASC Z365 Management of Work-Related Musculoskeletal Disorders (Accredited Standards Committee, 2002) and describe the elements of an ergonomics program and process for managing work-related musculoskeletal disorders (WMSDs) to reduce frequency and disability.

An ergonomics program for WMSDs has the following components:

- Employer responsibilities
- Training
- Employee involvement
- Injury and hazard surveillance
- Evaluation and management of WMSD cases
- Job analysis
- Job design and intervention

### 3.1   Employer Responsibilities

Effective implementation of a managed ergonomics process will require establishing priorities for prevention and control activities. The choice of priorities will depend on the progress made in addressing workplace factors and on the extent of problems already present in the workplace. Some will focus first on management of diagnosed WMSD cases and evaluation and intervention of the corresponding jobs. Others, in worksites with few or no WMSD cases or high employee turnover, may move straight to implementing proactive job surveys (e.g., employee interviews, checklists) so that potential problems associated with particular jobs can be identified and addressed before new WMSD cases appear.

The level and breadth of training and employee involvement may directly depend on how the program is initiated and progresses over time. As Figure 2 shows, there are three surveillance outcomes that could lead to job analysis. An employer may focus first on the employee reports and, in doing so, provide the appropriate training and employee involvement to accomplish this goal. As one moves to surveillance using existing records and job surveys, participants may change and so may the

corresponding level and breadth of training and employee involvement.

### 3.2   Training

Periodic training is necessary so that employees and managers can facilitate surveillance, job analysis, job design, and medical management. Be sure to provide training to appropriate management representatives and employees. Training may include:

- Recognition and reporting of the signs and symptoms of WMSDs that may be work related
- Record-keeping processes for reporting WMSDs
- Whom to contact for further assistance
- Roles and responsibilities in the surveillance procedures
- Recognition and management of WMSD risk factors
- Job analysis and design procedures
- Proper use, adjustment, and maintenance of tools, work equipment, and work stations
- Job interventions and best-work procedures and practices for minimizing risk of WMSDs

### 3.3   Employee Involvement

Give employees the opportunity to participate in the program for management of WMSDs. The following are examples of employee involvement:

- Submitting suggestions or concerns
- Participating in discussions related to their workplace and work methods
- Participating in employee surveys
- Participating in formal team meetings
- Using and operating tools and work equipment in the prescribed manner
- Participating in the design of work, equipment, and procedures
- Participating in the employer's WMSD problem-solving process
- Participating in WMSD education and training
- Notifying the employer of related WMSD symptoms and risk factors early

### 3.4   Injury and Hazard Surveillance

The results of surveillance are used to determine when and where job analysis is needed and where ergonomic interventions may be warranted. Each organization may want to establish criteria for when a job survey result or health surveillance data indicate the need for a job analysis. This information may be further used to assist in establishing job analysis and intervention priorities and assessing the program. Surveillance includes:

1. Initial review of existing records of work-related illnesses and injuries [e.g., Occupational Safety and Health Administration (OSHA) log and workers' compensation records] and at the start of surveillance, periodically thereafter. This

Program management
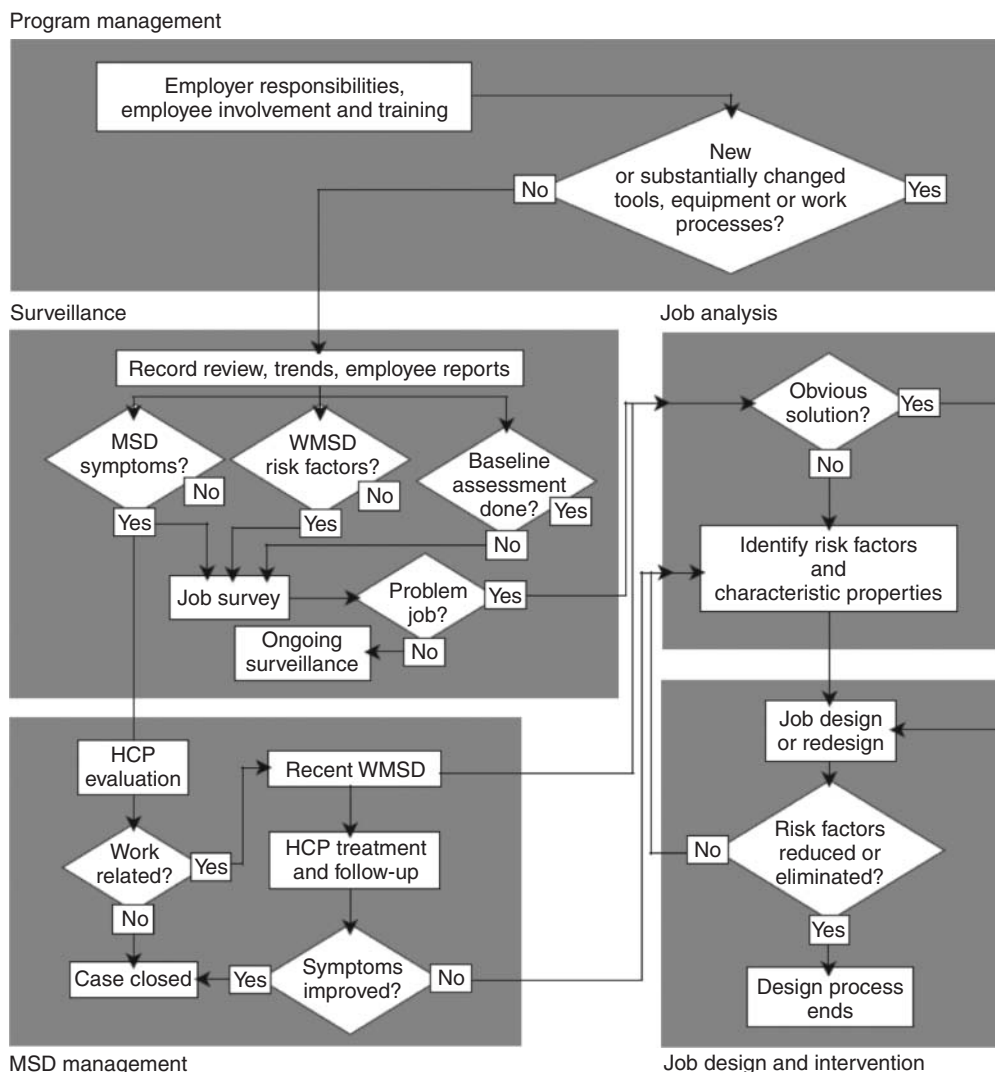


**Figure 2** Program for management of WMSD flowchart illustrating program elements. (Adapted from the 2002 final draft of ASC Z365 2002; Accredited Standards Committee, 2002.)

analysis will help determine where WMSDs are occurring and will help prioritize jobs needing further analysis.

2. Employee reports—There are two kinds of employee reports:
   a. Employee reports of WMSD symptoms
   b. Reports of employee concerns about WMSD risk factors
3. Job surveys—The aim of job surveys is to identify specific jobs and processes that may put employees at risk of developing WMSDs. Job surveys are considered a cursory or preliminary review of jobs, as compared to a more detailed

job analyses. Job surveys may include any of the following methods:
   a. Office walkthroughs
   b. Employee and supervisor interviews/questionnaires
   c. Computer workstation design assessment checklists
   d. Team problem-solving approaches

Job surveys can be incorporated into existing programs such as regular safety, health, team problem solving, or quality inspections and can expand their scope to include identification of WMSD risk factors. Results of job surveys may be applied to similar jobs within one or more departments or locations:

*Perform job surveys when*:

- New WMSD cases are reported, to help determine if risk factors exist across similar jobs that use similar tasks, equipment, tools, or processes. This might include a sampling of representative jobs.
- Employees report new MSD symptoms.
- Employees report WMSD risk factors.
- There is an unexplained high rate of turnover or absenteeism for a specific job. There may be many reasons for turnover or absenteeism not related to WMSDs.
- Surveillance activities are begun as a baseline assessment of job risk factors.
- A job, equipment, or process substantially changes to identify risk factors that may result from making these changes.
- New equipment or furniture or work processes are planned, purchased, or installed.

### 3.5 Evaluation and Management of WMSD Cases

Early assessment or establishing a diagnosis and initiating treatment may limit the severity, improve the effectiveness of the treatment, and allow for sufficient and timely recovery of the condition. Early identification of WMSDs can alert the employer to the need for job analysis of that employee's job or the need for further analysis if the job has already been evaluated.

***It is recommended that employers:***

- Examine existing policies, practices, and programs to ensure that they encourage prompt reporting of MSD symptoms or potential WMSD risk factors without reprisals.
- Once notified of recurrent or persistent MSD symptoms, facilitate a prompt evaluation of the symptomatic employee by an appropriate health care provider (HCP) consistent with state laws.
- Provide the HCP with a contact who is familiar with the job tasks.
- Provide HCPs the opportunity to become familiar with jobs and job tasks (e.g., site walkthroughs, review of job surveys, analysis reports, detailed job descriptions, job safety analyses, photographs, or videotapes).
- Ensure employee privacy and confidentiality regarding medical conditions identified during the assessment, as permitted by law.

***In addition, employers should:***

- Select or recommend HCPs with knowledge, experience, and training in workplace exposures and the evaluation and treatment of WMSDs.
- Whenever feasible, modify jobs, redesign the job, and/or accommodate employees with work restrictions as determined by a HCP.

*Note*: Refer to the Americans with Disabilities Act for guidance relevant to employees with disabilities.

***The HCP should:***

- Evaluate the symptomatic employee.
- Seek information and review materials regarding employee job activities.
- Be familiar with the management of WMSD cases or refer the employee to a HCP who is familiar with such management.

***Components of the HCP evaluation include:***

- A medical history (occupational and nonoccupational) which includes a complete description of symptoms.
- A description of work activities as reported by the employee and the employer.
- A review of exposure information relevant to the clinical findings.
- A physical examination appropriate to the presenting symptoms and history.
- An initial assessment or diagnosis and an appropriate treatment plan.
- Work restrictions or work modifications if appropriate.
- An opinion on the work-relatedness of the disorder based on professional guidelines [e.g., *A Guide to the Work-Relatedness of Disease* (Kusnetz and Hutchinson, 1979)].

***Employees with WMSDs should:***

- Provide input to and follow the treatment plan recommended by the HCP, including work restrictions.

The HCP should follow up with symptomatic employees to document symptom improvement or resolution or reevaluate the employee who may not have improved. The time frame for this follow-up depends on the type, duration, and severity of the employee symptoms. If symptoms do not improve within the expected time frame, the employee should be referred to an appropriate medical specialist and/or the job should be analyzed again.

When the employer has determined, based on the medical evaluation and exposure information (job description, walk through, etc.), that an employee has a WMSD, he or she should perform a job analysis of the employee's job or a sample of representative jobs and include input from the symptomatic employee.

### 3.6 Job Analysis

Job analyses are more detailed studies of the work than job surveys. Job analyses identify potential exposures to work-related risk factors and evaluate their characteristic properties.

***Perform job analyses:***

- When it is suspected that an MSD is work related.
- When a problem job is identified from a records review, a trend of WMSDs, or job surveys.

- When a problem persists after changes have been implemented.
- During the design or acquisition phase of equipment, processes, or jobs.

Work-related risk factors are present at varying levels for different jobs and tasks. The mere presence of a risk factor does not necessarily mean that an employee performing a job is at undue risk of injury. Generally, the greater the exposure is to a single risk factor or combination of risk factors, the greater the probability of a WMSD. For example, the risk associated with the first three risk factors may be increased in the presence of cold temperature (see Bernard, 1997; National Research Council & Institute of Medicine, 2001).

### Consider the following work-related risk factors in job analysis:

- Force and contact stress. Can include tasks other than keyboarding work as well, including gripping heavy file folders and heavy manual materials handling work
- Posture and motions
- Vibration
- Cold temperature

### Evaluate WMSD risk factors for the following exposure properties of the physical stresses listed above by qualitative or quantitative approaches:

- Magnitude
- Repetition
- Duration
- Recovery

Also consider work organization factors that can alter the characteristic properties or effects of physical stress exposure.

Work-related risk factors may pose minimal risk of injury if sufficient exposure is not present or if sufficient recovery time is provided. However, if there is sufficient exposure and insufficient recovery time, there will be a risk of injury. Reducing exposure to risk factors will result in reduced probability or severity of WMSDs. When work-related risk factors and their corresponding exposure properties are identified and prioritized from a job survey or analysis, job design or redesign, including feasible engineering or administrative changes, can eliminate or reduce exposure to work-related risk factors. The decision regarding which specific risk factor to reduce in job design or redesign is based on the scientific evidence, professional judgment, technical feasibility, and input from employees and management.

### 3.7 Job Design and Intervention

#### The job design and intervention process ends when:

1. exposures to work-related risk factors are reduced or eliminated as much as practical or surveillance indicates that the problem is under control or
2. appropriate exposure limits have been identified and met.

## 4 OFFICE FURNITURE DESIGN

Visual display terminals have been a subject of some concern as their use in business and industry has become almost universal. VDT technology improves productivity and simplifies work, but it also has the potential to cause problems when poor workplace design is coupled with high keying rates. Most of the reported problems have involved dedicated or full-time operators who use their VDTs for 4 or more hours a day.

Complaints have included back, neck, and wrist pains, eye strain, headaches, and stress. These symptoms are often associated with the fatigue and discomfort that can result from poor installation of VDT equipment. Applying ergonomic principles to the design of VDT workstations can alleviate many of these problems.

A well-designed VDT workstation will allow the operator to sit with good posture, see the screen clearly, and reach the keyboard and document easily. Operator comfort and sufficient room to work are key factors in improving productivity and reducing complaints. The best workstation designs allow independent height adjustment of the screen, keyboard, and chair. Many manufacturers now offer workstations and furniture designed specifically to meet the ergonomic needs of VDT users.

The diagrams and guidelines on the following pages give ergonomic considerations for selecting, installing, and adjusting VDTs and VDT workstations.

### 4.1 Seating and Viewing Considerations

The preferred chair type is a swivel chair on a five-point base, with a rounded front edge on the seat, easily height adjustable by the operator. Position the monitor so that the gaze angle to the center of the screen ranges between $15°$ and $20°$ below horizontal eye level. Always take into account the vision requirements of VDT operators who wear glasses or bifocals. The following considerations are adapted from ANSI/HFES 100 and BIFMA G1-2002 (BIFMA, 2002; HFES, 2007):

*Seat Height*. Seat height should be adjustable by the user within the recommended range of 15–22 in. (38–56 cm). If the operator is too short to keep both feet flat on floor in the suggested height range, provide a foot rest.

*Seat Depth*. Adequate seat depth supports the thighs and allows the user to sit back far enough to use the lower portion of backrest without creating pressure on back of the knees. If nonadjustable, seat depth should be no greater than 17 in. (43 cm). If adjustable, seat pan adjustment range should include 17 in. (43 cm) or less.

*Seat Pan Angle*. Seats may be designed with a fixed or adjustable seat angle (e.g., recline backward or forward from horizontal). If fixed, seats should be within the range from $0°$ (horizontal) to $4°$ rearward. If adjustable, seats should include some part of the range from $0°$ to $4°$ rearward.

*Backrest Height*. The backrest provides support for the back in various postures. The top of the

backrest should be at least 18 in. (45 cm) above the compressed seat height.

*Lumbar Support*. This helps maintain the natural curvature of the spine at the small of the back. The lumbar support area of the backrest should be located between 6 and 10 in. (15 and 25 cm) above the compressed seat height.

*Seat Pan–Backrest Angle*. Studies have shown that a recline angle of 30° from vertical reduces fatigue. The torso-to-thigh angle should be at least 90°. If adjustable, the backrest should recline at least 115° from vertical.

*Armrest Height*. Proper armrest height supports the neck and shoulders. Armrests should be adjustable from 7 to 11 in. (18–27 cm) above compressed seat height. All armrests should be detachable.

*Eye-to-Screen Distance*. Preferably at least 20 in. (51 cm); minimum 12 in. (30 cm).

*Angle between Upper Arm and Forearm*. Elbow angle between 70° and 135° is recommended.

*Work Surface Height*. Should accommodate the user population. Minimum range of adjustability should be 28 inches (56 cm to 72 cm) from floor.

### 4.1.1 Seated Postures

It is expected that VDT users will frequently change working postures to maintain comfort and productivity. Four reference postures (see Figure 3) are recognized and commonly observed at computer workstations. Movement within these postures is encouraged.

These working postures are acceptable as long as the workstation has been properly adjusted to the employee. Standing posture can occur when working at standing workstations or when getting out of a chair to do other work, for example, retrieving items from a printer.

### 4.2 Work Surface and Seated Clearance Considerations

The keyboard should be thin and detached from the console, and the mouse or track ball should be at the same level as the keyboard. Clearance guidelines are designed to accommodate upright, reclined, and declined sitting postures. The items that follow are adapted from ANSI/HFES 100 and BIFMA G1-2002 (BIFMA, 2002; HFES, 2007):

*Work Surface Width*. At least 27.5 in. (70 cm) wide.

*Palm Rest Depth*. Minimum 1.5 in. (3.8 cm).

*Input Device Support Surface*. For sitting postures, adjust in height per work surface height recommendations.

*Thigh Clearance (Height)*. If not adjustable, no less than 27 in. (68 cm) at front edge of work surface and 25 in. (64 cm) at 17 in. (43 cm) rearward from front edge of work surface. If adjustable, it should include a height clearance of 27 in. (68 cm) as part of the adjustment range.

*Thigh Clearance (Width)*. No less than 20 in. (50 cm).

*Knee and Feet Clearance (Depth)*. At knee level, no less than 17 in. (44 cm) deep and no less than 23.5 in. (60 cm) deep at foot level. Work surface depth should allow for knee and feet clearances and a viewing distance to monitor of at least 19.7 in. (50 cm).

### 4.2.1 Chair Selection Tips

- Chair adjustment controls should be easily operable from a seated position.
- The chair and adjustment mechanisms should be rugged.
- Supply chairs with both detachable and adjustable armrests. Remove them if they interfere with the task.
- Seat should be padded for comfort.
- Several chair styles should be available to accommodate different sizes and preferences of users. Seat pan adjustment is often an optional accessory so beware that one size chair may not fit all.
- Back-tilt tension should be adjustable.
- The chair should permit alterations in posture and freedom of movement.



Reclined sitting  Upright sitting  Declined sitting  Standing

(105–120°)  (between 90 and 105°)  (>90°)
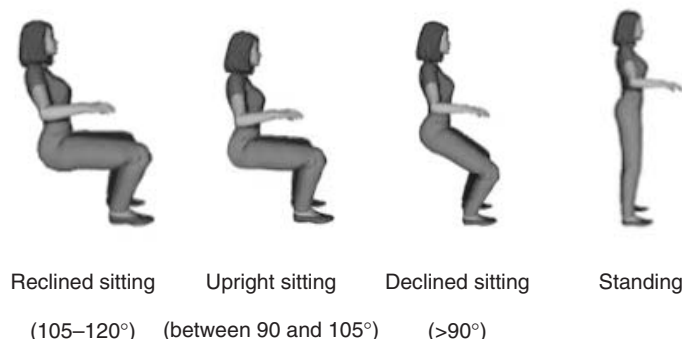
**Figure 3** Reference working postures. (Based on, and adapted from ANSI/HFES 100-2007; HFES, 2007.)

- The backrest should be contoured to conform with the curve of the lower spine.
- Chair fabric should allow ventilation.
- Be sure that repair service is readily available.

### 4.2.2  Minimizing Glare

- Position screen at right angles to windows.
- Adjust the tilt and swivel of the monitor.
- Reduce bright outside light by means of curtains, drapes, or blinds.
- Adjust lighting levels to the range of 200–500 lux (20–50 foot candles).
- Use parabolic diffuser grids or indirect lighting to help reduce overhead lighting glare.
- Provide work surfaces with an antiglare (matte) finish.
- Moveable task lights are often helpful.
- Screen filters and/or hoods can also be used if necessary.

### 4.2.3  Additional VDT Considerations

- *VDT Stands*. Height adjustability is preferred. Liquid crystal displays (LCDs) take up much less room and are much lighter than cathode ray tube (CRT) monitors.
- *Color Displays/Monitors*. Select a light highlight color that contrasts with the characters.
- *Black-and-White Monitor*. Rare these days, but select a light background and dark characters.
- *Flicker*. Screen should be readable with no perceptible flicker (rate at which images are "refreshed" on a screen from scanning of the electron gun). Not an issue with LCD flat-panel displays.
- *Printers*. Acoustical enclosures are recommended if sound levels exceed 55 dBA.
- *Ventilation*. Additional ventilation or air conditioning may be needed to overcome heat generated by many VDT workstations in a room. LCD flat-panel displays are much more energy efficient than CRTs.
- *Cables and Cords*. Should be concealed, covered, or out of the way.
- *Training*. Train all operators how to adjust chair, workstation height, and VDT position.

Computers in the workplace include desktop units on workstation furniture in office and work-at-home environments and laptop or notebook computers used virtually anywhere. Either way, discomfort associated with computer use can be traced to improper workstation adjustment and use. Surveys have shown that people who operate computers and VDTs are more comfortable and experience less discomfort when their workstations are adjusted properly. The importance of getting a good ergonomic match between the operator and the work is clear. But how do you create that match?

## 5  GETTING A GOOD ERGONOMIC MATCH BETWEEN OPERATOR AND WORKSPACE

### 5.1  Talk to Your Employees

An investment in office furniture with the latest ergonomic features can be wasted unless operators are taught to adjust their workstations correctly and unless management follows through to see that the adjustments are made. Keyboard work is demanding. Let your employees know that you are concerned with their comfort and you want to minimize the physical stress of working with the computer. Figure 4 shows the factors that you need to consider to ensure operator comfort.

### 5.2  Making Height Adjustments

Two methods are common for performing computer workstation assessments. They include observational techniques that estimate correct height through knowledge of "neutral" posture and direct measurement techniques.

### 5.3  Direct Measurement Techniques

Measure each operator individually to determine the appropriate height adjustments for their workstations. Seat the operator on a table or desk as shown in Figure 4, so that the edge of the tabletop just touches the back of the knees.

### 5.4  Operator Measurements (Dainoff and Dainoff, 1986)

*A = Knee Height*

Measure from the crease behind the knee to the bottom of the heel. Make sure the person is wearing the type of shoes normally worn on the job.

*B = Elbow Height*

Measure from a fixed surface, that is, tabletop, to the tip of the elbow. The person should be relaxed but sitting up straight. This measurement is easier if the person holds the upper arm against the body and reaches the hand toward the neck.
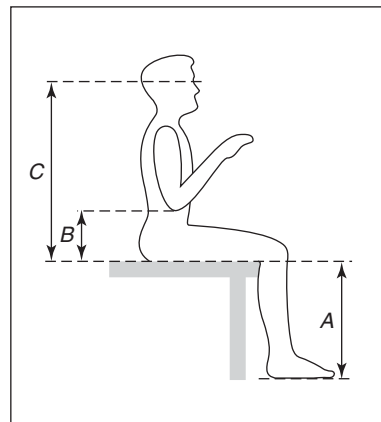


**Figure 4**  Operator measurements.

$C = Eye\ Height$

Measure from a fixed surface, that is, tabletop, to the eyes. Again, the person should be relaxed but sitting up straight.

### 5.5 Adjust the Chair Height

Once you have measured knee height ($A$), elbow height ($B$), and eye height ($C$), set the height of the chair front at knee height ($A$) initially. The seat pan may drop an inch or two when the operator sits down. If this is the case, raise the seat pan to offset the height change.

It is important the employee be trained on every chair adjustment feature. Some adjustment features are optional, for example, the seat pan adjustment.

Manufacturers may offer different size chairs, for example, small, medium, and large chairs to allow for longer legs. Make sure the employee has been fitted with the right chair.

If the seat is too high and cannot be lowered to the appropriate level, get a footrest and adjust the seat so that the vertical distance between the footrest and the front edge of the seat is equal to knee height ($A$).

If the seat pan has a tilt mechanism, the operator should tilt the seat to the most comfortable angle for work. In jobs that require a lot of data entry, such as word processing, some operators prefer a forward-tilted seat. For less-intensive keyboard work, many operators prefer a backward-tilted seat. Tilting the seat pan usually changes the height of the seat; readjust the front edge of the chair to knee height ($A$).

### 5.6 Position the Keyboard and Mouse

The center (or home) row of the keyboard should be adjusted to a height equal to knee height plus elbow height ($A + B$) above the floor, as shown in the lower portion of Figure 5. If a footrest is necessary, its height should also be added. The intent is to place the center row level with the tip of the elbow, thus keeping the forearms in a horizontal position.

If the keyboard height is not adjustable, raise or lower the chair height so that the difference in height between the chair seat and the keyboard is equal to elbow height ($B$). Provide footrests if needed.
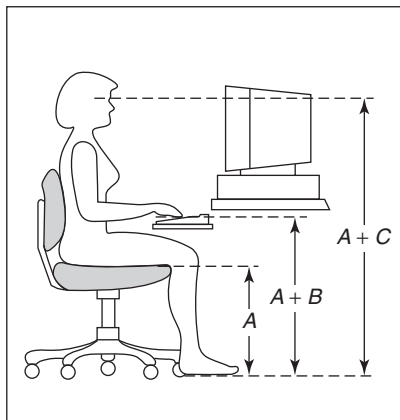
If the keyboard is thin (1–1.5 in.), place it about 2 in. back from the edge of the table. If the operator is using a thicker keyboard, provide a padded palm rest.

The mouse or input device should be at the same level as the keyboard. If using a keyboard tray, the tray should be wide enough to accommodate the mouse.

### 5.7 Position the Monitor

There are a variety of visual displays used in offices and they include CRT monitors and flat-panel LCD monitors. Configuration can be a single monitor or dual monitor set-up.

- Raise or lower the display so that the top of the screen is level with or slightly below the eyes, about equal to knee height plus eye height ($A + C$). If the operator wears bifocals or trifocals, a lower position may be more comfortable.

- Position the display at least 20 in. away from the operator's eyes or at arms length.

- For tasks in which the operator must read documents in addition to looking at the screen, move the visual display right or left of center to make room for a document holder (see Figure 5).

- Darken the screen while the operator checks for light reflectance or glare. Tilt the screen to eliminate as much glare or reflectance as possible. If the screen is right or left of center, moving it to the other side may help reduce glare.

### 5.8 Dual-Monitor Guidelines

- Both monitors should be matched in size and quality (luminance and contrast). If not matched in size, center viewing angle for documents on both screens should be the same.

- Flat-panel displays should not be paired with CRT monitors if possible.

- Both monitors should be placed at the same height and viewing distance. Viewing distance to each monitor should be a minimum of 20 in. or arms length away.

- Place both monitors as close to each other as possible.

- Provide adjustable monitor stands that are secure and allow for adjusting vertical height, screen tilt, and screen angle.

- Set up one monitor as the primary and the other as the auxiliary screen. Place the computer screen that is used more frequently closer to the center viewing angle and the auxiliary monitor to the side, left or right, and slightly angled toward the employee.

### 5.9 Laptop Computers

Laptop computers are no longer just for people who spend a large portion of their time away from a traditional office. Workers who rarely leave their office are using them too. Unfortunately this has led to complaints of back, neck, and wrist pain because the



**Figure 5** Desired workstation heights.

laptop is designed for portability, not ergonomics. With the keyboard and screen attached as one unit, the user must decide between a comfortable head and neck position or a comfortable wrist and arm position.

When discussing the use of laptop computers, there are two situations to consider:

1. An operator in an office environment with a docking station, external monitor, keyboard, and mouse
2. A mobile worker who uses laptops in airports, hotels, or offices without any external devices.

Operators in an office environment with an external device should follow the same height adjustment guidelines mentioned above. Mobile worker solutions are more challenging.

### 5.10 Positioning the Laptop

Positioning a laptop can be a challenge as placing the laptop low (in your lap or on a desk) for comfortable arm position means that you have to tilt your neck forward to view the screen; raising the screen to an acceptable level means that your hands are now reaching too high.

Some prefer placing the laptop on the work surface directly in front of the operator with the back elevated slightly to raise the display height. This can be accomplished inexpensively using specially designed laptop stands or three-ring binders with the binder at the back of the laptop. This also angles the keyboard, which may or may not be desirable. Tilting the screen too far may increase glare from overhead lights. Screen distance would follow the same guidelines as above.

Other operators prefer raising the entire laptop using a monitor stand or other means so the screen is at eye level and using an external keyboard and mouse. Inexpensive and portable monitor stands and external keyboards are readily available from mobile worker ergonomic accessory vendors and websites.

### 5.11 Other Considerations

Instruct the operator to:

Use a light touch when keying or using the mouse.

Use the index and middle fingers instead of the thumb to move the cursor via the touch screen. Move the hand toward the touch screen to eliminate stretching the fingers and alternate between hands.

Take short breaks every 20–30 min.

Use a bag with wheels when transporting the laptop. If the operator must carry the laptop, use a bag with a wide shoulder strap and alternate between shoulders.

Minimize the weight by carrying only what is needed. Reduce the number of peripherals such as disc drives and CD-ROM drives.

## 6   EYE STRAIN AND FATIGUE

While CTS may be the most infamous and possibly the most costly of all WMSDs, it is NOT the most prevalent malady of those who spend most of the working day interfacing with a computer. That distinction goes to yet another acronym, CVS, or computer vision syndrome.

The American Optometric Association (AOA) defines CVS as that "complex of eye and vision problems related to near work which are experienced during or related to computer use" and one that is very common among VDT workers (AOA, 2010). The following studies illustrate the importance of vision and visual fatigue on performance and safety associated with computer work:

- Visual symptoms occur in 50–90% of VDT workers while a National Institute for Occupational Safety and Health (NIOSH) study showed that 22% of VDT workers suffer from the more traditional musculoskeletal disorders (NIOSH, 1981).
- A survey of optometrists indicated that 10 million primary eye care examinations are provided annually in this country because of visual problems at VDTs (Sheedy, 1992).
- A 2000 NIOSH study concluded the addition of supplemental breaks such as a 5-min break during each hour decreased musculoskeletal discomfort, eye strain, and mood and did not affect performance in data entry workers (Galinski et al., 2000). A follow-up study published in 2007 provided further evidence that supplementary breaks reliably minimize discomfort and eyestrain without impairing productivity (Galinski et al., 2007).

On the human side of the VDT, the visual symptoms of an unsound interface have been broadly classified as "asthenopia," which is Greek for "MY EYES HURT," or as defined by the *Dictionary of Optometry and Visual Science*, "a subjective complaint of uncomfortable, painful and irritable vision" (Millodot, 1997).

On the process side of the VDT the symptoms of an unsound visual interface are work mistakes and lost productivity. A study that examined the relationship between the vision of computer workers and their productivity was conducted by the School of Optometry at the University of Alabama in Birmingham (Daum et al., 2004) and found a direct correlation between vision correction and process speed and accuracy.

### 6.1   CVS: Symptoms, Causes, and Controls

#### 6.1.1   Eye Muscle Fatigue/Strain

The energy for all structural movement in the body is provided by muscles and the movement of the eye is no different. The two primary muscle groups that are impacted by near work like viewing a computer screen are the large extraocular muscles that provide for the multidirectional movement of the eyeball and the smaller ciliary muscles that are attached to the lens capsule of the eye and provide the force necessary to change the lens shape when we focus on an object. Much of the eye irritation attributable to near-viewing activities like computer work is associated with the fatiguing and straining of these muscle groups.

### 6.1.2 Extraocular Muscle Strain/Fatigue

To understand how the extraocular muscles become fatigueg and strained, we must review the concept of the "resting point of convergence," which is that distance at which the eyes no longer need to cross to view an object. For most people the resting point of convergence is ≈40 in.

As the object we are viewing moves closer than 40 in., the medial rectus muscles will contract and pull the eyeballs inward toward the nose. This movement is called convergence and allows the eyes to maintain the alignment of the object we are viewing on the same place in both retinas, thus preventing double images. The closer the object, the stronger the contraction of the muscle and the more the muscle strain. Extended near viewing requires a prolonged contraction of the medial rectus muscles and increases the risk of muscle fatigue.

### 6.1.3 Ciliary Muscle Strain/Fatigue and Blurred Vision

To understand how the ciliary muscles become fatigue and strained, we must discuss yet another resting point, the "resting point of accommodation," which is defined as that point where the eye focuses when there is nothing to focus on. The resting point of accommodation will vary slightly from person to person but is ≈31 in. for young people and will increase with age. At this distance, the ciliary muscles, which focus the lens, will not be contracted.

As the object we are trying to focus on moves closer than ≈31 in., the small ciliary muscles will begin to contract to flex the lens into focus. The closer the focal point, the stronger the contraction of these muscles and the more opportunity for muscle strain. The risk of ciliary muscle fatigue also increases when our visual work area requires frequent changes of focal distances. These changes in focal distance are accommodated by the ciliary muscles repeatedly contracting and relaxing to change the shape of the lens and thus keep the visual object in focus.

### 6.2 Control Options for Eye Muscle Strain and Fatigue

#### 6.2.1 Monitor Distance

The closer the monitor, the more convergence and accommodation are required. By moving the monitor back, the load on both the extraocular and ciliary muscles will be reduced.

If you can see what you're looking at, the screen is not too far away. The only practical limit on how far away the monitor can be is the size of the letters and the workstation configuration. Fortunately many software programs allow us to change the font size, enabling us to write and edit with a larger font which is then changed before printing.

#### 6.2.2 Monitor Height

Because the resting points of convergence and accommodation both move inward with a downward gaze angle, lowering the monitor reduces the demand on these muscular systems. For example, a horizontal viewing angle has a resting point of convergence of ≈45 in. while a 40° downward viewing angle has a resting point of convergence of only 32 in.

#### 6.2.3 Focal Distance

- As much as possible keep all frequently accessed visual targets at a similar focal distance and in the same vertical plane.
- The use of a vertically oriented document holder helps to keep both the monitor screen and hard copy in approximately the same focal distance.

#### 6.2.4 Vision Breaks

Another way to reduce visual stress is to take "vision" breaks by looking at something that is well beyond your resting points of accommodation and convergence. We are all familiar with the term 20/20 as the descriptor of normal visual acuity. Just add another 20—20/20/20—and you have a great memory jogger for a visual work/rest regime.

Every 20 min take 20 s to focus on an object at least 20 ft away (Anshel, 1998).

## 7 MOBILE WORKERS: MANAGING SAFETY OF TELECOMMUTERS

Implementing a managed safety process is critical to optimizing the working environment of telecommuters, reducing the risk of claims and injury costs, and increasing profits. Key stakeholders inside and outside the organization are essential to the success of this program. Obtaining accurate and complete injury data and hazard information to effectively manage telecommuter safety is a challenge for managers. Three surveillance approaches are recommended by Robertson et al. (2003):

1. *Employee Reports*. Prompt reporting of hazards, injuries, or symptoms to the employer is important for treatment and prevention. However, some telecommuters are reluctant to do so, fearing that reporting work-related hazards or injuries may result in the cancellation of the telecommuting agreement. Rather than report a work-related injury, some may visit their personal physician and rely on health insurance to pay the bill.

2. *Review Existing Records*. Records such as workers compensation claims, reports, and OSHA logs provide valuable information. Check with your workers compensation (WC) insurer to make sure worker injuries occurring off-site are properly coded and tracked in your itemized loss statements.

3. *Job Surveys*. These include checklists and surveys dealing with hazards. Employers may not know what hazards exist in the home environment unless the worker voluntarily offers the information. Most companies rely on self-assessments of at-home workplaces.

**Table 3  Telecommuter Safety Program Evaluation**

1.  Do you offer guidelines for setting up a home office, including equipment and ergonomic accessories, and provide general recommendations?

2.  Do you have self-assessment surveys for ergonomics, computer workstations, and home hazards? If so, are these surveys administered online or by hard copy?

3.  What do you do with surveys after you receive them? What kind of follow-up exists to determine whether hazards are corrected?

4.  How is survey data collected, analyzed, and used for improving safety at off-site environments?

5.  Do you have a policy addressing what ergonomic accessories and office furniture will be paid for by the company?

6.  Do you offer training programs for work-at-home workers that include risk factors, ergonomic solutions, symptom recognition, and reporting? If so, are the training programs administered via the intranet, hard copy, or other means?

7.  Do you assess whether training is completed and learning has taken place?

8.  Is there a procedure for reporting computer and systems problems that impact the work-at-home employee? Are these problems promptly resolved? Do you know that for sure?

9.  How do work-at-home employees report symptoms and general health concerns they feel are work related? Do they feel they can do so without reprisal of job action? Is confidentiality of reports maintained?

10.  Does your WC insurer offer site coding in their claims databases for identifying injuries that occur to at-home or off-site workers? Do you use these data for determining safety and risk management priorities for off-site workers?

11.  Do you have a return-to-work strategy for disabled workers who work at home or off-site? Are workers able to receive quality health care? How do you know?

12.  Do work-at-home employees communicate regularly with their managers and peers, and are they kept current on company happenings?

If you have a safety program that addresses work-at-home employees, evaluate its effectiveness by answering the questions in Table 3 (Robertson et al., 2003). If an answer is "no" or "I don't know," target the item for improvement.

## 7.1  Tips for Working at Home

If you are considering a work-at-home policy, there are several issues to consider in order to maintain a safe and comfortable work-at-home environment.

### 7.1.1  Planning the Workspace

Identify a location that provides you with a physically separate workspace, preferably away from the flow of activity in your house. Interruptions by family members can be distracting.

When planning your space needs, a good rule of thumb for space allowance is to identify, at a minimum, a 6 × 6-ft space for your primary work area. Expect space requirements to grow depending on what you need for references or storage. Lateral files typically have a footprint of 36 in. × 18 in. while vertical files are 15 in. × 18 in. Bookcases require additional space.

Bookcases and filing cabinets should be placed so that one needs to stand up to access them. Walk around periodically. Do not sit continuously throughout the day. Plan movement into your office design and recognize that this adds to the space requirements.

Do not put your office in a small room without windows. A closed room needs two doors out for life safety. Ideally, you should have ready access to a view greater than 12 ft away. A window makes this easy. The longer view will allow the eye muscles to relax.

Avoid placing the computer next to a window. Windows that are close by create problems with visually demanding work because of the glare. It is best to find a space on a north wall. Be careful of extension cords and wiring that crosses the travel area, as they can produce trip and fall hazards. All cables and extension cords should be fastened up and out of the way.

Be sure you have a lockable door and can control entry into your work area. Try to have an understanding with family members or roommates that you need privacy to conduct business in a professional manner.

Your work area should have at least two means of egress. One way out can be a window if you have a safe means of getting from the window to the ground.

Select a location with access to sufficient electrical power outlets. If you have any questions about electrical supply, have a licensed electrician evaluate your needs and install additional outlets if necessary. Residential-type extension cords are not a good choice; look for a cord with a minimum of 14-gauge wire. If a power strip is used, look for types with surge and overload protection.

### 7.1.2  Selecting Furniture

Select your furniture carefully, especially your desk and chair. If your company provides furniture, know in advance where you intend to place it to be sure it will fit. If you are purchasing the furniture yourself, check with your manager or someone who is familiar with getting surplus furniture. Your desk will need to accommodate your computer, keyboard, phone, paper, references, stapler, sundry items like pen holders and paper clips, and possibly fax, CD drive, scanner, and printer; therefore desktop dimensions are important.

Watch out for the cheap office furniture in advertising fliers. This furniture offers little flexibility in monitor placement and adjustment. Those with cubby holes for the components can create problems if you have a large

terminal, want to use a document holder, or want to use a slant board to hold books or other large references. Sometimes leg space is inadequate as well.

If you have a typical VDT monitor, you will need a work surface with at least 30 in. depth. A work surface with less depth is going to create problems. It is not unusual to find that the depth of the terminal combined with the depth of the keyboard exceeds 24 in. In this case, you will need to install a keyboard support or tray.

Do not place the monitor to the side of the keyboard. This is a poor solution because your neck was not designed to be held in a twisted position and you will eventually begin to develop neck and shoulder pain.

The desk may have a fixed-height work surface or it may be adjustable. Adjustable is better because you will be able to set it at the correct height for you. Fixed-height desks or workstations are usually in the range of 28–29 in. This is a problem for many people. Some may find the keyboard is too high, even when using a standard office chair adjusted to its highest point. This requires an adjustable keyboard holder to bring the keyboard down to a comfortable position.

Keyboard trays or holders should be at least 26 in. wide and at least 10 in. deep or more. Keyboard holders or trays have some serious trade-offs. They are generally not as stable as a desk top and can be loose or bouncy. Trays push you away from the working surface and everything you have on that surface. The phone is harder to reach, you often have to stretch out your arm and get into awkward positions just to write, and you will find yourself leaning and stretching out to read documents.

Select a solid, substantial desk or workstation that does not tip over when loaded up or when an overloaded drawer is pulled out. Beware of raised edges, and look for good leg clearance (at least 17 in. deep at the knee) and a matte finish. Center-drawer desks are not a good choice because the drawer will not allow the keyboard to be adjusted to the correct height for you and still leave adequate leg clearance. A table is better, as long as the surface has cantilever support or is otherwise designed so there is no part of the frame impinging on leg room.

The chair is a critical component to your home office. A typical chair will create problems. Look for a commercial office chair with height adjustability, back-tilt mechanism, lumbar support, and a seat pan that is the right width and length. Select one wisely after trying some out. Most office chairs adjust in the range of 16–21 in. Even at 16 in., about 15% of the female population and 2% of the male population will need footrests.

Your chair should have a five- or six-point swivel base with wheels and a rounded or waterfall front edge. Some seat pans are strongly contoured; these can be a problem for some people. Be careful that armrests do not stop you from bringing yourself up close to the keyboard. If the chair has armrests, be sure they are neither too low, in which case you will be slumping in the chair all day, nor too high, in which case your shoulders will be raised into an unnatural posture. Armrest adjustability is preferable. The backrest should not be so wide that your elbows bump it.

Filing cabinets can be dangerous for young children. If upper drawers are pulled out and children climb on them, the cabinet can tip. Look for a means to secure the filing cabinet such as securing one to another or securing to a wall.

## 7.2 Using a Laptop Computer at Home

Many laptops lack the image clarity of a full-size VGA monitor and can create eye discomfort. Docking systems and simply attaching a full-size terminal are good solutions for those whose work requires a substantial amount of visual interaction with the screen. A full-size keyboard and mouse or other pointing device should be used as well.

If you are using a laptop, even with a detached keyboard, work surface depth of 24 in. should be suitable because the units are seldom more than 12 in. deep. The following steps can minimize the onset of eye fatigue and strain when using your laptop at home:

- Take "mini" breaks by focusing on a distant object for a few seconds before continuing work on your screen.
- Keep the screen clean at all times, using appropriate antistatic cleaning materials.
- It is better to make keyboard position your primary concern. If the keyboard is not separate, this will mean tipping the display back.
- Reflective lighting may be a source of annoyance for laptop users. Use drapes, shades, or blinds to control glare. Use indirect light whenever possible while avoiding intense or uneven lighting in your field of vision.
- Keep your head in a comfortable position, not overly turned or tilted. Adjust the screen brightness and contrast levels that allow you to comfortably view the screen. If you experience fatigue or visual discomfort after following these suggestions, consult an eye care specialist and inform that specialist of your computer use.

It is certainly a challenge to maintain comfortable hand and arm positions while using a laptop. The following recommendations may help:

- Change your position often to avoid discomfort and muscle fatigue. If you begin to feel uncomfortable, stop and rest.
- Take periodic breaks and stretch your arms, hands, and fingers. Many computer users find that frequent, short breaks are of greater benefit than fewer, longer breaks.
- Type with a light touch. Do not pound the keys. Make sure you are not pushing down on the keys harder than necessary.
- Keep your wrists in a straight, nonrigid position. Never position your wrists in an exaggerated angle or in a position that causes tension in your wrists.

- Your hands and wrists should be free to move when typing. Do not rest your wrists on a palm rest, a table, or your thighs while typing on a laptop.
- Keep your fingers relaxed and nonrigid when operating your laptop or an input device. Pay particular attention to your ring finger, pinkie finger, and thumb. Make sure you are not tensely holding them up in the air or scrunching them into the side of your hand when using your input device or typing.

Working with a laptop keyboard for long periods can be uncomfortable and fatiguing. Especially problematic is a laptop keyboard for someone who must work with numbers. A regular-size and configuration number pad as a peripheral is essential for those who work with numbers on a laptop computer.

### 7.3  Consider Your Environment

If you have a regular light-emissive terminal, the ambient lighting around the screen should not exceed 500 lux (50 foot candles). If you have a flat-panel display, you can increase the lighting to around 750 lux (75 foot candles). Bare incandescent bulbs do not make a visually comfortable workstation. Indirect fluorescent lighting or fluorescent lighting with diffusers that train the light directly downward are the best choice.

Avoid having any bright-light sources in your immediate field of view. The preferable location for light sources is behind you, over a shoulder at an angle, or at a right angle to you so that you do not see a reflection in the screen.

Walls and wall coverings should be nonreflective. Some walls have enamel paint or shiny wallpaper that can be very reflective. Avoid the impulse to put framed artwork or photographs in your immediate field of view because they tend to have a relatively high reflectance.

Most noise at home will come from televisions, stereos, and conversation. Demanding complete quiet while you work in the kitchen is unreasonable. Locating your office out of the mainstream of activity will allow your family or roommates to conduct normal lives.

The home office should have adequate ventilation. If the home has a forced hot-air system or central air, a duct should be in the work area.

Most home carpeting and carpet pads are softer and less durable than commercial carpeting used in offices. Your chair will not roll as easily and may be a problem for you to easily change position as you perform your tasks. A solid carpet protector can be helpful but can also be a problem if the chair rolls too easily.

If your office is below grade, have your work area tested for radon.

### 7.4  Making a Good Ergonomic Fit

Once you have installed your furniture and equipment, it is important to adjust your workstation to fit you.

A good ergonomic fit for your workstation includes a chair height adjustment which permits your feet to rest flat against the floor and the work surface for your keyboard to be about 1 in. lower than your elbow height.

If workstation design does not allow adequate adjustability for keyboard height, it may be necessary to adjust chair height to elevate your elbow about an inch above your keyboard, and support your feet with a footrest.

Position the monitor for a moderate downward gaze angle and at fingertip reach, or about 20 in. from your eyes. If you are a hunt-and-peck typist, it might be easier for you to have a closer, lower monitor so you are not moving your head and neck up and down. Eyes move fairly easily through an arc of about $30°$, so a fairly low monitor reduces $A$ and $A + B$ repeated neck motion. For those who touch type, a monitor at a higher position will probably be more comfortable.

A word on eye wear: A very common problem is presbyopia, the loss of the eye's ability to see close objects clearly. Presbyopia is usually corrected with bifocals or trifocals. If you are a touch typist or know the keyboard so well that you need not do more than glance at it occasionally, it would behoove you to get monocular lenses to replace the bifocals while working. The strength of the monocular lens should be set for the distance from your eyes to your screen. If you are a hunt-and-peck typist, special bifocal lenses for VDT use are a good option. In this case, the top lens is set for the terminal distance and the lower lens for the keyboard.

Document holders are often a case of personal preference. In most cases, the home worker or hoteler will be composing rather than transcribing, so it is often unnecessary to be concerned with a document holder. If your work involves a lot of transcription from a printed document, it will be very important to have a document holder. Generally, document holders are designed to be at the side of the terminal or between the terminal and keyboard. The location is a matter of personal preference but can be influenced by whether you are a touch typist, whether you have presbyopia, and what type of display you are using.

If your work surface is the wrong height and you decide to buy a keyboard holder or tray, be sure that you get one wide enough to hold your keyboard and mouse/mouse pad or track ball. This is usually about 28 in. unless you have a split keyboard or some other type that is wider than a standard expanded keyboard. Avoid situations where you must reach out for the mouse, especially with the shoulder raised. If you have multiple computers and terminals, it is best to get an "A–B" switch for your keyboard so you do not clutter the desktop with keyboards and mice.

Wrist rests are not for everyone and in some cases can be a problem. A wrist rest provides a soft place to relax the hands when not typing. It should not be used to support the hands while typing. Hands should be cupped and above the keyboard when typing while the wrist is straight or very slightly extended.

## 8  INTERVENTION STRATEGIES

Office ergonomics studies have revealed a variety of contributing factors to musculoskeletal and visual discomfort among computer users. These factors include

increased job demands and more hours working at a computer (e.g., Bernard et al., 1994; Faucett and Rempel, 1994), sustained awkward head and arm postures (Marcus and Gerr, 1996; Tittiranonda et al., 1999), increased levels of psychological stress and poor psychosocial work environment, (e.g., Bongers et al., 1993; Carayon and Smith, 2000; Faucett and Rempel, 1994; Marcus and Gerr, 1996), work organizational factors (e.g., Lassen et al., 2004; Punnett and Bergqvist, 1997), a lack of specific ergonomic features in the workstations and office buildings, and poor lighting (e.g., Daum et al., 2004; Nelson and Silverstein, 1998; Sauter et al., 1990). Typically, these studies are cross-sectional in design and only describe the work, safety, and health experiences of computer and office workers at one time period (Demure et al., 2000). Recently, the literature in workplace interventions intended to prevent or reduce musculoskeletal and visual symptoms among computer users has grown. However, few longitudinal field or lab studies have examined the effects of office ergonomics interventions on workers' health, safety, and performance (Brewer et al., 2006; Buckle, 1997; Karsh et al., 2001). Although there is a growing interest among employers to improve office workplaces, studies that investigated the effects of workstation, eyewear, and behavioral interventions on upper body musculoskeletal and visual symptoms are of mixed quality (Brewer et al., 2006; Karsh et al., 2001).

There is some evidence, however, that ergonomics training (Brisson et al., 1999) in workstation and building design (e.g., Aarås et al., 2001; Hagberg et al., 1995; Nelson and Silverstein, 1998; Sauter et al., 1990) can prevent or reduce musculoskeletal and visual discomforts and symptoms in office environments. One method for reducing the prevalence of musculoskeletal and visual symptoms is to provide specialized ergonomics training and workstation changes. Office ergonomics training helps employees to understand proper workstation set-up and postures (e.g., Bohr, 2000; Brisson et al., 1999; Ketola et al., 2002; Verbeek, 1991). Green and Briggs (1989) showed that merely providing adjustable furniture alone may not prevent the onset of overuse injury. However, a significant decrease in WMSDs and visual discomfort has been observed when workers were given an adjustable/flexible work environment coupled with ergonomics training (Amick et al., 2003; Robertson et al., 2008, 2009). Further, the provision of control over the work environment through adjustability and knowledge may enhance worker effectiveness as well as their health and safety (Hedge and Ray, 2004; McLaney and Hurrell, 1988; O'Neill, 1994; Robertson et al., 2008, 2009; Smith and Bayehi, 2003). Recent findings of a randomized control trial 15-day longitudinal laboratory study, replicating an 8-h customer service job, further supported these field intervention studies. It was observed in this study that the trained group who used sit/stand workstations exhibited minimal and significantly lower musculoskeletal and visual discomfort compared to a nontrained reference group also working in the same sit/stand configuration. Moreover, the trained group had significantly higher performance and effectiveness than the reference group as exhibited by

higher accuracy and quality control scores (Robertson et al., 2010). These findings, which indicate the ability to mitigate symptoms, change behaviors, and enhance performance through training combined with a sit/stand workstation, have implications for preventing discomforts in office and computer workers (Robertson et al., 2009). These results are also supported by the conclusions of a systematic review conducted by Brewer et al. (2006) with their critical and strict 11-point inclusion criteria of 31 studies. In this review only moderate evidence was observed for no effect of workstation adjustment alone and no effect of rest breaks and exercise on musculoskeletal or visual health. A positive effect of alternative pointing devices was found. They further concluded that for all other workplace interventions a mixed or insufficient evidence of effect was observed (Brewer et al., 2006).

When examining other office ergonomic intervention studies that may have weaker designs, such as nonrandomized control trial or no controlled group, positive and significant results have been observed consisting of various office ergonomics interventions (e.g., Dainoff et al., 1999; Hedge and Ray, 2004; Sauter et al., 1990; Smith and Bayehi, 2003; Smith and Carayon, 1996; Vink et al., 2009). Certainly these findings are limited in their generalizability due to lack of internal and external validity; however, they do provide insightful and useful information describing the effects of various office ergonomics interventions in either field settings or a control laboratory study with limited exposure variables consisting of real world issues. Other related areas of workplace interventions, such as participatory ergonomics interventions, have also shown mixed results (Rivilis et al., 2008). However, given mixed results of the effects of office and computer interventions, it appears that, when possible in this field research, having the ability to control some of the threats to validity can strengthen the study design and provide more definitive conclusions. Bridging laboratory with field intervention studies can also provide an important link between building conceptual models regarding office workers and associated risks and providing effective interventions and programmatic ergonomic recommendations. These and future studies grounded in high-quality research can all contribute to reducing and preventing symptoms among workers and providing an injury-free, comfortable, and productive work environment.

## 9 CONCLUDING REFLECTIONS

What are we to make of the current state of knowledge regarding office ergonomics? On the one hand, when rigorous methodological criteria are applied to the vast literature in the area, it is difficult to identify specific biomechanical risk factors in WMSDs. Hence, psychosocial variables are proposed as potential explanatory variables. And yet, we have an extended case study by the chief of experimental medicine at Beth Israel Deaconess Medical Center in Boston describing his own disabling wrist injury which he attributes to "banging clumsily at the (laptop) keyboard for many hours at a

time" (Groopman, 2007, Chapter 7). At the same institution, Boiselle et al. (2008) describe a prevalence rate of 58% of repetitive stress symptoms among radiologists who spend more than 8 h per day interacting with a standard radiological archiving and communication system (PACS). Symptom rates were reduced by introduction of and training on ergonomic workstations and chairs. The senior author of this chapter has similar personal knowledge of highly motivated professionals who suffered disabling wrist injuries attributed to prolonged computer use, injuries which were ameliorated by traditional ergonomic solutions. At the very least, these case studies would seem to question simple psychosocial explanations based on secondary gains through public expression of pain symptoms.

Two possible approaches to this dilemma are proposed for consideration.

### 9.1 More Complete Characterization of Exposure

It is conceivable that current Cochrane-based methodologies for attributing risk are more appropriate for traditional disease vectors than for describing the complexities of office work. An alternative approach to characterizing exposure relies on Ashby's (1956) principle of requisite variety, which states that the variety of the measurement system must match the variety of the system to be measured.

Using the terminology of Section 1., we might define a condition of maladaptive perception–action cycles if there is a mismatch of coordination of postural and supporting workstation degrees of freedom resulting in configurations of working postures in excess of some hypothetical three-dimensional spatial comfort zone (e.g., awkward posture), and such cycles proceed at a rate/pace in excess of a hypothetical temporal comfort zone. The causes of these presumed mismatches might be physically based (e.g., anthropometic variation, physical disability) and/or cognitive/behaviorally based (e.g., lack of understanding regarding adjustment mechanisms and the need to work at an appropriate work pace, lack of motivation to engage in appropriate adaptive behavior).

The outcome of such maladaptive PACS could be discomfort and pain. This may either resolve with the passage of time or progress to a medically significant event (diagnosis, compensation claim). Such pain might result from underlying tissue damage.

However, at the same time, we might consider that the impact of psychosocial factors on expressed pain might act through at least two pathways. The first involves the possibility that combinations of psychosocial factors might create and/or enable maladaptive perception–action cycles (e.g., increased work pace, lack of training, poor supervisory relationship, job content, equipment). The second is that emotional reactions to psychosocial stress have a direct influence on levels of pain (Marras et al., 2009). Unfortunately, the methodological challenges in subjecting the aforementioned complex nonlinear relationships to analysis by rigorous scientific method are considerable. Of particular difficulty is that the training and coaching required

to achieve adaptive control are also the same variables responsible for Hawthorne effects.

### 9.2 Focus on Quality and Performance

A second approach is to avoid the context of medical diagnosis and treatment and focus instead on achieving optimal and high-quality work performance. By this approach, maladaptive perception–action cycles are regarded as problems of management and work organization. As has been seen in Section 8, there is considerable evidence that the ergonomic practices discussed in Sections 3 through 7 can result in improvements in work performance and quality. Accordingly, reduction of discomfort/pain is a desirable management goal and, as such, is aligned with, rather than in conflict with, safety and health concerns (see, e.g., Dul and Neumann, 2009). In fact, in an extensive telephone survey of over 28,000 working adults, Stewart et al. (2003) determined that common pain at work is responsible for an annual loss of $61.2 billion in productivity in the United States, the bulk of which comes from reduced performance while at work. It might, therefore, be argued that use of ergonomic knowledge to provide an appropriate work system ought to be focused on optimizing overall work performance and quality and, having done so, the safety and health benefits are likely to result "for free."

## REFERENCES

Aarås, A., Horgen, G., Bjørset, H. H., Ro, O., and Walsøe, H. (2001), "Musculoskeletal, Visual and Psychosocial Stress in VDU Operators before and after Multidisciplinary Ergonomic Interventions. A 6 Years Prospective Study—Part II," *Applied Ergonomics*, Vol. 32, No. 6, pp. 559–571.

Accredited Standards Committee. (2002), *ASC Z365 Management of Work-Related Musculoskeletal Disorders* (Final Draft), National Safety Council, Washington, DC.

American Optometric Association (2010), *Computer Vision Syndrome* (*CVS*), available: http://www.aoa.org/x5374.xml, accessed September 29, 2010.

Amick, B. C. (2003), "Effect of Office Ergonomics Intervention on Reducing Musculoskeletal Symptoms," *Spine*, Vol. 28, pp. 2706–2711.

Anshel, J. (1998), *Visual Ergonomics in the Workplace*, Taylor & Francis, London.

Ashby, W. R. (1956), *An Introduction to Cybernetics*, Chapman & Hall, London.

Atroshi, I., Gummesson, C., Johnsson, R., Ornstein, E., Ranstam, J., and Rosen, I. (1999), "Prevalence of Carpal Tunnel Syndrome in a General Population," *Journal of the American Medical Association*, Vol. 282, pp. 153–158.

Bernard, B., Ed. (1997), *Musculoskeletal Disorders and Workplace Factors: A Critical Review of Epidemiologic Evidence for Work-Related Musculoskeletal Disorders of the Neck, Upper Extremity, and Low Back*, National Institute for Occupational Safety and Health, Cincinnati, OH.

Bernard, B., Sauter, S., Fine, L., Petersen, M., and Hales, T. (1994), "Job Task and Psychosocial Risk Factors for Work-Related Musculoskeletal Disorders among Newspaper Employees," *Scandinavian Journal of Work, Environment and Health*, Vol. 20, pp. 417–426.

Bernstein, N. A. (1996), "On Dexterity and Its Development," in *Dexterity and Its Development*, M. Latash and M. T. Turvey, Eds., Lawrence Earlbaum Associates, Mahwah, NJ, pp. 1–244.

Bohr, P. (2000), "Efficacy of Office Ergonomics Education," *Journal of Occupational Rehabilitation*, Vol. 10, No. 4, pp. 243–255.

Boiselle, P. M., et al. (2008), "Repetitive Stress Symptoms in Radiology: Prevalence and Response to Ergonomic Interventions," *Journal of the American College of Radiology*, Vol. 5, No. 8, pp. 919–923.

Bongers, P. M., de Winter, C. R., Kompier, M. A., and Hildebrandt, V. H. (1993), "Psychosocial Factors at Work and Musculoskeletal Disease," *Scandinavian Journal of Work, Environment and Health*, Vol. 19, pp. 297–312.

Brewer, S., et al. (2006), "Workplace Interventions to Prevent Musculoskeletal and Visual Symptoms and Disorders among Computer Users: A Systematic Review," *Journal of Occupational Rehabilitation*, Vol. 16, pp. 325–358.

Brisson, C., Montreuil, S., and Punnett, L. (1999), "Effects of an Ergonomic Training Program on Workers with Video Display Units," *Scandinavian Journal of Work, Environment and Health*, Vol. 25, pp. 255–263.

Buckle, P. (1997), "Upper Limb Disorders and Work: The Importance of Physical and Psychosocial Factors," *Journal of Psychosomatic Research*, Vol. 43, pp. 17–25.

Business and Institutional Furniture Manufacturers Association (BIFMA) (2002), *BIFMA G1-2002, Ergonomics Guideline for VDT (Visual Display Terminal) Furniture Used in Office Work Spaces*, BIFMA, Grand Rapids, MI.

Carayon, P., and Smith, M. J. (2000), "Work Organization and Ergonomics," *Applied Ergonomics*, Vol. 31, No. 6, pp. 649–662.

*Compact Edition of the Oxford English Dictionary* (1971), Oxford University Press, New York.

Dainoff, M., and Mark, L. S. (2001), "Affordances," in *Encyclopedia of Ergonomics*, W. Karwowski, Ed., Taylor and Francis, London.

Dainoff, M., Paasche, J., Simons, K., and Terlag, K. (1999), "Periodic Standing as a Relief from Fatigue in Office Work," in *Applied Ergonomics: Case Studies II*, D. C. Alexander, Ed., Institute of Industrial Engineers, Norcross, GA.

Dainoff, M. J. (2005), "Chairs and furniture," in *The Occupational Ergonomics Handbook: Interventions, Controls, and Applications in Occupational Ergonomics*, 2nd ed., W. Karwowski and W. S. Marras, Eds., CRC Press, Boca Raton, FL.

Dainoff, M. J. (2008), "Ecological Ergonomics," in *Ergonomics and Psychology*, O. Y. Chebykin, G. Z. Bedny, and W. Karwowski, Eds., CRC Press, Boca Raton, FL.

Dainoff, M. J., and Dainoff, M. H. (1986), *People and Productivity: A Manager's Guide to Ergonomics in the Electronic Office*, Holt Rinehart and Winston of Canada, Toronto.

Daum, K. M., et al. (2004). Productivity Associated With Visual Status Of Computer Users. *Optometry*, Vol. 75, No. 1, pp. 1–15.

Dembe, A. E. (1996), *Occupation and Disease—How Social Factors Affect the Conception of Work-Related Disorders*, Yale University Press, New Haven, CT.

Demure, B., Luippold, R. S., Bigelow, C., Ali, D., Mundt, K. A., and Liese, B. (2000), "Video Display Terminal Workstation Improvement Program, I. Baseline Associations between Musculoskeletal Discomfort and Ergonomic Features of Workstations," *Journal of Occupational and Environmental Medicine*, Vol. 42, pp. 783–791.

Dul, J., and Neumann, W. P. (2009), "Ergonomics Contributions to Company Strategies," *Applied Ergonomics*, Vol. 40, No. 4, pp. 745–752.

Faucett, J., and Rempel, D. (1994), "VDT-Related Musculoskeletal Symptoms: Interactions between Work Posture and Psychosocial Work Factors," *American Journal of Industrial Medicine*, Vol. 26, pp. 597–612.

Ferguson, D. (1971), "An Australian Study of Telegraphists' Cramp," *British Journal of Industrial Medicine*, Vol. 28, pp. 280–285.

Galinski, T. L., Swanson, N. G., Sauter, S. L., Hurrell, J. J., and Schliefer, L. M. (2000), "A Field Study of Supplementary Rest Breaks for Data-Entry Operators," *Ergonomics*, Vol. 43, No. 5, pp. 622–638.

Galinski, T. L., Swanson, N. G., Sauter, S. L., Dunkin, R., Hurrell, J. J., and Schliefer, L. M. (2007), "Supplementary Breaks and Stretching Exercises for Data Entry Operators: A Follow-Up Field Study," *American Journal of Industrial Medicine*, Vol. 50, No. 7, pp. 519–527.

Galison, P. (1997), *Image and Logic*, University of Chicago Press, Chicago.

Gerr, F., Marcus, M., and Monteilh, C. (2004), "Epidemiology of Musculoskeletal Disorders among Computer Users: Lesson Learned from the Role of Posture and Keyboard Use," *Journal of Electromyography and Kinesiology*, Vol. 14, No. 1, pp. 25–31.

Gerr, F., Monteilh, C. P., and Marcus, M. (2006), "Keyboard Use and Musculoskeletal Outcomes among Computer Users," *Journal of Occupational Rehabilitation*, Vol. 16, No. 3, pp. 265–277.

Gibson, J. J. (1950), *The Perception of the Visual World*, Houghton-Mifflin, Boston.

Gibson, J. J. (1979), *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston.

Green, R. A., and Briggs, C. A. (1989), "Effect of Overuse Injury and the Importance of Training on the Use of Adjustable Workstations by Keyboard Operators," *Journal of Occupational Medicine*, Vol. 31, No. 6, pp. 557–562.

Griffiths, K. L., Mackey, M. G., and Adamson, B. J. (2007). The Impact Of A Computerized Work Environment On Professional Occupational Groups And Behavioral And Physiological Risk Factors For Musculoskeletal Symptoms: A Literature Review. *Journal of Occupational Rehabilitation*, Vol. 17, No. 4, pp. 743–765.

Groopman, J. (2007). *How Doctors Think*. Boston, Houghton Mifflin.

Hagberg, M., Silverstein, B. A., Wells, R., Smith, M. J., Hendrick, H. W., Carayon, P., et al. (1995). *Work-Related Musculoskeletal Disorders (WMSDs): A Reference Book for Prevention*. London: Taylor & Francis.

Hedge, A., and Ray, E. J. (2004). Effects of an Electronic Height-Adjustable Work Surface On Computer Worker Musculoskeletal Discomfort And Productivity. Paper presented at the Human Factors and Ergonomics Society 48th Annual Meeting, Santa Monica, CA.

Human Factors and Ergonomics Society (HFES) (2007). *ANSI-HFES 100-2007 Human Factors Engineering of Computer Workstations*. HFES, Santa Monica, California:.

IJmker, S., Huysmans, M. A., Blatter, B. M., van der Beek, A. J., van Mechelen, W., and Bongers, P. M. (2007), "Should

Office Workers Spend Fewer Hours at Their Computer? A Systematic Review of the Literature," *Occupational and Environmental Medicine*, Vol. 64, No. 4, pp. 211–222.

Karsh, B., Moro, F. B. P., and Smith, M. J. (2001), "The Efficacy of Workplace Ergonomic Interventions to Control Musculoskeletal Disorders: A Critical Examination of the Peer-Reviewed Literature," *Theoretical Issues in Ergonomics Sciences*, Vol. 2, No. 1, pp. 23–96.

Ketola, R., et al. (2002), "Effects of Ergonomic Intervention in Work with Video Display Units," *Scandinavian Journal of Work, Environment and Health*, Vol. 28, No. 1, pp. 18–24.

Kusnetz, S., and Hutchinson, M. K., Eds. (1979), *A Guide to the Work-Relatedness of Disease (DHHS (NIOSH), rev. ed.,* Publication No. 79-116), U.S. Department of Health and Human Services, Centers for Disease Control, National Institute for Occupational Safety and Health, Cincinnati, OH.

Lassen, C. F., et al. (2004), "Elbow and Wrist/Hand Symptoms among 6943 Computer Operators: A 1-Year Follow-Up Study (the NUDATA Study)," *American Journal of Industrial Medicine*, Vol. 46, pp. 521–533.

Latash, M. L., and Turvey, M. T., Eds. (1996), *Dexterity and Its Development*, Lawrence Erlbaum Associates, Mahwah, NJ.

Lucire, Y. (2003), *Constructing RSI: Belief and Desire*, UNSW Press, Sydney.

Maeda, K. (1977), "Occupational Cervicobrachial Disorder and Its Causative Factors," *Journal of Human Ergology*, Vol. 6, pp. 193–202.

Marcus, M., and Gerr, F. (1996), "Upper Extremity Musculoskeletal Symptoms among Female Office Workers: Associations with Video Display Terminal Use and Occupational Psychosocial Stressors," *American Journal of Industrial Medicine*, Vol. 29, pp. 161–170.

Marras, W. S., Cutlip, R. G., Burt, S. E., and Waters, T. R. (2009), "National Occupational Research Agenda (NORA) Future Directions in Occupational Musculoskeletal Disorder Health Research," *Applied Ergonomics*, Vol. 40, No. 1, pp. 15–22.

McLaney, M. A., and Hurrell, J. J. (1988), "Control, Stress, and Job Satisfaction in Canadian Nurses," *Work and Stress*, Vol. 2, pp. 217–224.

Meister, D. (1989), *Conceptual Aspects of Human Factors*, Johns Hopkins University Press, Baltimore, MD.

Millodot, T. (1997), *Dictionary of Optometry and Visual Science*, 7th ed., Butterworth-Heinemann, Oxford.

National Research Council and Institute of Medicine (2001), *Musculoskeletal Disorders and the Workplace: Low Back and Upper Extremities*, National Academy Press, Washington, DC.

Nelson, N. A., and Silverstein, B. A. (1998), "Workplace Changes Associated with a Reduction in Musculoskeletal Symptoms in Office Workers," *Human Factors*, Vol. 40, No. 2, pp. 337–350.

Newell, K. M. (1996), "Change in Movement and Skill: Learning, Retention, and Transfer," in *Dexterity and Its Development*, M. Latash and M. T. Turvey, Eds., Lawrence Earlbaum Associates, Mahway, NJ.

National Institute for Occupational Safety and Health (NIOSH) (1981), *Potential Health Hazards of Video Display Terminals*, DHHS (NIOSH) Publication No. 81-129, NIOSH, U.S. Department of Health and Human Services, Centers for Disease Control, Cincinnati, OH.

Nubar, Y., and Contini, R. (1961), "A Minimal Principle In Biomechanics," *Bulletin of Mathematical Biophysics*, Vol. 23, pp. 377–391.

O'Neill, M. J. (1994), "Work Space Adjustability, Storage and Enclosure as Predictors of Employee Reactions and Performance," *Environment and Behavior*, Vol. 26, pp. 504–526.

Palmer, K. T., Harris, E. C., and Coggon, D. (2007), "Carpal Tunnel Syndrome and Its Relation to Occupation: A Systematic Literature Review," *Occupational Medicine (London)*, Vol. 1, pp. 57–66.

Punnett, L., and Bergqvist, U. (1997), "Visual Display Unit Work and Upper Extremity Musculoskeletal Disorders: A Review of Epidemiological Findings," in *National Institute for Working Life-Ergonomic Expert Committee*, Document No. 1, Arbetslivsininstitutet, Solna, Sweden, pp. 1–173.

Reed, E. S., and Bril, B. (1996), "The Primacy of Action in Development," in *Dexterity and Its development*, M. Latash and M. T. Turvey, Eds., Lawrence Erlbaum Associates, Mahwah, NJ.

Richter, J. M., Mathiassen, S. E., Slijper, H. P., Over, E. A., and Frens, M. A. (2009), "Differences in Muscle Load between Computer and Non-Computer Work among Office Workers," *Ergonomics*, Vol. 52, No. 12, pp. 1540–1555.

Rivilis, I., et al. (2008), "Effectiveness of Participatory Ergonomic Interventions on Health Outcomes: A Systematic Review," *Applied Ergonomics*, Vol. 39, No. 3, pp. 342–358.

Robertson, M. M., Maynard, W. S., and McDevitt, J. R. (2003), "Telecommuting: Managing the Safety of Workers in Home Office Environments," *Professional Safety*, Vol. 48, No. 4, pp. 30–36.

Robertson, M. M., Huang, Y. H., O'Neill, M. J., and Schleifer, L. M. (2008), "Flexible Workspace Design and Ergonomics Training: Impacts on the Psychosocial Work Environment, Musculoskeletal Health, and Work Effectiveness among Knowledge Workers," *Applied Ergonomics*, Vol. 39, No. 4, pp. 482–494.

Robertson, M. M., et al. (2009), "The Effects of an Office Ergonomics Training and Chair Intervention on Worker Knowledge, Behavior and Musculoskeletal Risk," *Applied Ergonomics*, Vol. 40, pp. 124–135.

Robertson, M. M., Ciriello, V. M., Garabet, A., (2010), "Enhancing Musculoskeletal Health and Performance through Training Combined with a Sit-Stand Workstation," Proceedings of the Seventh International Conference on Prevention of Work-Related Musculoskeletal Disorders, University of Angers, Angers, France.

Root-Bernstein, R. S. (1989), *Discovering*, Harvard University Press, Cambridge, MA.

Sauter, S. L., Dainoff, M. J., and Smith, M. J. (1990), *Promoting Health and Productivity in the Computerized Office*, Taylor and Francis, London.

Sheedy, J. E. (1992), "Vision Problems at Video Display Terminals: A Survey of Optometrists," *Journal of the American Optometric Association*, Vol. 63, pp. 687–692.

Silverstein, M. A., Silverstein, B. A., and Franklin, G. M. (1996), "Evidence for Work-Related Musculoskeletal Disorders: A Scientific Counterargument," *Journal of Occupational and Environmental Medicine*, Vol. 38, pp. 477–484.

Smith, M. J., and Bayehi, A. D. (2003), "Do Ergonomics Improvements Increase Computer Workers' Productivity? An Intervention Study in a Call Centre," *Ergonomics*, Vol. 46, No. 1, pp. 3–18.

Smith, M. J., and Carayon, P. (1996), "Work Organization, Stress, and Cumulative Trauma Disorders," in *Beyond Biomechanics: Psychosocial Aspects of Musculoskeletal Disorders in Office Work*, S. D. Moon and S. L. Sauter, Eds., Taylor & Francis, London, pp. 23–42.

Stewart, W. F., Ricci, J. A., Chee, E., Morgenstein, D., and Lipton, R. (2003), "Lost Productive Time and Cost Due to Common Pain Conditions in the US Workforce," *Journal of the American Medical Association*, Vol. 290, pp. 2443–2454.

Thomsen, J. F., Gerr, F., and Atroshi, I. (2008), "Carpal Tunnel Syndrome and the Use of Computer Mouse and Keyboard: A Systematic Review," *BMC Musculoskeletal Disorders*, Vol. 9, p. 134.

Tittiranonda, P., Rempel, D., Armstrong, T., and Burastero, S. (1999), "Effect of Four Computer Keyboards in Computer Users with Upper Extremity Musculoskeletal Disorders," *American Journal of Industrial Medicine*, Vol. 35, pp. 647–661.

Turvey, M. T., Shaw, R. E., and Mace, W. (1978), "Issues in the Theory of Action: Degrees of Freedom, Coordinative Structures, and Coordination," in *Attention and Performance*, Vol. VII, J. Requin, Ed., Lawrence Earlbaum Associates, Hillsdale, NJ, pp. 557–595.

Van Rijn, R. M., Huisstede, B. M., Koes, B. W., and Burdorf, A. (2009), "Associations between Work-Related Factors and the Carpal Tunnel Syndrome—A Systematic Review," *Scandinavian Journal of Work, Environment and Health*, Vol. 35, No. 1, pp. 19–36.

Verbeek, J. (1991), "The Use of Adjustable Furniture: Evaluation of an Instruction Program for Office Workers," *Applied Ergonomics*, Vol. 22, No. 3, pp. 179–184.

Vicente, K. J. (1999), *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*, Lawrence Erlbaum Associates, Mahwah, NJ.

Village, J., Rempel, D., and Teschke, K. (2005), "Musculoskeletal Disorders of the Upper Extremity Associated with Computer Work: A Systematic Review," *Occupational Ergonomics*, Vol. 5, No. 4, pp. 205–218.

Vink, P., Konijn, I., Jongejan, B., and Berger, M. (2009), "Varying the Office Work Posture between Standing, Half-Standing and Sitting Results in Less Discomfort," in *Ergonomics and Health Aspects, HCII 2009*, B.-T. Karsh, Ed., Springer-Verlag, Berlin, pp. 115–120.

Waersted, M., Hanvold, T. N., and Veiersted, K. B. (2010), "Computer Work and Musculoskeletal Disorders of the Neck and Extremity: A Systematic Review," *BMC Musculoskeletal Disorders*, Vol. 11, No. 1, p. 79.

Wahlström, J. (2005), "Ergonomics, Musculoskeletal Disorders and Computer Work," *Occupational Medicine* (*London*), Vol. 55, No. 3, pp. 168–176.

# CHAPTER 57

# HUMAN FACTORS AND ERGONOMICS IN HEALTH CARE

**Pascale Carayon, Bashar Alyousef, and Anping Xie**
**Department of Industrial and Systems Engineering**
**University of Wisconsin—Madison**
**Madison, Wisconsin**

The U.S. health care industry is going through major changes under increasing pressures from the public for higher quality and lower cost. A major part of the 2010 health care reform is the push toward health information technology, which is assumed to provide major patient care benefits. The Health Information Technology for Economic and Clinical Health (HITECH) Act provides resources for health care organizations to implement health information technology (IT). Other pressures are exerted on the health care industry to redesign its systems, structures, and processes to better meet the needs of society in a safe, reliable, and efficient manner. For instance, the Joint Commission, which is the major organization that accredits health care organizations in the United States, has a set of national patient safety goals, such as improving the accuracy of patient identification and improving the safety of using medications. The national patient safety goals are updated on a regular basis. Every health care organization that is accredited by the Joint Commission is expected to implement systems, structures, and processes for meeting the national patient safety goals. There has also been a call by the National Academy of Engineering and the Institute of Medicine (IOM) to create a partnership between health care and industrial and systems

engineering, including human factors and ergonomics, in order to address the problems experienced by the health care industry (Reid et al., 2005). Multiple reports issued by the IOM have clearly highlighted the need for human factors professionals and researchers to be part of the effort to redesign health care systems, structures, and processes (IOM, 2006; Institute of Medicine Committee on Data Standards for Patient Safety, 2004; Institute of Medicine Committee on Quality of Health Care in America, 2001; Institute of Medicine Committee on the Work Environment for Nurses and Patient Safety, 2004; Kohn et al., 1999). For instance, the IOM (2006) report on medication errors calls for the consideration of human factors principles in designing health information technologies that can prevent medication errors, such as computerized provider order entry or bar coding medication administration technologies.

Numerous studies provide information on the various quality-of-care problems experienced across the continuum of care in the United States as well as in other countries. McGlynn et al. (2003) conducted phone interviews with a representative sample of adults living in 12 metropolitan areas of the United States; participants shared information about their health care experiences. In addition, medical record data were available for a

subset of the respondents. The researchers found that overall participants received 55% of recommended care; the level of recommended care provided was pretty consistent across the different categories of care: preventive care (53.5%; e.g., mammographic screening for breast cancer, documentation of smoking status), acute care (53.5%; e.g., prophylactic antibiotics given on day of hip repair surgery), and chronic care (56.1%; e.g., diet and exercise counseling for diabetic patients). Another set of quality-of-care problems received major public attention with the IOM report "To Err is Human: Building a Safer Health System" (Kohn et al., 1999). According to this report, between 44,000 and 98,000 people die each year as a result of (preventable) medical errors. A Canadian study on adverse events in hospital patients found that 7.5% of hospital admissions had an adverse event, such as unplanned readmission and hospital-acquired infection or sepsis (Baker et al., 2004). Thirty-seven percent of the adverse events were judged to be preventable. The Centers for Disease Control and Prevention has estimated the overall annual direct medical costs of health care–associated infections to U.S. hospitals at between $28.4 and $33.8 billion (Scott, 2009). Based on current practice and technology, it has been estimated that prevention strategies can eliminate 20–70% of health care–associated infections. These studies and others provide significant information about the need to redesign health care systems and processes in order to improve quality of care.

This chapter reviews various ways that the human factors and ergonomics (HFE) discipline can provide useful information for analyzing and redesigning health care systems and processes. The chapter begins with a description of the health care industry, including the various factors contributing to complexity in health care. We then review issues related to defining and involving end users in health care system design and other HFE improvement efforts. Given the complexity in health care, a major section of this chapter examines various systems approaches to health care. A separate section of the chapter examines HFE of medical devices and information technology. Special attention is given to patient safety and the role of HFE in understanding contributing factors to medical errors and error reporting and recovery mechanisms.

## 1 CHARACTERISTICS OF HEALTH CARE INDUSTRY

It is important to understand the unique characteristics the health care industry in order to know how human factors and ergonomics concepts, models, and methods can be applied or need to be modified and adapted for optimal outcomes.

### 1.1 Health Services Industry

In 2008, the health services industry was the largest in the United States, providing 14.3 million jobs (Bureau of Labor Statistics, 2010). The industry is comprised of the following segments (Bureau of Labor Statistics, 2010):

- Hospitals that employ 34.6% of all workers, 72% of hospital workers working in institutions with more than 1000 workers

- Nursing and residential care facilities, nursing aides providing the majority of direct care
- Physician offices, with physicians and surgeons having private practices or working in groups of physicians with similar or different specialties
- Offices of dentists
- Offices of other health practitioners, such as chiropractors
- Home health care services—one of the fastest growing sectors of the economy
- Outpatient care centers, such as dialysis centers and free-standing outpatient surgery centers
- Other ambulatory health care services
- Medical and diagnostic laboratories

The health services industry has 40 occupations and professions in the following categories: (1) management, business, and financial occupations (4.3% of employment); (2) professional and related occupations (43.8%); (3) service occupations (32.2%); and (4) office and administrative support functions (17.7%). The employment in the health services industry is expected to grow by 22.5% between 2008 and 2018. Some of the fastest growing occupations include home health aides (growth of 53.3%), physician assistants (growth of 41.3%), physical therapist assistants and aides (growth of 35.9%), medical assistants (growth of 35.1%), and occupational therapist assistants and aides (growth of 33.0%). A number of factors contribute to the growth in employment in the health services industry (Sultz and Young, 2001). The aging population is a primary factor that increases the demand for workers who provide long-term care, such as nursing home care and home health care. The increasing implementation of medical and nonmedical technology has also implications for the number and skill requirements of the health care work force (Sultz and Young, 2001). Health care changes have shifted health care delivery sites from acute-care hospitals to ambulatory, home care, and long-term-care settings (Sultz and Young, 2001).

The health care industry is a "people-intensive" industry involving many different types and categories of workers, patients and their families, and communities and society-at-large. As Van Cott (1994, p. 56) says, "the health-care system is people-centered and people-driven." Therefore, the discipline of human factors and ergonomics has much to offer to improve the performance, quality, and safety of the health care system. The current health care system is very decentralized: It is comprised of a range of subsystems connected with each other and including caregivers and patients (Institute of Medicine Committee on Quality of Health Care in America, 2001). The subsystems include hospitals, community pharmacies, clinics, laboratories, and long-term facilities. Because of the variety of systems and subsystems, different goals, values, beliefs, and norms of behavior are at work in the health care systems (Van Cott, 1994).

### 1.2 Complexity of Health Care

Different dimensions of system complexity have been identified (Carayon, 2006; Perrow, 1984; Vicente, 1999): (a) large problem spaces, (b) social, (c) heterogeneous

perspectives, (d) distributed, (e) dynamic, (f) potentially high hazards, (g) many coupled subsystems, (h) automated, (i) uncertain data, (j) mediated interaction with computers, and (k) disturbances. Health care systems possess many of the characteristics of system complexity (see Table 1).

### 1.2.1 Breadth and Depth of Health Care

Health care is composed of many different elements. The International Statistical Classification of Diseases and Related Health Problems (most commonly known by the abbreviation ICD) includes 155,000 codes for various diseases, symptoms, and others.

The United States spends a large amount of its gross domestic product (GDP) on health care: in 2009, health care expenditures represented 17.33% of the GDP, or $2.5 trillion (Truffer et al., 2010). The health care expenditures are distributed across hospital care (31%), physician and clinical services (21%), prescription drugs (10%), nursing home and home health (9%), and other services (29%).

### 1.2.2 Health Care as a Sociotechnical System

Health care is basically a sociotechnical system in which people have a preponderant role (see Section 2), for example, as providers, patients, families, and purchasers. People are customers and consumers of health care, and people are producers of health care. Effective functioning of health care depends largely on people and the communication and coordination among various health care staff and patients.

### 1.2.3 Heterogeneous Perspectives in Health Care

Workers in health care come from different background and may have different values regarding health care, its delivery, and its quality and safety. For instance, a study by Thomas et al. (2003) shows the discrepancy in the attitudes of critical-care physicians and nurses with regard to teamwork. A total of 90 physicians and 230 nurses from eight nonsurgical intensive care units (ICUs) in six hospitals were surveyed. Thirty-three percent of the nurses rated the quality of communication and collaboration with physicians as high or very high, whereas 73% of the physicians rated the quality of communication and collaboration with nurses as high or very high. Physicians were more likely than nurses to agree with the statement that "Input from ICU nurses about patient care is well received in my unit." Such discrepancy in perceptions and attitudes between workers in health care can have numerous consequences, such as dissatisfaction and poor well-being, and may also affect expectations regarding performance and ultimately the quality and safety of care provided to patients.

### 1.2.4 Distributed Health Care

People involved in the delivery of health care may be located in different places. One type of long-term care in which people are geographically dispersed is home services provided by home health care agencies (Wunderlich and Kohler, 2001). The home is fast becoming the primary site of care for most persons with acute or chronic illnesses. Indeed, with the current trend toward ambulatory procedures and increasing technologies, one-half of patients once cared for in hospitals now receive their care at home. The home health arena is unique because it has so many workers of various skill levels who are dispersed over large geographical areas. Workers in home care function in a geographically distributed environment, in which few workers ever see other members of their team on a day-to-day basis. Home health care workers have very high turnover rates (Wunderlich and Kohler, 2001), which

**Table 1  Complexity of Health Care Systems**

| Dimensions of System Complexity | Examples in Health Care |
|---|---|
| 1. Large problem spaces | ICD system includes 155,000 codes for diseases and other symptoms<br>Health care expenditures = 17% of GDP in the United states |
| 2. Social | People intensive<br>People centered<br>People driven |
| 3. Heterogeneous perspectives | Different goals, beliefs, values, and behavior norms<br>Different cultures and subcultures |
| 4. Distributed | Geographical dispersion (e.g., home care)<br>Telemedicine |
| 5. Dynamic | Changes in medical knowledge and technology<br>Delay between ''action'' and ''outcome'' (e.g., primary care and preventive medicine) |
| 6. Potentially high hazards | Patient safety and medical errors |
| 7. Many coupled subsystems | Both tight and loose coupling |
| 8. Automated | High levels of automation in certain parts of health care (e.g., radiotherapy) |
| 9. Uncertain data | Imperfect information<br>Imperfect knowledge<br>Patient factors |
| 10. Mediated interaction with computers | Medical devices and technologies (e.g., endoscopic technologies) |
| 11. Disturbances | Unanticipated events |

may be due to their poor working conditions and low quality of working life (Feldman, 1993; Wunderlich and Kohler, 2001). Stone and Wiener (2001) highlight the major issues affecting the long-term-care frontline workers, including difficulties in recruiting and retaining workers. They also discuss many of the negative effects of turnover, such as poorer quality and/or unsafe care, increased stress and frustration on the workers, reduced opportunities for on-the-job training and learning, and less peer support. Reasons for the turnover problem of home health care workers include poor working conditions, perceptions of inadequate resources, and a suboptimal environment for providing good safe care (Aiken et al., 2001; Blegen, 1993; Simmons et al., 2001). The discipline of human factors can provide the models and tools for improving working conditions and therefore improving retention and reducing turnover of home care workers. HFE can also be applied to understanding and improving quality and safety of home health care (Henriksen et al., 2009).

Telemedicine is one form of organizing care that puts distance between the patient and the health care providers. According to the American Telemedicine Association (2011, para. 1), "Telemedicine is the use of medical information exchanged from one site to another via electronic communications to improve patients' health status". A recent application of telemedicine for intensive care has been rapidly expanding (Lilly and Thomas, 2010). The tele-ICU care model can complement existing on-site ICU care activities. Patients are monitored by the tele-ICU staff (e.g., an intensivist and a critical-care nurse). There is much discussion about whether and how telemedicine such as tele-ICU can achieve positive results with regard to patient outcomes (quality of care) and cost, and very little is known about the HFE aspects of telemedicine. Such geographical separation poses unique challenges to communication and coordination between health care providers, which would benefit from input and advice from the HFE community.

### 1.2.5 Health Care as a Dynamic System

Complex systems are dynamic and rapidly changing. The health care industry has seen lots of changes in medical knowledge and technology that have been precipitated by large investments in biomedical research (Institute of Medicine Committee on Quality of Health Care in America, 2001). Much of the medical knowledge and information on evidence-based practice comes from randomized controlled trials (RCTs). Chassin (1998) shows that the number of publications from RCTs referenced in Medline has increased exponentially from 1966 to 1995: in 1966, there were about 100 RCT articles, whereas in 1995, there were over 100,000 RCT articles. The volume and complexity of this information pose unique challenges to health care practitioners and managers and add to the complexity of the health care system. Another time factor of system complexity is the time lag between action and response. Some health care action can lead very quickly to a response. For instance, using an electric heart defibrillation leads to an immediate physical reaction on the part of the patient.

The administration of certain drugs can also lead to immediate physiological reactions. There can also be long delays between actions done to the patient and their consequences. In addition, the consequences may not be easily linked to specific actions or may not be visible to the health care provider who performed the actions in the first place. This is particularly the case in primary and ambulatory care. For instance, the health effects of preventive services provided by primary care physicians may not be measured for many years.

### 1.2.6 Health Care as a Hazardous Industry

Operating a complex sociotechnical system can produce various hazards. In health care, the main hazards are those done to the patients, such as medical errors and lack of patient safety (see Section 5).

### 1.2.7 System Coupling in Health Care

Perrow (1984, pp. 93-94) defines the following characteristics of coupling in systems. First, "tightly coupled systems have more time-dependent processes," whereas "In loosely coupled systems, delays are possible." Second, the sequencing of tasks or process steps in tightly coupled systems is somewhat fixed and invariant. Third, in tightly coupled systems, there is typically only one way of designing and performing a process. One characteristic of loosely coupled systems is equifinality, that is, they have a common objective that can be achieved with different processes and tasks. Fourth, tightly coupled systems have little slack and bugger. A health care system can either be tightly coupled or loosely coupled (Cook, 2004). An example of tight coupling is the sequence of specific steps in the performance of a surgical procedure, for example, the surgery cannot occur before the anesthetic has been administered. An example of loose coupling is the preoperative process once a surgery has been decided: A number of tasks must occur, but not necessarily in a specific tight temporal sequence; for instance, the physical exam that the patient needs to have before the surgery can occur within a certain window of time, typically 1–3 weeks.

### 1.2.8 Automation in Health Care

Complex systems tend to be highly automated. Whereas automation is not widespread over the health care industry, there are parts of health care with high degree of automation. Radiotherapy is an example of patient care that relies on different forms of automation, such as connection between information from imaging devices and treatment devices. See Section 4 for a discussion of medical devices and information technology.

### 1.2.9 Uncertainty in Health Care

Uncertainty is another characteristic of complex system and is highly present in health care. The sources and types of uncertainty in health care include imperfect information, imperfect knowledge regarding medical treatment, and patient factors (e.g., impact of treatment on a particular patient). Uncertainty increases when patient-related information is not available or not provided in a timely manner (Schultz et al., 2007).

### 1.2.10 Technology-Mediated Interaction in Health Care

In health care, much of the interaction is mediated by devices and technologies. For instance, it is not possible to measure directly blood pressure of a patient during a surgery: There is a piece of equipment and a display that provide information on the patient's blood pressure. That type of technology-mediated interaction between the "worker" and the "work object" highlights the importance of cognitive ergonomics in health care. See Section 4 for a discussion of medical devices and information technology.

### 1.2.11 Unanticipated Events and Disturbances in Health Care

In complex systems, disturbances are very present and workers need to deal with unanticipated events. In order to maintain safety and performance, workers need to adapt to those events quickly. Health care, of course, is filled with unanticipated events to which workers need to react quickly in order to ensure the safety of the patients and maintain adequate performance. Weinger et al. (2003) conducted a study of nonroutine events in anesthesia. A nonroutine event was defined as "any event that is perceived by clinicians or skilled observers to deviate from ideal care for that specific patient in that specific clinical situation (Weinger et al., 2003, p. 110)." In-person surveys of anesthesia providers show that 27% of anesthesia cases ($N = 277$) had at least one nonroutine event (total of 98 nonroutine events). Data from quality improvement reporting systems in two hospitals yielded information on 135 events. An analysis of all of these 233 nonroutine events produced information on the factors contributing to those events: patient disease/unexpected response (67%); provider supervision, knowledge, experience, and judgment (33%); surgical issues (26%); logistical or system issues (19%); inadequate preoperative patient preparation (17%); equipment failure or usability (16%); coordination/communication (15%); and patient positioning (9%).

The complexity of the health care system has much impact on the practice of HFE. For instance, HFE projects in health care require one to spend more time on managing and implementing change than on deciding the content of change (Hignett, 2003).

### 1.3 Standardization

The report published in 2001 by the IOM on "Crossing the Quality Chasm" (Institute of Medicine Committee on Quality of Health Care in America, 2001) and the 1999 IOM report on "To Err is Human: Building a Safer Health System" (Kohn et al., 1999) emphasize the importance of standardization of care processes, which is believed to have the potential to reduce medical errors and improve quality of care (Field and Lohr, 1900; Reiling and Chernos, 2007). It is unclear whether we need more or less standardization (Brunsson and Jacobsson, 2000). Pros and cons of standardization are discussed in the literature. The IOM considers standardization an important redesign strategy that is assumed to reduce reliance on short-term memory and

allow those unfamiliar with a given process to use it safely (Kohn et al., 1999). Hospital administrators may use standardization to optimize their control over the hospital's increasingly complex structure (Timmermans and Berg, 2003). On the other hand, a survey of members of the American College of Physicians showed that although guidelines were thought to be able to improve the quality of care by 70% of respondents, 43% of respondents believed they would increase health care cost and 34% of respondents believed they would make medical practice less satisfying (Tunis et al., 1994). Critics argue that health care standardization will repeat the mistakes of scientific management by stifling professionals' creativity and autonomy, undermining clinical expertise, and rendering the profession vulnerable to oversight, substitution, and interference (Timmermans and Berg, 2003). In an environment with preset rules and regulations, patients may become numbers and interact with impersonal technologies and technicians, and health care workers bemoan the removal of mystery or excitement from their work lives (Reiser, 1978).

Timmermans and Berg (2003) proposed a useful approach to health care standardization, which is neither a "well-oiled machine" nor a "stifled robotscape." They stressed the importance of the content of standardization (i.e., what is standardized) and the manner in which standardization is designed and implemented (i.e., how it is standardized). With regard to content of standardization, several elements have been discussed, such as the physical design of health care settings, the design of medical devices, and health care procedures.

The physical design of a health care setting can have an impact on the quality of care. When leaders at St. Joseph's Community Hospital were working on the design of the new hospital facility, they looked for ways of designing the hospital and its spaces to ensure maximizing efficiency and quality and safety of care (Reiling et al., 2004). One of the hospital design principles was the complete standardization of patient rooms. According to Reiling and Chernos (2007), standardization of patient rooms should take every detail into consideration, including the location of the gas outlets, the bed controls, the cupboard where the latex gloves are stored, the charting process, and the switches on light fixtures. However, even if the ultimate goal is standardization, the design of patient rooms may vary because of physical constraints or design intentions (Saucier, 2010). Therefore, it is critical to identify and prioritize the elements of patient rooms that will benefit most from standardization. This type of standardization of the physical layout of health care settings can reduce the need for health care workers to memorize the location of equipment and supplies; this may contribute to efficiencies as well as reduce the likelihood of skill-based errors.

In the context of medical devices, lack of standardization may cause problems for users (Ward and Clarkson, 2007). For instance, an audit of the range of infusion devices in six National Health Service (NHS) hospitals shows that, on average, each hospital uses 31 different types of infusion devices [National Patient

Safety Agency (NPSA), 2004]. In a study of clinical needs assessment of intravenous (IV) therapy, standardization of IV pumps and accessories was a recommended strategy to reduce medication errors (Scroggs, 2008). Standardization of medical devices in a health care organization may support end users' performance: If there is only one type of each medical device, end users have to learn to use only this one type of medical device. In some instances, it may be important to allow end users to customize the interface of medical devices. A study of anesthesia alarms shows that tailoring can lead to improved performance if there is time to tailor the alarm, the means for adapting the alarm are present, and the benefits appear to outweigh the costs (Watson et al., 2004). Randell (2003) conducted an observational study of the customization of medical devices by ICU nurses. She found that nurses performed the following types of device customization: (1) customization to overcome limitations of the device and provide adequate patient care, (2) use of pen and paper to ease the use of the device (e.g., post-it notes attached to devices), and (3) change in the procedure for the usage of the device. Further research is needed to understand the safety and performance benefits and problems associated with standardization versus customization of medical devices.

Standardization of health care procedures is a highly controversial issue. Evidence-based medicine is the main attempt to accomplish this standardization at the level of the profession; this has led to an increasing number of clinical practice guidelines being implemented (Amalberti and Hourlier, 2007). By encouraging the use of current best evidence in making decisions about patient care, evidence-based medicine is assumed to lead to better and more efficient care, improved health care outcomes, better educated patients and clinicians, a scientific base for public policy, a higher quality of clinical decisions, and better coordinated research activities (Timmermans and Berg, 2003). However, many studies argue that guidelines may do little to change practice behavior (Donaldson et al., 1999; Grol et al., 1998; Woolf et al., 1999). One of the main reasons for noncompliance with clinical practice guidelines is that individual clinical autonomy takes precedence over the normative and prescriptive aspect of the guidelines (Timmermans and Berg, 2003). Ambiguities in the content of the guidelines (i.e., what is to be done, exceptions to the guidelines) and responsibility for the guideline implementation can also contribute to poor compliance with guidelines (Gurses et al., 2008).

In addition to what is standardized (content of standardization), it is important to consider the manner in which standardization is implemented. The IOM recommends an approach where systems and processes are designed for the usual, but the unusual is being recognized and planned for (Institute of Medicine Committee on Quality of Health Care in America, 2001). Similarly Walker and Carayon (2009) proposed a balanced approach to standardization. They suggested that, in addition to supporting appropriate standardization of routine tasks, value-added processes support intentional variation based on the uncertainty inherent in the patient's condition, the strength of the available evidence, patients' needs, local factors, and providers' professional judgment.

In summary, standardization may be an important principle of health care system design. The content of what can be standardized and how standardization can be implemented are two important elements of standardization. A balance between standardization and customization needs to be achieved; more research is needed to understand the HFE benefits and costs of standardization versus customization.

## 2 END USERS IN HEALTH CARE

The involvement and participation of users is a critical principle of human factors and ergonomics. In this section, we discuss some of the challenges to the application of this principle in health care: definition and determination of the users, involvement of laypersons who do not have medical/health care knowledge, and challenges related to participatory ergonomics.

### 2.1 End User Involvement in Health Care System Design

An overriding principle of human factors is to center the design process around the user, therefore creating a user-centered design (Meister and Enderwick, 2001; Norman, 1988). In the design of health care systems, the variety of potential end users needs to be considered in the design cycle.

### 2.1.1 System Design Process

There is much controversy about the system design process and its characteristics and components. However, the system design process can be conceptualized as being organized around four questions (Meister and Enderwick, 2001): analysis of the design problem, generation of alternative solutions, analysis of alternative solutions, and selection of preferred solution. From a human factors point of view, Wickens et al. (2004) describe major stages of system design in which human factors can provide important useful information: (1) front-end analysis, (2) iterative design and test and system production, (3) implementation and evaluation, and (4) system operation and maintenance and system disposal.

- *Front-end analysis*, including definition of the users, of the functions to be achieved by the system, of the environmental conditions under which the system will be used, and of the users' preferences or requirements for the system. This stage will typically include user analysis and task analysis.

An example of front-end analysis is the human factors and a macroergonomic approach used to study and analyze ultrasonic central venous catheter (CVC) guidance, placement, and care (Alvarado et al., 2008). In this study, ICU staff (physicians, nurses, and other technical personnel) were asked to participate in

individual observations and to complete a questionnaire assessing their current knowledge of CVC placement and care. The objective of this analysis was to understand the ICU work system and the CVC insertion tasks performed and to produce useful information to develop recommendations for redesigning CVC insertion work system and care processes in the ICU.

- *Iterative design and test and system production*: Initial specifications are used to create initial design or prototypes. At this stage, human factors input typically consists of identification of human factors criteria to the list of system requirements (e.g., usability requirements), function allocation, and design of support materials.

At the stage of creating and testing initial system specifications, it may be useful to conduct a more extensive HFE analysis, such as a proactive risk assessment (Carayon et al., 2009a). For instance, a hospital conducted an FMEA (failure mode and effects analysis) to examine the potential safety implications of implementing a new Smart IV pump technology (Wetterneck et al., 2006). This FMEA identified a number of design and implementation issues that were addressed before the new device was implemented (Carayon et al., 2008; Wetterneck et al., 2006).

- *Implementation and evaluation*: Various methods for system change implementation use basic human factors principles (e.g., participatory ergonomics). The evaluation should consider human factors variables (human performance, health and safety, and well-being).

Examples of human factors evaluations of system redesign include the evaluation of electronic health record (EHR) implementation in small clinics (Carayon et al., 2009b) and the evaluation of computerized provider order entry (CPOE)/EHR implementation in ICUs (Hoonakker et al., 2010). Both of these evaluations use multiple HFE data collection methods, including observational methods, interviews with key personnel, focus groups, a survey questionnaire, as well as other methods to assess medication errors and adverse drug events.

- *System operation and maintenance and system disposal*: Various human factors activities occur at those stages, for example, ensuring the reliability and functioning of medical equipment and devices for safe operations and designing an appropriate system for hazardous materials (e.g., needles).

After the implementation of a new system or a new technology, there may be lingering or new human factors issues that are discovered during the actual use phase. For instance, a study of nurses' use of bar code medication administration (BCMA) technology led to the discovery of many human factors problems, some of which are related to the design of the technology (Carayon et al., 2007). Data were collected using structured observations of medication administration and short interviews with nurses. Overall nurses were satisfied with the BCMA technology, but several potentially unsafe tasks and work-arounds were discovered. Recommendations for continued technology redesign and enhanced training came out from this post–technology implementation analysis.

### 2.1.2 End Users of Health Care Systems

Human factors engineers and ergonomists focus on the interactions between humans and other elements of the work system. In health care systems, the humans are varied: the health care providers and clinicians, the patients and their families, and other types of workers (e.g., housekeeping, biomedical engineering, purchasing, and administration). This large variety adds to the complexity of health care systems. For example, a single device such as an infusion pump is used by multiple users: The *nurse* programs the pump when administering medication to a patient; the *patient* is connected to the infusion via tubing; and the *biomedical engineer* maintains the pump and ensures its calibration. This example shows that a single device is used by different users performing different tasks. It is also important to note that this variety of end users is often related to a variety of physical and organizational settings. In the infusion pump example, the physical environment in which the pump is used varies from patient rooms to engineering laboratory. From an organizational viewpoint, the various users have probably received different level and type of training with regard to the usage of the pump.

Defining the "user" in a HFE project in health care is critical (Hignett, 2003). Every person is a potential user of the health care system in any country, but only a small proportion of a country's population is directly in "contact" with the health care system. Another complicating factor is the fact that very often the patients are not directly paying for the "health care service." All of these factors make the definition of the user a difficult task for the user-centered designer (Hignett, 2003). A structured review from a social science perspective of the published literature about user involvement in health care technology development and assessment from 1980 to 2005 in peer-reviewed journals (Syed et al., 2008) found that (1) the users of medical devices include clinicians, patients, care providers, and family members; (2) different kinds of medical devices are developed and assessed by user involvement and persons with different disabilities and impairments; (3) the user involvement occurs at different stages of the medical device technology life cycle and the degree of user involvement is in the order of design stage, testing and trials stage, deployment stage, and concept stage; and (4) methods most commonly used for capturing users' perspectives are usability tests, interviews, and questionnaire surveys.

## 2.2 Involvement of Laypersons

Recently, providing patient-centered care has been a focus of many health care organizational restructuring and quality improvement efforts. Patient-centered care has the potential to improve health status and can increase the efficiency of care by reducing diagnostic tests and referrals (Lutz and Bowers, 2000). The concept of patient-centered care covers two dimensions: (1) reorganization of services around patients' needs, requirements, wants, and expectations and (2) understanding patient-perceived needs, priorities, and expectations for health care (Lutz and Bowers, 2000). Patient-centered care should ensure access and continuity of health care, increase opportunities for patients to participate in the care process, provide self-management support, and coordinate care between settings (Bergeson and Dean, 2006).

Because of the increasing shift toward patient involvement in the care process, there has been much interest in examining the use of health care technologies by laypersons. For instance, automated defibrillators can be found in a variety of places, such as airports and other public places. Such devices need to be designed for people who do not have medical and clinical training. Callejas et al. (2004) show that naïve users and video-trained users were able to safely used two types of automated defibrillators.

Home care is a crucial extension of the health care system. Many health care activities occur at home, including self-help activities (following a balanced diet, maintain a balance of rest and activity, and engaging in the mental stimulation necessary to promote cognitive function) and self-care activities (primary and secondary preventions activities recognized and endorsed by health professionals). There are many people involved in home care, such as patient, family members, professional caregivers, nurses, physical therapist, and nursing assistants. Many home health care activities are carried out by the patient only or aided by family members, these laypeople are expected to collaborate with professionals in order to get adequate information and share roles in decision making and critical responsibilities to carry out health and healing practices (Zayas-Caban and Brennan, 2007). Therefore, the principles of HFE play a vital role to improve patient safety in home care through understanding the conditions and environment at home, laypeople cognitive and physical capabilities, communication methods, tools and technologies, and involvement of laypeople in designing and implanting home care tools and technologies (Henriksen et al., 2009; Zayas-Caban and Brennan, 2007).

## 2.3 Participatory Ergonomics in Health Care

Participatory ergonomics is a powerful method for involving the end users in system design (Wilson, 1995). Participation has been used in a variety of human factors processes, such as implementing ergonomic programs (Wilson and Haines, 1997). According to Noro and Imada (1991), participatory ergonomics is a method in which end users of ergonomics (workers, nurses, patients) take an active role in the identification and analysis of ergonomic risk factors as well as the design and implementation of ergonomic solutions. Participatory ergonomics was used to involve nurses in a process of developing and evaluating a nursing bag system for home care nurses (Lee et al., 2006). The participatory ergonomics approach began with seeking input from the entire nursing population that would be affected and then forming a working group to represent this population. The working group carried out an iterative process to develop a prototype for the nursing population to evaluate.

Evanoff and colleagues have conducted studies on participatory ergonomics in health care (Bohr et al., 1997; Evanoff et al., 1999). In one study, they examined the implementation of participatory ergonomics teams in a medical center (Bohr et al., 1997). Three groups participated in the study: a group of orderlies from the dispatch department, a group of ICU nurses, and a group of laboratory workers. Overall, the team members for the dispatch and the laboratory groups were satisfied with the participatory ergonomics process, and these perceptions seem to improve over time. However, the ICU team members expressed more negative perceptions. The problems encountered by the ICU team seem to be related to the lack of time and the time pressures due to the clinical demands. A more in-depth evaluation of the participatory ergonomics program on orderlies showed substantial improvements in health and safety following the implementation of the participatory ergonomics program (Evanoff et al., 1999). The studies by Evanoff and colleagues demonstrate the feasibility of implementing participatory ergonomics in health care but highlight the difficulty of the approach in a high-stress, high-pressure environment, such as an ICU, where patient needs are critical and patients need immediate or continuous attention. More research is needed to develop HFE methods for implementing participatory ergonomics programs in health care. Those programs should lead to improvements in human and organizational outcomes (e.g., reduced work-related musculoskeletal disorders) as well as improved quality and safety of care. This research should consider the high-pace, high-pressure work environment of health care.

## 3 HUMAN FACTORS SYSTEMS APPROACHES APPLIED TO HEALTH CARE

Human factors experts working in health care agree on the need to adapt and adopt systems approaches in health care systems (Bogner, 2004; Cook and Woods, 1994; Vincent, 2004). In this section, we review selected human factors systems approaches that have been applied to health care.

## 3.1 Work System Model

The work system model developed by Carayon and Smith describes the many different elements of work (Carayon and Smith, 2000; Smith and Carayon-Sainfort, 1989). The work system is comprised of five elements: the *individual* performing different *tasks* with various
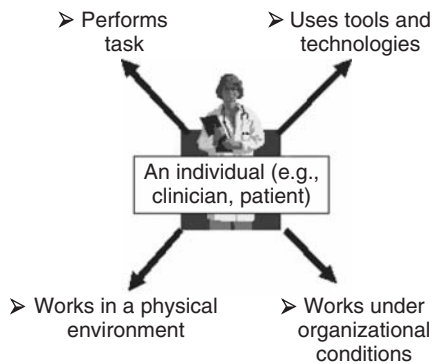
**Figure 1** Adapted version of the work system model (Carayon and Smith, 2000; Smith and Carayon-Sainfort, 1989).

*tools and technologies* in a *physical environment* under certain *organizational conditions* (see Figure 1).

Given the complexity of the health care system (see Section 1), it is important to adopt a systems approach to the analysis of health care systems (Vincent, 2004). The work system model (Carayon and Smith, 2000; Smith and Carayon-Sainfort, 1989; Smith and Carayon, 2000) and its extension, the SEIPS [Systems Engineering Initiative for Patient Safety] model of work system and patient safety (Carayon et al., 2006b), has been applied to the analysis of medical errors, such as wrong-site surgery (Carayon et al., 2004a). An example of the application of the work system model is provided below: see Example 1 for how the work system model can be applied to the analysis of an ICU nurse. Gurses and colleagues provide additional information about the work system elements and performance obstacles experienced by ICU nurses (Gurses and Carayon, 2007, 2009; Gurses et al., 2009).

*Example 1: Work System Analysis of an ICU Nurse:* The following is a brief overview of the different work elements of an ICU nurse job.

*Task.* The tasks performed by the ICU nurses include (but are not limited to) direct patient care, continuous patient status assessment, carrying out physician orders, medication administration, and family interaction.

*Organizational Factors.* There is a range of organizational factors that are important to understand the job of an ICU nurse. Conflict among nurses and between physicians and nurses has been correlated with high stress and workload in ICUs (Gray-Toft and Anderson, 1981). The studies by Knaus, Rousseau, Shortell, Zimmerman, and colleagues have shown the importance of "caregiver interaction," which is a composite concept that includes several dimensions, such as communication and coordination (Knaus et al., 1986; Shortell et al., 1994).

*Environment.* Noise and other sensory disruptions abound in the modern ICU setting (Topf, 2000). The physical environment is often crowded and messy, with no one available to help with immediate clean-up of the

environment or equipment. The noise, the housekeeping, the level of constant activity, the size of the rooms or physicians' and nurses' personal space (if any), patients/staff coming and going, and crowds of people waiting to get a moment of the physician's or the nurse's time and attention may all make the physical environment more difficult to carry out tasks.

*Equipment and Technology.* The technology, tools, and equipment of modern ICUs have been identified as possible causes of errors and problems (Bracco et al., 2000). The availability of needed supplies, the type of supplies, tools, technology desired, the working condition of the equipment, and the new technology available or unavailable are but some of the tools and technology issues that can increase workload. Additionally, training and time for acclimation are needed to learn all the new tools and technology.

*How to balance the work system of an ICU nurse, in particular with regard to workload?*

As an example in the ICU setting, in efforts to reduce workloads and balance the overall work system, the physicians and nurses might review how often physical assessments are performed on the patients and who performs them. Typically, both the nurse and the physician perform a patient physical assessment every hour, or as needed by the patient's condition. Under this system, both the physician and the nurse perform the patient physical assessment and enter it into the patient's records. The process takes the health care provider at least several minutes or more, depending on the patient's status, out of every hour. This takes time away from other tasks and/or professional activities associated with the patient's care. In addition, others, such as specialty consultation services, may be waiting to review the patient record currently in use for the patient assessment. To balance this problem, the ICU physicians and nurses may redesign the patient assessment system based on clinical expertise and cooperation among the physicians and the nurses involved.

### 3.2 Care Processes

As highlighted in the SEIPS model of work system and patient safety (Carayon, et al., 2006b) and in the classic model of quality of care by Donabedian (1988), care processes are critical for understanding quality-of-care outcomes. According to Donabedian (1988), a care process is what is being done to the patient. A care process can be analyzed using the work system model (Carayon et al., 2006b; Smith and Carayon-Sainfort, 1989): It involves various people (e.g., physician, nurse, patient) who perform a range of care tasks (e.g., nurse administering medication) using various tools and technologies (e.g., physician using a stethoscope) in a physical environment (e.g., patient room) under certain organizational conditions (e.g., coordination between primary-care physician and specialist). Therefore, the performance of care processes is influenced by the various work system characteristics of different people involved in the process.

As an example, the outpatient surgery process is comprised of major steps, including patient work-up prior

to day of surgery, five steps occurring on the day of surgery (i.e., patient admission and preparation, patient surgery, first-stage patient recovery, second-stage patient recovery and discharge), and patient recovery at home (Carayon et al., 2006a). At each stage of the process, various performance obstacles occur that can affect the quality of care provided to the surgery patients. A survey of outpatient surgery staff highlighted communication to patient and coordination among various providers as the major performance obstacles (Carayon et al., 2006a). A follow-up study examined the specific phase of preoperative outpatient surgery and identified a range of facilitators and obstacles to information flow in the care transitions (Schultz et al., 2007).

Care transitions in the patient journey can be particularly vulnerable because of information flow problems (Coleman, 2003; Coleman and Berenson, 2004). However, care transitions also represent an opportunity for error detection and recovery (Cooper, 1989; Perry, 2004; Wears et al., 2003). Given the many challenges presented by care transitions, there needs to be further research to understand the functions of care transitions and their role in patient safety and quality of care. Human factors research in this area is only beginning (Patterson et al., 2004; Patterson and Wears, 2010) and should be further encouraged.

### 3.3 Levels of System Analysis

There is increasing recognition in the human error literature of the different levels of factors that can contribute to human error and accidents (Rasmussen, 2000). If the various factors are aligned "appropriately" like "slices of Swiss cheese," accidents can occur (Reason, 1990). According to Rasmussen (1997), the sociotechnical factors involved in safety include (1) government, (2) regulators and associations, (3) company, (4) management, (5) staff, and (6) work. These different levels of sociotechnical factors interact and influence each other to produce accidents. Karsh and Brown (2010) have proposed a macroergonomic model of patient safety that highlights the need to describe, measure, and analyze different system levels.

### 3.4 Health Care Team as a System

Over recent decades, teams have been a strategic choice for many organizations in various areas such as aviation, military, industry, health care, nuclear power plants, and engineering project teams; many of these teams carry out complex and difficult tasks (Salas et al., 2008). Work teams are defined as "small groups of interdependent individuals who share responsibility for outcomes for their organizations" (Sundstrom et al., 1990, p. 120).

In health care, people work together in a variety of teams, for example, multidisciplinary teams caring for patients with specific clinical conditions (e.g., ICU team) and operating room teams. Members of the teams often come from different disciplines and educational background (Kohn et al., 1999). For instance, multidisciplinary rounds (MDR) represent a patient-centered model of care that can facilitate interdisciplinary collaboration in the ICU that is a highly complex and dynamic work setting where interdisciplinary collaboration can

have a significant impact on patient outcomes. Teamwork can take different forms (e.g., autonomous work groups versus manager-led teams) and be implemented in various modes (e.g., temporary versus permanent teams) (Sainfort et al., 2001b).

The 1999 IOM Report on "To Err is Human—Building a Safer Health System" (Kohn et al., 1999) recommends the establishment of team training programs for personnel in critical-care areas, such as an emergency department, ICU, and operating room. The report published in 2001 by the Institute of Medicine on "Crossing the Quality Chasm" (Institute of Medicine Committee on Quality of Health Care in America, 2001) goes one step further and emphasizes the need for developing effective teams in order to meet the six challenges of providing safe, effective, efficient, personalized, timely, and equitable care.

The HFE discipline has contributed significantly to the science and practice of teams, teamwork, and team performance (Salas et al., 2008). Salas and colleagues (2008) summarized the following six major findings in the area of team performance over the past five decades from a human factors perspective: (1) shared cognition is important in team performance; (2) shared cognition can be measured; (3) team training promotes teamwork and enhances team performance; (4) team performance can be modeled; (5) researchers have defined factors that influence team performance such as shared collective orientation; and (6) well-designed technology can improve team performance. There has been significant human factors research on teamwork in health care. For example, Guerlain et al. (2002) have developed a system for evaluating performance of surgery teams in the operating room. Helmreich and colleagues (Helmreich and Merritt, 1998; Helmreich and Schaefer, 1994) have examined the performance of operating room teams.

The Department of Defense (DoD) and the Agency for Healthcare Research and Quality (AHRQ) developed a systematic approach of Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS™) to integrate teamwork into practice and improve team performance in health care (King et al., 2008). Accordingly, TeamSTEPPS™ is learning material designed to improve patient outcomes by educating teamwork among health care providers (Guimond et al., 2009). Figure 2 shows the TeamSTEPPS instructional framework (King et al., 1999).

## 4 HFE OF MEDICAL DEVICES AND INFORMATION TECHNOLOGY

In health care, technologies are often seen as an important solution to improve quality of care and reduce or eliminate medical errors (Bates and Gawande, 2003; Kohn et al., 1999). These technologies include organizational and work technologies aimed at improving the efficiency and effectiveness of information and communication processes (e.g., computerized order entry provider systems and electronic medical record systems) and patient care technologies that are directly involved
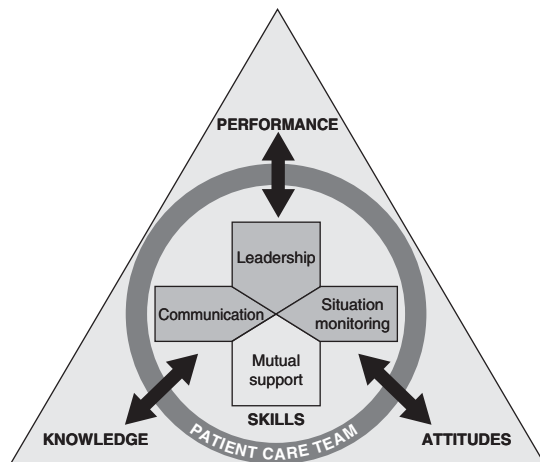
**Figure 2** TeamSTEPPS framework for health care teams (King et al., 2008)

in the care processes (e.g., bar code technology or smart infusion pump technology for medication administration). The 1999 IOM report recommends adoption of new technology, such as bar code technology, to reduce medication errors (Kohn et al., 1999). However, implementation of new technologies in health care has not been without troubles or work-around. For example, the study of Patterson and colleagues (2002) shows some of the negative side effects of bar code medication administration technology, such as degraded coordination between nurses and physicians. Technologies can change the way work is being performed, and because health care work and processes are complex, negative consequences of new technologies are possible (Cook, 2002). In this section, we focused on the following HFE issues related to medical devices and information technology: (1) design of technology (e.g., usability), (2) impact of technology on the work system, and (3) implementation of technology.

### 4.1 Health Care Technology Design

The human factors characteristics of the health care technologies' design should be studied carefully (Battles and Keyes, 2002). An experimental study by Lin et al. (2001) showed the application of human factors engineering principles to the design of the interface of an analgesia device. Results showed that the new interface led to the elimination of drug concentration errors and the reduction of other errors. A study by Effken et al. (1997) shows the application of a human factors engineering model, that is, the ecological approach to interface design, to the design of a hemodynamic monitoring device. New health care technologies may bring their own "forms of failure" (Battles and Keyes, 2002; Cook, 2002; Reason, 1990). For instance, bar coding technology can prevent patient misidentifications, but the possibility exists that an error during patient registration may be disseminated throughout the information system and may be more difficult to detect and correct

than with conventional systems (Wald and Shojania, 2001).

New digital technologies such as surgical navigation or robotic systems are changing the clinical working systems dramatically. The risk–benefit assessment of those emerging technologies is difficult. Such an assessment has to consider the utilization of the technologies and their impact on task completion as well as the long-term impact, such as the patient's condition five years after intervention. For instance, Cao and Taylor (2004) examined the impact of the introduction of a new robotic technology on performance and communication patterns in the operating room (OR) team. The new technology, a remote master–slave surgical robot, removes the surgeon from the surgical site. Results of the human factors analysis show large differences in amount and type of information required by the surgeon in order to accomplish the procedure with and without the robot. In the robotic work system, the surgeon has additional tasks to perform and additional decisions to make, therefore increasing the cognitive load.

One important health care technology design characteristic is the usability of a medical device. To what extent does the medical device perform in a safe, easy-to-use manner? For a given task and a unified group of users, usability is a measure of the fast (efficient), simple (effective), and satisfying use of a technical device (Bevan et al., 1991; Dumas and Redish, 1993; Ravden and Johnson, 1989; Staggers, 2003). In order to assess usability, various methods of usability engineering have been developed (Mayhew, 1999; Nielsen, 1992, 1993; Whiteside et al., 1988; Wiklund, 1993a). For the evaluation of the usability of medical devices, specific methods have been used, such as user questionnaire, usability inspection methods, and usability tests (Wiklund, 1993b, 1995). With the major push toward health information technology implementation throughout the U.S. health care system, the issue of usability of health information technologies such as electronic medical record and computerized provider order entry has received increasing attention (Koppel and Kreda, 2010; Li et al., 2006).

### 4.2 Health Care Technology Implementation

It is important to address issues of usability and usefulness of health care technology; however, in order to ensure the technology is used successfully and effectively, we also need to address implementation issues (Carayon and Karsh, 2000; Davis, 1993; Karsh, 1997; Venkatesh et al., 2002). The manner in which a new technology is implemented is as critical to its success as its capabilities and characteristics (e.g., usability) [see, e.g., Eason (1982) and Smith and Carayon (1995)]. For instance, inadequate planning when introducing a new technology designed to decrease medical errors has led to technology falling short of achieving its patient safety goal (Kaushal and Bates, 2001; Patterson et al., 2002). To promote end-user acceptance and effective use of technology, it is imperative to apply a human factors systems approach to technology implementation (Karsh, 2004). This proactive approach emphasizes simultaneous design of the technology and the work system, which aims at

achieving a balanced work system and preventing errors from happening in the first place.

### 4.2.1 Impact of Health Care Technology on the Work System

Whenever implementing a technology, it is essential to consider the work system factors that affect technology acceptance and satisfaction in the implementation plan (Karsh, 2004; Karsh and Holden, 2007; Smith and Carayon, 1995). We need to examine the potential positive and negative influences of the technology on the work system (Battles and Keyes, 2002; Sheridan and Thompson, 1994; Smith and Carayon-Sainfort, 1989). The implementation of technology in an organization has both positive and negative effects on the job characteristics that ultimately affect individual outcomes (quality of working life, such as job satisfaction and stress, and perceived quality of care delivered or self-rated performance) (Carayon and Haims, 2001). In a study of the implementation of an EHR system in a small family medicine clinic (Carayon et al., 2009b), a number of issues were examined: impact of the EHR technology on work patterns, employee perceptions related to the EHR technology and its potential/actual effect on work, and the EHR implementation process (Carayon and Smith, 2001). Employee questionnaire data showed that employees perceived increased dependency on computers and a small increase in perceived quantitative workload was found. This result was confirmed by the work analysis data, which indicated a dramatic increase in the amount of time spent on computers by the various job categories. While the EHR implementation did not change the amount of time spent by physicians on patient care, the work of clinical and office staff changed significantly. For clinical and office staff, the main differences between the pre- and the post-EHR implementation were decreases in time spent on distributing charts, transcription, and other clerical tasks.

### 4.2.2 Principles of Health Care Technology Implementation

The most common reason for failure of technology implementations is that the implementation process is treated solely as a technological problem, and the human and organizational issues are ignored or not recognized (Eason, 1988). Lorenzi and her colleagues (2009) discussed the implementation of an EHR system in small ambulatory practice settings. They divided the implementation process into several stages, including decision, selection, preimplementation, implementation, and postimplementation. Through this process, the authors emphasized the importance of developing a flexible change management strategy; identifying a champion; assessing and redesigning workflow; understanding financial issues; conducting training; and evaluating the implementation process. In a similar vein, Blumenthal and Epstein (1996) comment that "the science of behavior modification" has not been applied much in the health care field. It is important to consider the human and organizational aspects that can hinder or foster technological change and use in health care systems.

The way change is implemented (i.e., the process of implementation) is central to the successful adaptation of organizations to changes (Karsh, 2004; Korunka et al., 1993; Tannenbaum et al., 1996). A "successful" technology implementation can be defined by its "human" and organizational characteristics: reduced/limited negative impact on people (e.g., stress, dissatisfaction) and on the organization (e.g., delays, costs, reduced performance) and increased positive impact on people (e.g., acceptance of change, job control, enhanced individual performance) and on the organization (e.g., efficient implementation process). Various principles for the successful implementation of technological change have been defined in the human factors and ergonomics and business research literature (see Table 2).

*Employee participation* is a key principle in organizational change (Coyle-Shapiro, 1999; Korunka et al., 1993; Smith and Carayon, 1995). There is research and theory demonstrating the potential benefits of participation in the workplace. Benefits include increased employee motivation and job satisfaction, enhanced performance and employee health, more rapid implementation of technological and organizational change, and more thorough diagnosis and solution formation for ergonomic problems (Gardell, 1977; Lawler, 1986; Lawler III, 1986; Noro and Imada, 1991; Wilson and Haines, 1997). End-user involvement in the design and implementation of a new technology is a good way to help ensure a successful technological investment. Korunka and his colleagues (Korunka and Carayon, 1999; Korunka et al., 1993, 1997) have empirically demonstrated the crucial importance of end-user involvement in the implementation of technology to the health and well-being of end users. Previous research has made the distinction between *active participation*, where the employees and the end users are actively participating in the implementation of the new technology, and *passive participation*, where the employees and end users are *informed about and communicated with* regarding the new technology (Carayon and Smith, 1993).

*Communication* among end users, decision makers, and technical support is also important. It promotes the transmission of information about the implementation to end users to reduce uncertainty and enables end users to obtain quick *feedback* to questions about new technologies through the implementation process (Karsh and Holden, 2007). Feedback is an important element in order to change behavior (Smith and Smith, 1966) and has been emphasized as an important organizational design element in the health care literature (Evans et al., 1998; McDonald et al., 1996). The importance of feedback in managing the change process is also echoed in the literature on quality management [see, e.g., the plan–do–check–act cycle proposed by Deming (1986)].

Technology implementation in health care systems should be considered as an evolving process that requires considerable *learning* and adjustment (Mohrman et al., 1995). An aspect of learning in the context of technological change is the type and content of *training* (e.g., Frese et al., 1988; Gattiker, 1992). Training, which is designed to promote the transfer of knowledge and skills into work practice, can improve

**Table 2 Principles of Technology Implementation**

| Principles of Technology Implementation | Description | Examples in Health Care |
|---|---|---|
| Employee participation | Extent to which health care staff is involved in various decisions and activities related to the technology implementation (active participation)<br>Information/communication about the technology implementation (passive participation) | Involve nurses and pharmacists in implementing a bar code medication administration system |
| Communication and feedback | Extent to which health care staff is kept informed of the technology implementation through various means of communication and extent to which feedback is sought after/during the technology implementation | Develop structured communication networks between supervisors and health care staff to deal with the new technology<br>Provide feedback to health care staff to show them that their ideas are taken seriously |
| Learning and training | Extent and nature of the training provided to the health care staff and extent of learning by the health care staff<br>Use of simulation techniques | Design a science-based training program to train certified nursing assistants in nursing homes<br>Create a new simulated health care work system with new processes and structures<br>Use a simulator to help physicians try out new functions on a new computerized order entry system |
| Top-management commitment | Extent to which top management directly participate in the technology implementation | Conduct structured program for technology implementation and show health care staff where responsibility for the different aspects of the changes lie<br>Make resources available to the implementation of each of the other principles |
| Project management | Activities related to the organization and management of the technology implementation itself<br>Use of pilot testing | Analyze the current health care system into which the new technology will be implemented<br>Facilitate the implementation team to design and implement the change process<br>Implement pilot testing to debug the new technology in the context of use in nursing units, operating rooms, or outpatient clinics |

self-efficacy and demonstrate usability and usefulness of new technologies (Karsh and Holden, 2007). A questionnaire survey of 244 family practice residents' perceptions regarding the use of EHR showed that those residents who felt that the EHR-related training was adequate were more likely to report benefits due to the EHR, such as decreased time to review past records, increased documentation accuracy, and increased consistency of health maintenance (Aaronson et al., 2001). *Simulation* is an effective method of training delivery in health care (Issenberg et al., 1999; Nishisaki et al., 2007; Woodward et al., 2010). It is suggested to be designed into the implementation process as a means for planning, training, and continuous learning (Karsh and Holden, 2007).

The prerequisite to implementing all principles discussed above is the *commitment of top management*. True top-management commitment enables resources available to promote other strategies of implementation design and to ensure a successful implementation (Karsh, 2004; Karsh and Holden, 2007). An important indicator of top-management commitment is the presence of a structured program for implementation (Smith and Carayon, 1995).

It is suggested that *project management* concepts and methods (e.g., project structure, roles, timeline) be utilized for technology implementation (Carayon et al., 2009b; Lorenzi et al., 2009). Keys to a successful implementation project include (1) analyzing needs and preferences of medical providers and key administrators; (2) selecting a strong physician leader to champion the

project; (3) hiring a project manager with dedicated time to lead the project; (4) forming a project leadership team of key personnel from clinical, office, and information system staff; (5) gathering needs of other users early in the planning process; and (6) obtaining buy-in by clinicians and office staff early in the process. In many implementations of health care technology, *pilot testing* is conducted to identify additional problems that are not uncovered during proactive system analysis. Besides, it is also applied to the design of technologies themselves and other support systems to ensure successful implementation.

## 5   PATIENT SAFETY AND MEDICAL ERRORS

The discipline of HFE has much to offer to the understanding, reduction, and prevention of medical errors and therefore the improvement of patient safety (as well as employee safety) (Bogner, 1994).

### 5.1   Patient Safety

The safety of health care is very much discussed across the world, for example, in Australia (McNeil and Leeder, 1995) and the United Kingdom (U.K. Department of Health, 2002). In the United States, the 1999 publication of a report by the IOM has raised the level of public awareness regarding medical errors and patient safety (Kohn et al., 1999).

A 2000 report published by the U.K. Department of Health provides some data on the extent to which the English health care system fails to provide high-quality, safe care (U.K. Department of Health, 2002). About 400 people die or are seriously injured in adverse events involving medical devices. About 10,000 people report having experienced serious adverse reactions to drugs. The U,K, National Health Service pays around 400 million a year settlement of clinical negligence claims. Data for the United States indicate that "Preventable adverse events are a leading cause of death in the United States" (Kohn, et al., 1999, p. 26). It has been suggested that at least 44,000 and perhaps as many as 98,000 Americans die in hospitals each year as a result of medical errors. Much debate has occurred around the validity of those numbers (Leape, 2000). Whereas there is disagreement regarding the frequency of medical errors in health care, most people agree that system changes need to occur to improve the quality and safety of care (Institute of Medicine Committee on Quality of Health Care in America, 2001).

Different HFE approaches to patient safety emphasize the characteristics of the system (or structure) in which care processes occur and which lead to patient outcomes (Bogner, 2004, 1994; Moray, 1994). Therefore, the discipline of HFE has an important role to play in helping in the human-centered design of systems and processes in order to achieve both positive individual and organizational outcomes as well as improved patient outcomes (improved quality and safety of care) (Sainfort et al., 2001a).

### 5.2   Human Errors in Health Care

In the HFE literature, numerous models of and approaches to human error have been developed to understand the mechanisms leading to accidents and injuries. One of the most prevalent human error models was defined by Rasmussen (1983) and Reason (1990). This model defines two types of human error: (1) slips and lapses and (2) mistakes. In turn, mistakes can be categorized as resulting from either rule-based behavior or knowledge-based behavior. This taxonomy of human error has been successfully applied to analyze and evaluate accidents in a range of domains, including the nuclear industry (Moray, 1997; Rasmussen, 1982), aviation (Helmreich and Merritt, 1998), and more recently health care (Reason, 2000; Sexton et al., 2000).

Another important distinction brought up by the human error literature is that of active and latent errors (Reason, 1990). Active errors have effects that are felt or seen immediately and are associated with the performance of the "front-line operators," such as nurses. Latent errors are more likely to be related to organizational and management factors that are removed in both time and space from the front-line operations. The distinction between active and latent failures, or between the "sharp end" and the "blunt end" (Cook and Woods, 1994), has led to the recognition of the importance of organizational, management, and procedural factors in errors and accidents. This had led to the development of a number of models describing the "chain of events" that can lead to an accident or an adverse outcome. For instance, Vincent and colleagues (1998) proposed an organizational accident model that identifies the following chain of events: latent failures (i.e. management decision, organizational processes) influence conditions of work (i.e. workload, supervision, communication, equipment, knowledge/ability), which in turn can lead to unsafe acts or active failures (i.e., omissions, action slips/failures, cognitive failures, violations) that can lead to accidents or adverse outcomes if the barriers or defense mechanisms are insufficient.

Over the last decades, error recovery has been a focus of research in domains other than health care, including aviation, traffic control, process industry, and human computer interaction. In health care, error recovery mechanisms are important in a range of processes (e.g., medication) in order to improve patient safety and quality of care (Kanse et al., 2006). There are three phases involved in error recovery: (1) the detection of the failures or at least the immediate resulting deviation or problem, (2) followed by explanation of the problem and its causes, and (3) countermeasures aimed at returning to the normal situation or at least limiting the consequences, including recurrences, or even entirely skipping one or both of these last two phases, Figure 3 shows a graphical representation of the error recovery process (Kanse et al., 2006).

Numerous work system factors influence error recovery: person-related factors such as experience and knowledge, technical factors such as the design of the workplace, equipment and interfaces, and organizational factors such as culture, work design, and procedures and
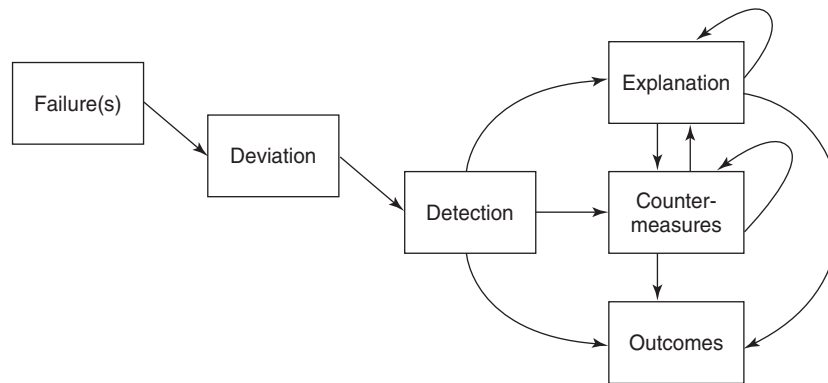
**Figure 3** Model of error recovery process (Kanse et al., 2006).

management priorities (Kanse, et al., 2006). Improving these factors may contribute to (complete or partial) recovery once an error or failure has occurred, thus preventing or reducing the negative consequences of that error or failure (Van Der Schaaf and Kanse, 2000). Human factors and ergonomic principles can improve work system factors that influence error recovery through better understanding human cognitive and physical abilities and training to improve person-related factors; technology design, implementation and evaluation, cognition, and situation awareness to improve technical factors; and job design and task analysis to improve organizational factors.

In health care, error recovery has been studied in domains such as cardiac surgery (Carthey et al., 2001), hospital pharmacy (Jonathan et al., 2002; Kanse et al., 2006), critical-care units (Dykes et al., 2010), and surgery (Catchpole et al., 2008).

### 5.3 Error Reporting

Medical errors are a leading cause of death. The IOM report (Cohen, 2000) estimated that more than 1 million preventable errors occur annually in the United States and about 44,000–98,000 result in death. In addition, medical errors lead to loss of time, resources, and credibility and cost of delays and legal action (Holden and Karsh, 2007). Medical error reporting has been encouraged and implemented in order to enhance patient safety (Anderson et al. 2009; Cohen, 2000; Holden and Karsh, 2007; Karsh et al., 2006). The World Alliance for Patient Safety program of the World Health Organization (WHO) (2005) has published draft guidelines for adverse-event reporting and learning systems that can facilitate the improvement or development of reporting systems for patient safety.

Error reporting may enhance patient safety by providing information to guide the development of error prevention strategies and recommendations for work system and process redesign. Other benefits of error reporting systems include (1) opportunities to educate and train health care providers, management, patients, and family members about the importance of patient safety; (2) mechanism to provide manufacturer

feedback about issues with their products, devices, and technologies; and (3) demonstrated commitment to patient safety that can increase public assurance and awareness (Karsh et al., 2006). However, error reporting systems are not implemented widely in U.S. health care organizations (Farley et al., 2008). Even when a reporting system is implemented, the amount of underreporting to national monitoring organizations is estimated to range from 50 to 96% of discovered medical errors annually (Barach and Small, 2000).

Many researchers have examined barriers and concerns related to error reporting (Brady et al., 2009; Evans et al., 2006; Firth-Cozens, 2002; Holden and Karsh, 2007; Lawton and Parker, 2002; Uribe et al., 2002). A review of the literature categorizes these barriers into the following: (1) busyness and fatigue; (2) difficult reporting methods and lack of knowledge about the existence of the reporting system; (3) adverse consequences of reporting; and (4) lack of perceived system usefulness (Holden and Karsh, 2007). Several factors determine the success and usefulness of an event reporting system: (1) the culture of an organization; (2) the provision of standardized methodologies; (3) classification systems and tools for analysis; and (4) and the feedback given to staff (Kaplan and Rabin Fastman, 2003).

Health care professionals differ in their attitudes to reporting medication errors. Recent studies found that physicians are unlikely to report less serious medication errors; nurses and pharmacists are likely to report all types of errors, including less serious as well as serious medication errors, despite their fears of receiving disciplinary action (Sarvadikar et al., 2010).

Human factors play a vital role in enhancing error reporting systems (Johnson, 2007). Many studies used human factor theories and principles in designing, implementing, evaluating, and assessing error reporting systems such as usability evaluation, user-centered design, change management, workload assessment, and decision-making strategies. For instance, Karsh et al. (2006) used sociotechnical theory with focus on the concept of end-user design to explore barriers and facilitators for the design of a statewide medical error reporting system. Kaplan and colleagues (Battles et al., 1998; Callum et al., 2001) designed and

**Table 3 HFE Issues in Health Care Addressed in Other Chapters**

| HFE Issues in Health Care | Chapters of the Handbook |
| --- | --- |
| Musculoskeletal disorders of health care workers | See Chapters 27 and 28 on low back disorders and upper extremity and shoulder disorders. |
| Layout of health care space | See Chapter 21 on workplace design |
| Teamwork in health care | See Chapter 15 on job and team design |
| Job and organizational design and working conditions in health care | See Chapter 15 on job and team design and Chapter 18 on human factors in organizational design and management |
| More on . . . human error | See Chapter 26 on human error |
| More on . . . usability | See Chapter 46 on usability testing |
| More on . . . cognitive ergonomics | See Chapter 5 on information processing and Chapter 7 on decision making |

implemented a near-miss event reporting system for transfusion medicine aimed at improving transfusion safety. Johnson (2007) reviews the human factors of health care reporting systems by examining recent initiatives from the IOM and range of national patient safety agencies. He also discusses the problem of underreporting and introduces different architectures for error reporting systems.

## 6 FUTURE NEEDS FOR HFE IN HEALTH CARE

The discipline of HFE has much to offer in order to improve the performance, quality, and safety of health care systems. Given the people-intensive, people-centered, people-driven characteristic of health care, HFE can provide the models, concepts, and methods necessary to consider the people component of health care systems.

Some of the HFE models, concepts, and methods need to be adapted to the characteristics of health care (see Section 1). For instance, a key principle of HFE is user participation and involvement. Two characteristics of health care contribute to the difficulty of implementing this principle. First is the definition of the user (see Section 2). Second, health care is a very dynamic changing environment with much time pressure. This makes the application of participatory approaches difficult when the "users" have little time to spare and spend on those HFE activities (see Section 2.3). More work needs to be done to pursue and expand the effort of considering the unique characteristics of health care in HFE work. Hignett (2003) also argues for the discipline of HFE to develop "more context-sensitive methodology" for health care.

Another important issue in HFE in medicine is the necessity of combining HFE technical knowledge with health care knowledge. For instance, in observing medical work such as anesthesia processes, the observer needs to have knowledge of anesthesia in order to meaningfully interpret the activities (Norros and Klemola, 1999). On the other hand, Weinger et al. (1994) developed a highly structured task analysis that could be used by non–medically trained observers in the observation of anesthesiologist's work. Carayon and colleagues (2004b, 2004c) discussed criteria to consider when combining HFE and health care knowledge in observations

of clinical work. The impact of HFE on medicine will be strengthened by encouraging "encounters" between HFE and medicine knowledge and the development of collaborations between HFE and health care subject matter experts.

## 7 CONCLUSION

In this chapter, we have reviewed a number of HFE issues in health care; but a number of important HFE issues were not reviewed. For instance, the issue of working conditions and workload in health care has not been addressed in this chapter, but much has been written on this topic, in particular regarding nursing (Carayon et al., 2003; Institute of Medicine Committee on the Work Environment for Nurses and Patient Safety, 2004). Table 3 lists other important HFE issues in health care that are addressed by other chapters of this handbook.

## REFERENCES

Aaronson, J. W., Murphy-Cullen, C. L., Chop, W. M., and Frey, R. D. (2001). Electronic Medical Records: The Family Practice Resident Perspective. *Family Medicine*, Vol. 33, No. 2, pp. 128–132.

Aiken, L. H., et al. (2001), "Nurses' Reports on Hospital Care in Five Countries," *Health Affairs*, Vol. 20, No. 3, pp. 43–53.

Alvarado, C. J., Wood, K., and Carayon, P. (2008), "Pre-Implementation Technology Assessment of Ultrasonic Placement of Central Venous Catheter Insertion," in *Human Factors in Organizational Design and Management*, Vol. IX, L. Sznelwar, F. Mascia, and U. Montedo, Eds., IEA Press, Santa Monica, CA, pp. 81–86.

Amalberti, R., and Hourlier, S. (2007), "Human Error Reduction Strategies in Health Care," in *Handbook of Human Factors and Ergonomics in Healthcare and Patient Safety*, P. Carayon, Ed., Lawrence Erbaum Associates, Mahwah, NJ, pp. 561–577.

American Telemedicine Association (2011), "Telemedicine Defined," available: http://www.americantelemed.org/i4a/pages/index.cfm?pageid=3333, accessed November 4, 2011.

Anderson, B., Stumpf, P. G., and Schulkin, J. (2009), "Medical Error Reporting, Patient Safety, and the Physician," *Journal of Patient Safety*, Vol. 5, No. 3, pp. 176–179.

Baker, G. R., et al. (2004), "The Canadian Adverse Events Study: The Incidence of Adverse Events among Hospital Patients in Canada," *Journal of the Canadian Medical Association*, Vol. 170, No. 11, pp. 1678–1686.

Barach, P., and Small, S. D. (2000), "Reporting and Preventing Medical Mishaps: Lessons from Non-Medical Near Miss Reporting Systems," *British Medical Journal*, Vol. 320, No. 7237, pp. 759–763.

Bates, D. W., and Gawande, A. A. (2003), "Improving Safety with Information Technology," *The New England Journal of Medicine*, Vol. 348, No. 25, pp. 2526–2534.

Battles, J. B., and Keyes, M. A. (2002), "Technology and Patient Safety: A Two-Edged Sword," *Biomedical Instrumentation and Technology*, Vol. 36, No. 2, pp. 84–88.

Battles, J. B., Kaplan, H. S., Van Der Schaaf, T. W., and Shea, C. E. (1998), "The Attributes of Medical Event-Reporting Systems: Experience with a Prototype Medical Event-Reporting System for Transfusion Medicine," *Archives of Pathology and Laboratory Medicine*, Vol. 122, No. 3, pp. 231–238.

Bergeson, S. C., and Dean, J. D. (2006), "A Systems Approach to Patient-Centered Care," *Journal of the American Medical Association*, Vol. 296, No. 23, pp. 2848–2851.

Bevan, N., Kirakowski, J., and Maissel, J. (1991), "What Is Usability?" in *Human Aspects in Computing: Design and Use*, H. J. Bullinger, Ed., Elsevier, Amsterdam, The Netherlands, pp. 651–655.

Blegen, M. A. (1993), "Nurses' Job Satisfaction: A Meta-Analysis of Related Variables," *Nursing Research*, Vol. 42, No. 1, pp. 36–41.

Blumenthal, D., and Epstein, A. M. (1996), "Part 6: The Role of Physicians in the Future of Quality Management," *The New England Journal of Medicine*, Vol. 335, No. 17, pp. 1328–1331.

Bogner, M. S. (2004), "Understanding Human Error," in *Misadventures in Health Care—Inside Stories*, M. S. Bogner, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1–12.

Bogner, M. S., Ed. (1994), *Human Error in Medicine*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Bohr, P. C., Evanoff, B. A., and Wolf, L. (1997), "Implementing Participatory Ergonomics Teams among Health Care Workers," *American Journal of Industrial Medicine*, Vol. 32, No. 3, pp. 190–196.

Bracco, D., et al. (2000), "Human Errors in a Multidisciplinary Intensive Care Unit: A 1-Year Prospective Study," *Intensive Care Medicine*, Vol. 27, No. 1, pp. 137–145.

Brady, A. M., Malone, A. M., and Fleming, S. (2009), "A Literature Review of the Individual and Systems Factors That Contribute to Medication Errors in Nursing Practice," *Journal of Nursing Management*, Vol. 17, No. 6, pp. 679–697.

Brunsson, N., and Jacobsson, B. (2000), *A World of Standards*, Oxford University Press, Oxford.

Bureau of Labor Statistics, U.S. Department of Labor (2010), *Career Guide To Industries, 2010--2011 Edition, Healthcare*, available: http://www.bls.gov/oco/cg/cgs035.htm, accessed July 1, 2010.

Callejas, S., Barry, A., Demertsidis, E., Jorgenson, D., and Becker, L. B. (2004), "Human Factors Impact Successful Lay Person Automated External Defibrillator Use during Simulated Cardiac Arrest," *Critical Care Medicine*, Vol. 32, No. 9, Suppl., pp. S406–S413.

Callum, J. L., et al. (2001), "Reporting of Near-Miss Events for Transfusion Medicine: Improving Transfusion Safety," *Transfusion*, Vol. 41, No. 10, pp. 1204–1211.

Cao, C. G. L., and Taylor, H. (2004), "Effects of New Technology on the Operating Room Team," in *Work with Computing Systems 2004*, H. M. Khalid, M. G. Helander and A. W. Yeo, Eds., Damai Sciences, Kuala Lumpur, Malaysia.

Carayon, P. (2006), "Human Factors of Complex Sociotechnical Systems," *Applied Ergonomics*, Vol. 37, pp. 525–535.

Carayon, P., and Haims, M. C. (2001), "Information and Communication Technology and Work Organization: Achieving a Balanced System," in *Humans on the Net-Information and Communication Technology (ICT), Work Organization and Human Beings*, G. Bradley, Ed., Prevent, Stockholm, Sweden, pp. 119–138.

Carayon, P., and Karsh, B.-T. (2000), "Sociotechnical Issues in the Implementation of Imaging Technology," *Behaviour and Information Technology*, Vol. 19, No. 4, pp. 247–262.

Carayon, P., and Smith, M. J. (1993), "The Balance Theory of Job Design and Stress as a Model for the Management of Technological Change," paper presented at the Fourth International Congress of Industrial Engineering, Marseille, France.

Carayon, P., and Smith, M. J. (2000), "Work Organization and Ergonomics," *Applied Ergonomics*, Vol. 31, pp. 649–662.

Carayon, P., and Smith, P. D. (2001), "Evaluating the Human and Organizational Aspects of Information Technology Implementation in a Small Clinic," in *Systems, Social and Internationalization Design Aspects of Human-Computer Interaction*, M. J. Smith and G. Salvendy, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 903–907.

Carayon, P., Alvarado, C., and Hundt, A. S. (2003), *Reducing Workload and Increasing Patient Safety through Work and Workspace Design*, Institute of Medicine, Washington, DC.

Carayon, P., Schultz, K., and Hundt, A. S. (2004a), "Righting Wrong Site Surgery," *Joint Commission Journal on Quality and Safety*, Vol. 30, No. 7, pp. 405–410.

Carayon, P., et al. (2004b), "Observing Nurse Interaction with Medication Administration Technologies, " in *Advances in Patient Safety: From Research to Implementation*, K. Henriksen, J. B. Battles, E. Marks, and D. I. Lewin, Eds. Agency for Healthcare Research and Quality, Washington, DC.

Carayon, P., et al. (2004c), "Assessing Nurse Interaction with Medication Administration Technologies: The Development of Observation Methodologies," in *Work with Computing Systems 2004*, H. M. Khalid, M. G. Helander, and A. W. Yeo, Eds., Damai Sciences, Kuala Lumpur, Malaysia, pp. 319–324.

Carayon, P., Alvarado, C. J., Hundt, A. S., Springman, S., and Ayoub, P. (2006a), "Patient Safety in Outpatient Surgery: The Viewpoint of the Healthcare Providers," *Ergonomics*, Vol. 49, Nos. 5/6, pp. 470–485.

Carayon, P., et al. (2006b), "Work System Design for Patient Safety: The SEIPS Model," *Quality and Safety in Health Care*, Vol. 15, Suppl. I, pp. i50–i58.

Carayon, P., et al. (2007), "Evaluation of Nurse Interaction with Bar Code Medication Administration Technology in the Work Environment," *Journal of Patient Safety*, Vol. 3, No. 1, pp. 34–42.

Carayon, P., Wetterneck, T. B., Hundt, A. S., Rough, S., and Schroeder, M. (2008), "Continuous Technology Implementation in Health Care: The Case of Advanced IV Infusion Pump Technology," in *Corporate Sustainability as a Challenge for Comprehensive Management*, K. Zink, Ed., Springer. New York, pp. 139–151.

Carayon, P., Faye, H., Hundt, A. S., Karsh, B.-T., and Wetterneck, T. B. (2009a), "Patient Safety and Proactive Risk Assessment," in Ed., *Handbook of Healthcare Delivery Systems*, Y. Yuehwern, Taylor & Francis, New York.

Carayon, P., Smith, P. D., Hundt, A. S., Kuruchittham, V., and Li, Q. (2009b), "Implementation of an Electronic Health Records System in a Small Clinic: The Viewpoint of Clinic Staff," *Behaviour and Information Technology*, Vol. 28, No. 1, pp. 5–20.

Carthey, J., de Leval, M. R., and Reason, J. T. (2001), "The Human Factor in Cardiac Surgery: Errors and Near Misses in a High Technology Medical Domain," *The Annals of Thoracic Surgery*, Vol. 72, No. 1, pp. 300–305.

Catchpole, K., Mishra, A., Handa, A., and McCulloch, P. (2008), "Teamwork and Error in the Operating Room: Analysis of Skills and Roles," *Annals of Surgery*, Vol. 247, No. 4, pp. 699–706.

Chassin, M. R. (1998), "Is Health Care Ready for Six Sigma Quality?" *Milbank Quarterly*, Vol. 76, No. 4, pp. 565–591.

Cohen, M. R. (2000), "Why Error Reporting Systems Should Be Voluntary," *British Medical Journal*, Vol. 320, No. 7237, pp. 728–729.

Coleman, E. A. (2003), "Falling through the Cracks: Challenges and Opportunities for Improving Transitional Care for Persons with Continuous Complex Care Needs," *Journal of the American Geriatrics Society*, Vol. 51, No. 4, pp. 549–555.

Coleman, E. A., and Berenson, R. A. (2004), "Lost in Transition: Challenges and Opportunities for Improving the Quality of Transitional Care," *Annals of Internal Medicine*, Vol. 141, No. 7, pp. 533–536.

Cook, R. I. (2002), "Safety Technology: Solutions or Experiments?" *Nursing Economic*, Vol. 20, No. 2, pp. 80–82.

Cook, R. I. (2004), "Observational and Ethnographic Studies: Insights on Medical Error and Patient Safety," paper presented at the 6th Annual NPSF Patient Safety Congress, Boston, MA.

Cook, R. I., and Woods, D. D. (1994), "Operating at the Sharp End: The Complexity of Human Error," in *Human Error in Medicine*, M. S. Bogner, Ed., Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 255–310.

Cooper, J. B. (1989), "Do Short Breaks Increase or Decrease Anesthetic Risk?" *Journal of Clinical Anesthesiology*, Vol. 1, No. 3, pp. 228–231.

Coyle-Shapiro, J. A.-M. (1999), "Employee Participation and Assessment of an Organizational Change Intervention," *Journal of Applied Behavioral Science*, Vol. 35, No. 4, pp. 439–456.

Davis, F. D. (1993), "User Acceptance of Information Technology: System Characteristics, User Perceptions and Behavioral Impacts," *International Journal of Man-Machine Studies*, Vol. 38, No. 3, pp. 475–487.

Deming, W. E. (1986), *Out of the Crisis*, Massachusets Institute of Technology, Cambridge, MA.

Donabedian, A. (1988), "The Quality Of Care: How Can It Be Assessed?" *Journal of the American Medical Association*, Vol. 260, No. 12, pp. 1743–1748.

Donaldson, L. J., Hayes, J. H., Barton, A. G., Howel, D., and Hawthorne, M. (1999), "Impact of Clinical Practice Guidelines on Clinicians' Behaviour: Tonsillectomy in Children," *Journal of Otolaryngology*, Vol. 28, No. 1, pp. 24–30.

Dumas, J., and Redish, J. C. (1993), *A Practical Guide to Usability Testing*, Ablex Publishing, Norwood, NJ.

Dykes, P. C., Rothschild, J. M., and Hurley, A. C. (2010), "Medical Errors Recovered by Critical Care Nurses," *Journal of Nursing Administration*, Vol. 40, No. 5, pp. 241–246.

Eason, K. (1982), "The Process of Introducing Information Technology," *Behaviour and Information Technology*, Vol. 1, No. 2, pp. 197–213.

Eason, K. (1988), *Information Technology and Organizational Change*, Taylor & Francis, London.

Effken, J. A., Kim, M.-G., and Shaw, R. E. (1997), "Making the Constraints Visible: Testing the Ecological Approach to Interface Design," *Ergonomics*, Vol. 40, No. 1, pp. 1–27.

Evanoff, V. A., Bohr, P. C., and Wolf, L. (1999), "Effects of a Participatory Ergonomics Team among Hospital Orderlies," *American Journal of Industrial Medicine*, Vol. 35, pp. 358–365.

Evans, R. S., et al. (1998), "A Computer-Assisted Management Program for Antibiotics and Other Antiinfective Agents," *New England Journal of Medicine*, Vol. 338, No. 4, pp. 232–238.

Evans, S. M., et al. (2006), "Attitudes and Barriers to Incident Reporting: A Collaborative Hospital Study," *Quality and Safety in Health Care*, Vol. 15, No. 1, pp. 39–43.

Farley, D. O., et al. (2008), "Adverse-Event-Reporting Practices by US Hospitals: Results of a National Survey," *Quality and Safety in Health Care*, Vol. 17, No. 6, pp. 416–423.

Feldman, P. H. (1993), "Work Life Improvements for the Home Aide Work Force: Impact and Feasibility," *Gerontologist*, Vol. 33, No. 1, pp. 47–54.

Field, M. J., and Lohr, K. N. (1900), *Clinical Practice Guidelines: Directions for a New Program*, National Academy Press, Washington, DC.

Firth-Cozens, J. (2002), "Barriers to Incident Reporting," *Quality and Safety in Health Care*, Vol. 11, No. 1, p. 7.

Frese, M., et al. (1988), "The Effects of an Active Development of the Mental Model in the Training Process: Experimental Results in a Word Processing System," *Behaviour and Information Technology*, Vol. 7, No. 3, pp. 295–304.

Gardell, B. (1977), "Autonomy and Participation at Work," *Human Relations*, Vol. 30, pp. 515–533.

Gattiker, U. E. (1992), "Computer Skills Acquisition: A Review and Future Directions for Research," *Journal of Management*, Vol. 18, No. 3, pp. 547–574.

Gray-Toft, P., and Anderson, J. G. (1981), "The Nursing Stress Scale: Development of an Instrument," *Journal of Behavioral Assessment*, Vol. 3, No. 1, pp. 11–23.

Grol, R., Dalhuijsen, J., Thomas, S., Veld, C., Rutten, G., and Mokkink, H. (1998), "Attributes of Clinical Guidelines That Influence Use of Guidelines in General Practice: Observational Study," *British Medical Journal*, Vol. 317, No. 7162, pp. 858–861.

Guerlain, S., Shin, T., Guo, H., Adams, R., and Calland, J. F. (2002), "A Team Performance Data Collection and Analysis System," paper presented at the Human Factors and Ergonomics Society 46th Annual Meeting, Baltimore, MD.

Guimond, M. E., Sole, M. L., and Salas, E. (2009), "Team-STEPPS," *American Journal of Nursing*, Vol. 109, No. 11, pp. 66–68.

Gurses, A. P., and Carayon, P. (2007), "Performance Obstacles of Intensive Care Nurses," *Nursing Research*, Vol. 56, No. 3, pp. 185–194.

Gurses, A. P., and Carayon, P. (2009), "A Qualitative Study of Performance Obstacles and Facilitators among ICU Nurses," *Applied Ergonomics*, Vol. 40, No. 3, pp. 509–518.

Gurses, A. P., et al. (2008), "Systems Ambiguity and Guideline Compliance: A Qualitative Study of How Intensive Care Units Follow Evidence-Based Guidelines to Reduce Healthcare-Associated Infections," *Quality and Safety in Health Care*, Vol. 17, No. 5, pp. 351–359.

Gurses, A. P., Carayon, P., and Wall, M. (2009), "Impact of Performance Obstacles on Intensive Care Nurses Workload, Perceived Quality and Safety of Care, and Quality of Working Life," *Health Services Research*, Vol. 44, pp. 422–443.

Helmreich, R. L., and Merritt, A. C. (1998), *Culture at Work in Aviation and Medicine*, Ashgate, Aldershot, England.

Helmreich, R. L., and Schaefer, H. (1994), "Team Performance in the Operating Room," in *Human Error in Medicine*, M. S. Bogner, Ed., Lawrence Erlbaum Associates, Mahway, NJ, pp. 225–253.

Henriksen, K., Joseph, A., and Zayas-Caban, T. (2009), "The Human Factors of Home Health Care: A Conceptual Model for Examining Safety and Quality Concerns," *Journal of Patient Safety*, Vol. 5, No. 4, pp. 229–236.

Hignett, S. (2003), "Hospital Ergonomics: A Qualitative Study to Explore the Organizational and Cultural Factors," *Ergonomics*, Vol. 46, No. 9, pp. 882–903.

Holden, R. J., and Karsh, B.-T. (2007), "A Review of Medical Error Reporting System Design Considerations and a Proposed Cross-Level Systems Research Framework," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 49, No. 2, pp. 257–276.

Hoonakker, P., Cartmill, R., Carayon, P., and Walker, J. M. (2010), "Development of the Survey to Evaluate CPOE/EHR Implementation in ICUs," Technical Report No. 2010-01, Center for Quality and Productivity Improvement, University of Wisconsin—Madison, Madison, WI.

Institute of Medicine (IOM) (2006), *Preventing Medication Errors*, National Academy Press, Washington, DC.

Institute of Medicine Committee on Data Standards for Patient Safety (2004), *Patient Safety—Achieving a New Standard for Care*, National Academy Press, Washington, DC.

Institute of Medicine Committee on Quality of Health Care in America (2001), *Crossing the Quality Chasm: A New Health System for the 21st Century*, National Academy Press, Washington, DC.

Institute of Medicine Committee on the Work Environment for Nurses and Patient Safety (2004), *Keeping Patients Safe: Transforming the Work Environment of Nurses*, National Academy Press, Washington, DC.

Issenberg, S. B., et al. (1999), "Simulation Technology for Health Care Professional Skills Training and Assessment," *Journal of the American Medical Association*, Vol. 282, No. 9, pp. 861–866.

Johnson, C. (2007), "Human Factors of Health Care Reporting Systems," in *Handbook of Human Factors and Ergonomics in Healthcare and Patient Safety*, in P. Carayon, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 525–560.

Jonathan, A., Jane, C., Judith, C., Matthew, P., and Suzette, W. (2002), "Medication Errors: Causes, Prevention and Reduction," *British Journal of Haematology*, Vol. 116, No. 2, pp. 255–265.

Kanse, L., Van Der Schaaf, T. W., Vrijland, N. D., and Van Mierlo, H. (2006), "Error Recovery in a Hospital Pharmacy," *Ergonomics*, Vol. 49, No. 5, pp. 503–516.

Kaplan, H. S., and Rabin Fastman, B. (2003), "Organization of Event Reporting Data for Sense Making and System Improvement," *Quality and Safety in Health Care*, Vol. 12, Suppl. 2, pp. ii68–ii72.

Karsh, B.-T. (1997), "An Examination of Employee Participation during New Technology Implementation," *Human Factors and Ergonomics Society Annual Meeting Proceedings*, Vol. 41, pp. 767–771.

Karsh, B.-T. (2004), "Beyond Usability: Designing Effective Technology Implementation Systems to Promote Patient Safety," *Quality and Safety in Health Care*, Vol. 13, No. 5, pp. 388–394.

Karsh, B.-T., and Brown, R. (2010), "Macroergonomics and Patient Safety: The Impact of Levels on Theory, Measurement, Analysis and Intervention in Patient Safety Research," *Applied Ergonomics*, Vol. 41, No. 5, pp. 674–681.

Karsh, B.-T., and Holden, R. J. (2007), "New Technology Implementation," in *Handbook of Human Factors and Ergonomics in Healthcare and Patient Safety*, in P. Carayon, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 393–410.

Karsh, B.-T., Escoto, K. H., Beasley, J. W., and Holden, R. J. (2006), "Toward a Theoretical Approach to Medical Error Reporting System Research and Design," *Applied Ergonomics*, Vol. 37, No. 3, pp. 283–295.

Kaushal, R., and Bates, D. W. (2001), "Computerized Physician Order Entry (CPOE) with Clinical Decision Support Systems (CDSSs)," in *Making Health Care Safer: A Critical Analysis of Patient Safety Practices*, K. G. Shojania, B. W. Duncan, K. M. McDonald, and R. M. Wachter, Eds., Vol. Evidence Report/Technology Assessment, Agency for Healthcare Research and Quality, Washington, DC, pp. 59–69.

King, H., et al. (2008), "TeamSTEPPS™: Team Strategies and Tools to Enhance Performance and Patient Safety," in *Advances in Patient Safety: New Directions and Alternative Approaches*, K. Henriksen, J. B. Battles, M. A. Keyes, and M. L. Grady, Eds., Agency for Healthcare Research and Quality, Washington, DC.

Knaus, W. A., Draper, E. A., Wagner, D. P., and Zimmerman, J. E. (1986), "An Evaluation of Outcome from Intensive Care in Major Medical Centers," *Annals of Internal Medicine*, Vol. 104, pp. 410–418.

Kohn, L. T., Corrigan, J. M., and Donaldson, M. S., Eds. (1999), *To Err Is Human: Building a Safer Health System*, National Academy Press, Washington, DC.

Koppel, R., and Kreda, D. A. (2010), "Healthcare IT Usability and Suitability for Clinical Needs: Challenges of Design, Workflow, and Contractual Relations," *Studies in Health Technology and Informatics*, Vol. 157, pp. 7–14.

Korunka, C., and Carayon, P. (1999), "Continuous Implementations of Information Technology: The Development of an Interview Guide and a Cross-National Comparison of Austrian and American Organizations," *The International Journal of Human Factors in Manufacturing*, Vol. 9, No. 2, pp. 165–183.

Korunka, C., Weiss, A., and Karetta, B. (1993), "Effects of New Technologies with Special Regard for the Implementation Process Per-Se," *Journal of Organizational Behavior*, Vol. 14, No. 4, pp. 331–348.

Korunka, C., Zauchner, S., and Weiss, A. (1997), "New Information Technologies, Job Profiles, and External Workload as Predictors of Subjectively Experienced Stress and Dissatisfaction at Work," *International Journal of Human-Computer Interaction*, Vol. 9, No. 4, pp. 407–424.

Lawler, E. E. I. (1986), *High Involvement Management: Participative Strategies for Improving Organizational Performance*, Jossey-Bass, San Francisco.

Lawler III, E. E. (1986), *High Involvement Management: Participative Strategies for Improving Organizational Performance*, Jossey-Bass, San Francisco.

Lawton, R., and Parker, D. (2002), "Barriers to Incident Reporting in a Healthcare System," *Quality and Safety in Health Care*, Vol. 11, No. 1, pp. 15–18.

Leape, L. L. (2000), "Institute of Medicine Medical Error Figures Are Not Exaggerated," *Journal of the American Medical Association*, Vol. 284, No. 1, pp. 95–97.

Lee, E. J., Haddock, M., Yu, S., Back, C., and Yassi, A. (2006), "Developing and Evaluating a Nursing Bag System for Home Care Nurses Using a Participatory Ergonomics Approach," *Home Healthcare Nurse*, Vol. 24, No. 9, pp. 591–597.

Li, Q., Douglas, S., Hundt, A. S., and Carayon, P. (2006), "A Heuristic Evaluation of a Computerized Provider Order Entry (CPOE) Technology," in *Proceedings of the IEA2006 Congress*, R. N. Pikaar, E. A. P. Koningsveld, and P. J. M. Settels, Eds., Elsevier, Maastricht, The Netherlands.

Lilly, C. M., and Thomas, E. J. (2010), "Tele-ICU: Experience to Date," *Journal of Intensive Care Medicine*, Vol. 25, No. 1, 1 pp. 6–22.

Lin, L., Vicente, K. J., and Doyle, D. J. (2001), "Patient Safety, Potential Adverse Drug Events, and Medical Device Design: A Human Factors Engineering Approach," *Journal of Biomedical Informatics*, Vol. 34, No. 4, pp. 274–284.

Lorenzi, N. M., Kouroubali, A., Detmer, D. E., and Bloomrosen, M. (2009), "How to Successfully Select and Implement Electronic Health Records (EHR) in Small Ambulatory Practice Settings," *BMC Medical Informatics and Decision Making*, Vol. 9, p. 15.

Lutz, B. J., and Bowers, B. J. (2000), "Patient-Centered Care: Understanding Its Interpretation and Implementation in Health Care," *Scholarly Inquiry for Nursing Practice*, Vol. 14, No. 2, pp. 165–183.

Mayhew, D. J. (1999), *The Usability Engineering Lifecycle*, Morgan Kaufmann, San Francisco, CA.

McDonald, C. J., Overhage, J. M., Tierney, W. M., Abernathy, G. R., and Dexter, P. R. (1996), "The Promise of Computerized Feedback Systems for Diabetes Care," *Annals of Internal Medicine*, Vol. 124, 1S-II, Suppl., pp. 170–174.

McGlynn, E. A., et al. (2003), "The Quality of Health Care Delivered to Adults in the United States," *The New England Journal of Medicine*, Vol. 348, No. 26, pp. 2635–2645.

McNeil, J. J., and Leeder, S. R. (1995), "How Safe Are Australian Hospitals?" *Medical Journal of Australia*, Vol. 163, No. 6, pp. 472–475.

Meister, D., and Enderwick, T. P. (2001), *Human Factors in System Design, Development, and Testing*, Lawrence Erlbaum Associates, Mahwah, NJ.

Mohrman, S. A., Cohen, S. G., and Mohrman, A. M. J. (1995), *Designing Team-Based Organizations: New Forms of Knowledge and Work*, Jossey Bass, San Francisco, CA.

Moray, N. (1994), "Error Reduction as a Systems Problem," in *Human Error in Medicine*, M. S. Bogner, Ed., Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 67–91.

Moray, N. (1997), "Human Factors in Process Control," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salveny, Ed., Wiley, New York, pp. 1944–1971.

National Patient Safety Agency (NPSA) (2004), "Improving Infusion Device Safety," No. NPSA Safer Practice Notice 01, available: http://www.nrls.npsa.nhs.uk/resources/?EntryId45=59788, accessed July, 22, 2010.

Nielsen, J. (1992), "The Usability Engineering Lifecycle," *IEEE Computer*, Vol. 25, No. 3, pp. 12–22.

Nielsen, J. (1993), *Usability Engineering*, Morgan Kaufmann, Amsterdam, The Netherlands.

Nishisaki, A., Keren, R., and Nadkarni, V. (2007), "Does Simulation Improve Patient Safety? Self-Efficacy, Competence, Operational Performance, and Patient Safety," *Anesthesiology Clinics*, Vol. 25, No. 2, pp. 225–236.

Norman, D. A. (1988), *The Psychology of Everyday Things*, Basic Books, New York.

Noro, K., and Imada, A. (1991), *Participatory Ergonomics*, Taylor and Francis, London.

Norros, L., and Klemola, U.-M. (1999), "Methodological Considerations in Analysing Anaesthetists' Habits of Action in Clinical Situations," *Ergonomics*, Vol. 42, No. 11, pp. 1521–1530.

Patterson, E. S., and Wears, R. L. (2010), "Patient Handoffs: Standardized and Reliable Measurement Tools Remain Elusive," *The Joint Commission Journal on Quality and Patient Safety*, Vol. 36, No. 2, pp. 52–61.

Patterson, E. S., Cook, R. I., and Render, M. L. (2002), "Improving Patient Safety by Identifying Side Effects from Introducing Bar Coding in Medication Administration," *Journal of the American Medical Informatics Association*, Vol. 9, pp. 540–553.

Patterson, E. S., Roth, E. M., Woods, D. D., Chow, R., and Gomes, J. O. (2004), "Handoff Strategies in Settings with High Consequences for Failure: Lessons for Health Care Operations," *International Journal for Quality in Health Care*, Vol. 16, No. 2, pp. 125–132.

Perrow, C. (1984), *Normal Accidents—Living with High-Risk Technologies*, Basic Books, New York.

Perry, S. (2004), "Transitions in Care: Studying Safety in Emergency Department Signovers," *Focus on Patient Safety*, Vol. 7, No. 2, pp. 1–3.

Randell, R. (2003), "User Customisation of Medical Devices: The Reality and the Possibilities," *Cognition, Technology and Work*, Vol. 5, pp. 163–170.

Rasmussen, J. (1982), "Human Errors. A Taxonomy for Describing Human Malfunction in Industrial Installations," *Journal of Occupational Accidents*, Vol. 4, pp. 311–333.

Rasmussen, J. (1983), "Skills, Rules, and Knowledge; Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 13, No. 3, pp. 257–266.

Rasmussen, J. (1997), "Risk Management in a Dynamic Society: A Modelling Problem," *Safety Science*, Vol. 27, Nos. 2/3, pp. 183–213.

Rasmussen, J. (2000), "Human Factors in a Dynamic Information Society: Where Are We Heading?" *Ergonomics*, Vol. 43, No. 7, pp. 869–879.

Ravden, S., and Johnson, G. (1989), *Evaluating Usability of Human-computer Interfaces—A Practical Method*, Wiley, New York.

Reason, J. (1990), *Human Error*, Cambridge University Press, Cambridge.

Reason, J. (2000), "Human Error: Models and Management," *British Medical Journal*, Vol. 320, No. 7237, pp. 768–770.

Reid, P. R., Compton, W. D., Grossman, J. H., and Fanjiang, G. (2005), *Building a Better Delivery System. A New Engineering/Health Care Partnership*, National Academy Press, Washington, DC.

Reiling, J. G., and Chernos, S. (2007), "Human Factors in Hospital Safety Design," in *Handbook of Human Factors and Ergonomics in Healthcare and Patient Safety*, P. Carayon, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 275–286.

Reiling, J. G., Knutzen, B. L., Wallen, T. K., McCullough, S., Miller, R. H., and Chernos, S. (2004), "Enhancing the Traditional Design Process: A Focus on Patient Safety," *The Joint Commission Journal on Quality Improvement*, Vol. 30, No. 3, pp. 115–124.

Reiser, S. J. (1978), *Medicine and the Reign of Technology*, Cambridge University Press, Cambridge.

Sainfort, F., Karsh, B.-T., Booske, B. C., and Smith, M. J. (2001a), "Applying Quality Improvement Principles to Achieve Healthy Work Organizations," *Journal on Quality Improvement*, Vol. 27, No. 9, pp. 469–483.

Sainfort, F., Taveira, A. D., Arora, N., and Smith, M. J. (2001b), "Teams and Team Management and Leadership," in *Handbook of Industrial Engineering*, G. Salvendy, Ed., Wiley,. New York, pp. 975–994.

Salas, E., Cooke, N. J., and Rosen, M. A. (2008), "On Teams, Teamwork, and Team Performance: Discoveries and Developments," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 50, No. 3, pp. 540–547.

Sarvadikar, A., Prescott, G., and Williams, D. (2010), "Attitudes to Reporting Medication Error Among Differing Healthcare Professionals," *European Journal of Clinical Pharmacology,* Vol. 66, No. 8, pp. 843–853.

Saucier, R. (2010), "Design Standardization in the Private Neonatal Intensive Care Unit Room," *Newborn and Infant Nursing Reviews*, Vol. 10, No. 2, pp. 92–96.

Schultz, K., Carayon, P., Hundt, A. S., and Springman, S. (2007), "Care Transitions in the Outpatient Surgery Preoperative Process: Facilitators and Obstacles to Information Flow and Their Consequences." *Cognition, Technology and Work*, Vol. 9, No. 4, pp. 219–231.

Scott, R. D. (2009), *The Direct Medical Costs of Healthcare-Associated Infections in US Hospitals and the Benefits of Prevention*, Centers for Disease Control and Prevention, Atlanta, GA.

Scroggs, J. (2008), "Improving Patient Safety Using Clinical Needs Assessments in IV Therapy," *British Journal of Nursing*, Vol. 17, No. 19, pp. S22–28.

Sexton, J. B., Thomas, E. J., and Helmreich, R. L. (2000), "Error, Stress, and Teamwork in Medicine and Aviation: Cross Sectional Surveys," *British Medical Journal*, Vol. 320, No. 7237, pp. 745–749.

Sheridan, T. B., and Thompson, J. M. (1994), "People versus Computers in Medicine," in *Human Error in Medicine*, M. S. Bogner, Ed., Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 141–158.

Shortell, S. M., et al. (1994), "The Performance of Intensive Care Units: Does Good Management Make a Difference?" *Medical Care*, Vol. 32, No. 5, pp. 508–525.

Simmons, B. L., Nelson, D. L., and Neal, L. J. (2001), "A Comparison of the Positive and Negative Work Attitudes of Home Health Care and Hospital Nurses," *Health Care Management Review*, Vol. 26, No. 3, pp. 63–74.

Smith, K. U., and Smith, M. F. (1966), *Cybernetic Principles of Learning and Educational Design*, Holt, Rhinehart, and Winston, New York.

Smith, M. J., and Carayon, P. (1995), "New Technology, Automation, and Work Organization: Stress Problems and Improved Technology Implementation Strategies," *International Journal of Human Factors in Manufacturing*, Vol. 5, No. 1, pp. 99–116.

Smith, M. J., and Carayon, P. (2000), "Balance Theory of Job Design," in *International Encyclopedia of Ergonomics and Human Factors*, W. Karwowski, Ed., Taylor & Francis, London, pp. 1181–1184.

Smith, M. J., and Carayon-Sainfort, P. (1989), "A Balance Theory of Job Design for Stress Reduction," *International Journal of Industrial Ergonomics*, Vol. 4, No. 1, pp. 67–79.

Staggers, N. (2003), "Human Factors: Imperative Concepts for Information Systems in Critical Care," *AACN Clinical Issues*, Vol. 14, No. 3, pp. 310–319.

Stone, R. I., and Wiener, J. M. (2001), *Who Will Care for Us? Addressing the Long-Term Workforce Crisis*, Urban Institute and the American Association of Homes and Services for the Aging, Washington, DC.

Sultz, H. A., and Young, K. M. (2001), *Health Care USA: Understanding Its Organization and Delivery*, Aspen, , Gaithersburg, MD.

Sundstrom, E., de Meuse, K. P., and Futrell, D. (1990), "Work Teams: Applications and Effectiveness," *American Psychologist*, Vol. 45, No. 2, pp. 120–133.

Syed, G., Sarwar, S., and Ian, R. (2008), "User Involvement in Healthcare Technology Development and Assessment: Structured Literature Review," *International Journal of Health Care Quality Assurance*, Vol. 19, No. 6, pp. 500–515.

Tannenbaum, S. I., Salas, E., and Cannon-Bowers, J. A. (1996), "Promoting Team Effectiveness," In *Handbook of Work Group Psychology*, M. A. West, Ed., Wiley, New York, pp. 503–530.

Thomas, E. J., Sexton, J. B., and Helmreich, R. L. (2003), "Discrepant Attitudes about Teamwork among Critical Care Nurses and Physicians," *Critical Care Medicine*, Vol. 31, No. 3, pp. 956–959.

Timmermans, S., and Berg, M. (2003), *The Gold Standard: The Challenge of Evidence-Based Medicine and Standardization in Health Care*, Temple University Press, Philadelphia, PA.

Topf, M. (2000), "Hospital Noise Pollution: An Environmental Stress Model to Guide Research and Clinical Interventions," *Journal of Advanced Nursing*, Vol. 31, No. 3, pp. 520–528.

Truffer, C. J., et al. (2010), "Health Spending Projections through 2019: The Recession's Impact Continues," *Health Affairs*, Vol. 29, No. 3, pp. 522–529.

Tunis, S. R., et al. (1994), "Internists' Attitudes about Clinical Practice Guidelines," *Annals of Internal Medicine*, Vol. 120, No. 11, pp. 956–963.

U.K. Department of Health (2002), *An Organisation with a Memory: Report of an Expert Group on Learning from Adverse Events in the NHS*, U.K. Department of Health, London.

Uribe, C. L., Schweikhart, S. B., Pathak, D. S., Dow, M., and Marsh, G. B. (2002), "Perceived Barriers to Medical-Error Reporting: An Exploratory Investigation," *Journal of Healthcare Management*, Vol. 47, No. 4, pp. 263–279.

Van Cott, H. (1994), "Human Errors: Their Causes and Reduction," in *Human Error in Medicine*, M. S. Bogner, Ed., Lawrence Erlbaum Associates. Hillsdale, NJ, pp. 53–65.

Van Der Schaaf, T. W., and Kanse, L. (Artist) (2000), "Errors and Error Recovery," in *Human Error and System Design and Management*, P. F. Elzer, R. H. Kluwe and B. Boussoffara, Eds., Springer Verlag, London. pp. 27–38.

Venkatesh, V., Speier, C., and Morris, M. G. (2002), "User Acceptance Enablers in Individual Decision Making about Technology: Toward an Integrated Model," *Decision Sciences*, Vol. 33, No. 2, pp. 297–316.

Vicente, K. J. (1999), *Cognitive Work Analysis*, Lawrence Erlbaum Associates, Mahwah, NJ.

Vincent, C. A. (2004), "Analysis of Clinical Incidents: A Window on the System Not a Search for Root Causes," *Quality and Safety in Health Care*, Vol. 13, pp. 242–243.

Vincent, C. A., Taylor-Adams, S., and Stanhope, N. (1998), "Framework for Analysing Risk and Safety in Clinical Medicine," *British Medical Journal*, Vol. 316, No. 7138, pp. 1154–1157.

Wald, H., and Shojania, K. (2001), "Prevention of Misidentifications," in *Making Health Care Safer: A Critical Analysis of Patient Safety Practices*, D. G. Shojania, B. W. Duncan, K. M. McDonald, and R. M. Wachter, Eds., AHRQ publication No. 01-E058, Agency for Healthcare Research and Quality, Washington, DC, pp. 491–503.

Walker, J. M., and Carayon, P. (2009), "From Tasks to Processes: The Case for Changing Health Information Technology to Improve Health Care," *Health Affairs*, Vol. 28, No. 2, pp. 467–477.

Ward, J., and Clarkson, J. (2007), "Human Factors Engineering and the Design of Medical Devices," in *Handbook of Human Factors and Ergonomics in Healthcare and Patient Safety*, P. Carayon, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 367–382.

Watson, M., Sanderson, P., and Russell, W. J. (2004), "Tailoring Reveals Information Requirements: The Case of Anaesthesia Alarms," *Interacting with Computers*, Vol. 16, No. 2, pp. 271–293.

Wears, R. L., Perry, S. J., Shapiro, M., Beach, C., Croskerry, P., and Behara, R. (2003), "Shift Changes among Emergency Physicians: Best of Times, Worst of Times," in *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting* Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1420–1423.

Weinger, M. B., Herndon, O. W., Zornow, M. H., Paulus, M. P., Gaba, D. M., and Dallen, L. T. (1994), "An Objective Methodology for Task Analysis and Workload Assessment in Anesthesia Providers," *Anesthesiology*, Vol. 80, No. 1, pp. 77–92.

Weinger, M. B., Slagle, J., Jain, S., and Ordonez, N. (2003), "Retrospective Data Collection and Analytical Techniques for Patient Safety Studies," *Journal of Biomedical Informatics*, Vol. 36, Nos. 1/2, pp. 106–119.

Wetterneck, T. B., et al. (2006), "Using Failure Mode and Effects Analysis to Plan Implementation of Smart Intravenous Pump Technology," *American Journal of Health-System Pharmacy*, Vol. 63, pp. 1528–1538.

Whiteside, J., Bennett, J., and Holtzblatt, K. (1988), "Usability Engineering: Our Experience and Evolution," in *Handbook of Human-Computer Interaction*, M. Helander, Ed., North-Holland, Amsterdam, The Netherlands, pp. 791–817.

Wickens, C. D., Lee, J. D., Liu, Y., and Becker, S. E. G. (2004), *An Introduction to Human Factors Engineering*, 2nd ed., Prentice Hall, Upper Saddle River, NJ.

Wiklund, M. E. (1993a), "How to Implement Usability Engineering," *Medical Device and Diagnostic Industry*, Vol. 15, No. 9, pp. 68–73.

Wiklund, M. E. (1993b), "Usability Tests of Medical Products as a Prelude to the Clinical Trial," *Medical Device and Diagnostic Industry*, Vol. 3, No. 7, pp. 68–73.

Wiklund, M. E. (1995), *Medical Device and Equipment Design: Usability Engineering and Ergonomics*, Interpharm, Buffalo Grove, IL.

Wilson, J. R. (1995), "Ergonomics and Participation," in *Evaluation of Human Work*, 2nd ed., J. R. Wilson and E. N. Corlett, Eds., Taylor & Francis. London, pp. 1071–1096.

Wilson, J. R., and Haines, H. M. (1997), "Participatory Ergonomics," in *Handbook of Human Factors and Ergonomics*, G. Salvendy, Ed., Wiley, New York, pp. 490–513.

Woodward, H. I., et al. (2010), "What Have We Learned about Interventions to Reduce Medical Errors?" *Annual Review of Public Health*, Vol. 31, pp. 479–497.

Woolf, S. H., Grol, R., Hutchinson, A., Eccles, M., and Grimshaw, J. (1999), "Clinical Guidelines: Potential Benefits, Limitations, and Harms of Clinical Guidelines," *British Medical Journal*, Vol. 318, No. 7182, pp. 527–530.

World Health Organization (WHO) (2005), *WHO Draft Guidelines for Adverse Event Reporting and Learning Systems*, WHO, Geneva.

Wunderlich, G. S., and Kohler, P. O., Eds. (2001), *Improving the Quality of Long-Term Care*, National Academy Press, Washington, DC.

Zayas-Caban, T., and Brennan, P. F. (2007), "Human Factors in Home Care," in *Handbook of Human Factors and Ergonomics in Healthcare and Patient Safety*, in P. Carayon, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 883–897.

# CHAPTER 58

# HUMAN FACTORS AND ERGONOMICS IN MOTOR VEHICLE TRANSPORTATION

**Paul A. Green**
**University of Michigan Transportation Research Institute (UMTRI)**
**Ann Arbor, Michigan**

## 1 CHAPTER ORGANIZATION AND PHILOSOPHY

This chapter provides background material, design guidelines, reference data, and equations that human factors engineers can use to design and evaluate motor vehicles to make them safe, useful, and easy to use. In addition, this chapter also identifies areas in which human factors research is needed.

Coverage of this chapter is intended to be global, though until now much of the research has been conducted in the United States, Western Europe, and Japan. Accordingly, information is lacking on China and India, important vehicle markets.

## 2 DRIVING CONTEXT (PEOPLE, VEHICLES, ROADS)

### 2.1 Who Are the Users, the Drivers?

In their seminal paper, Gould and Lewis (1985) describe three key principles to design useful and easy-to-use systems. The first is "early focus on users and tasks." Given that principle, the relevant questions are: Who drives? What do people drive? Where and when do they drive?

Legally, almost any adult can drive, subject to passing a test. Thus, the age distribution of potential drivers should closely match the distribution of adults in that country, though there are countries with gender restrictions for licensing, such as Saudi Arabia. There is little data on the distribution of driver age in aggregate for the world or by country, except for the United States [Figure 1; Federal Highway Administration (FHWA), 2009a]. Notice that most Americans still drive beyond age 65. To provide a specific example, 5% of all drivers in the United States are ages 65–69 and 94% of those in that age group are licensed drivers. Thus, older adults commonly drive and should be expected to use everything produced, even if they are not the intended market segment. Older drivers are likely to be the most challenged group for ingress/egress, reach, and telematics use.

In the United States, licensing requirements for passenger car drivers typically consist of a vision test, usually static visual acuity, a test of knowledge of driving rules and laws, and a brief driving test [Eby and Molnar, 2008; Leonard, 2002; International Council of Ophthalmology (ICOPH), 2002, 2006; http://www.md support.org/library/drivingrequirements.html, retrieved
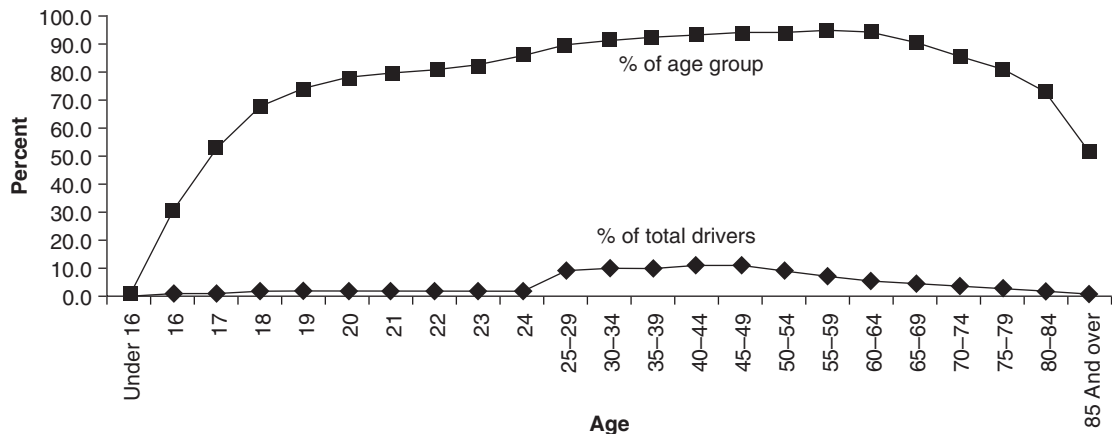
**Figure 1**   Age distribution of U.S. drivers (developed from FHWA, 2009a, Table DL-20).

September 9, 2010]. In other countries, licensing requirements vary quite widely, and for some, obtaining a license may require minimal skill, training, or knowledge. Corruption can also be an issue (Bertrand et al., 2008). Globally, licensing requirements are generally more stringent for drivers of buses, trucks (especially large trucks), and vehicles with trailers (http://en .wikipedia.org/wiki/European_driving_licence, retrieved September 9, 2010).

### 2.2   What Do People Drive?

How are vehicles classified? In the United States, for fuel economy comparisons, cars are grouped into size categories (e.g., minicompact, midsized) by the Environmental Protection Agency based on interior volume (http://www.fueleconomy.gov/feg/info.shtml#size classes). Europe has two schemes, one as part of its Euro Car Segment description (ranging from A to F, with F being the largest) and a second, similar five-category scheme as part of its Euro New Assessment Program (http://en.wikipedia.org/wiki/Vehicle_size_class). Thus, in Europe, a particular vehicle might be described as B class, which would be a subcompact. China and Japan have different schemes. Multiple schemes exist due to practical and historical national differences in disposable income for purchasing vehicles, the price of fuel, road width, and the availability of parking. Noteworthy, for example, is the Kei (Keijidosha) class of cars in Japan, which are less than 3.4 m long and have engines of less than 660 cc.

Trucks are classified by their gross weight and the number of axles. Large trucks are either straight trucks or tractor-trailers. Tractor-trailers vary in their cab design (conventional or cab over engine), have two to four axles for the cab, and vary in the length of the semitrailer. In the United States, tractor-trailers usually have a single trailer, commonly 40 ft long but varying anywhere from 28 to 53 ft. The largest tractor-trailer combinations are Australian road trains, where a single tractor may pull three or four trailers. In Europe and Japan, straight trucks are more common, because there is less space to maneuver a vehicle.

Finally, one needs to consider the wide range of mopeds, motorcycles, and various types of three-wheel vehicles, which are relatively more common outside North America, especially in Asia.

Which vehicles are most popular? According to the International Organization of Motor Vehicle Manufacturers (OICA) (http://www.oica.net/category/ production-statistics/), there were approximately 51.1 million cars, 7.8 million light commercial vehicles, 1.3 million heavy commercial vehicles, and 0.3 million heavy buses produced in 2009. Though cars predominate in number, heavy vehicles and buses tend to be much more expensive on average than cars and accumulate more mileage per year and over their lifetime.

As shown in Table 1, China is the largest market for motor vehicles, with sales in China increasing by 45% between 2008 and 2009. India, which also has a large

**Table 1  2009 Sales for New Cars by Country**

| Rank | Country | Volume | Rank | Country | Volume |
|---|---|---|---|---|---|
| 1 | China | 13,600,000 | 6 | France | 2,270,000 |
| 2 | United States | 10,400,000 | 7 | Italy | 2,160,000 |
| 3 | Japan | 4,600,000 | 8 | United Kingdom | 1,990,000 |
| 4 | Germany | 3,870,000 | 9 | Russia | 1,470,000 |
| 5 | Brazil | 3,140,000 | 10 | Canada | 1,460,000 |

*Source:* http://www.thetruthaboutcars.com/2010/01/car-sales-around-the-world-in-2009-mostly-down/

population, has not yet seen such rapid growth, but it is expected in the future.

The kinds of vehicles that are popular vary among regions of the world. In the United States, the best-selling vehicle for many years has been the Ford F-series pickup truck, with an estimated half million to be sold in 2010, extrapolating from published data (http://online .wsj.com/mdc/public/page/2_3022-autosales.html). The number 2 vehicle is the Chevrolet Silverado, another pickup truck, and vehicles 8 and 9 of the top-10 best sellers are two sport utility vehicles (SUVs) (Honda CR-V and Ford Escape). Pickup trucks, such as the Ford F series, are uncommon in other parts of the world.

Much smaller than the car and truck markets are the markets for construction equipment, agricultural equipment, forestry equipment, and mining equipment. To provide some perspective, in 2010 an estimated 169,300 tractors of all types will be sold in the United States and an estimated 10,300 self-propelled combines, extrapolating from published data (http://www.aem.org/ MarketInfo/Stats/Reports/AgTCR-US/). The design of these work vehicles (e.g., motor graders, asphalt pavers, tractor loader backhoes, feller bunchers) is quite specialized and beyond the scope of this chapter.

Finally, there is also a separate body of literature concerning the design of military vehicles. To find information on them, readers should use Google to search for publications from the U.S. Army Human Research and Engineering Directorate of the Army Research Laboratory in Aberdeen, Maryland.

## 2.3 On What Kinds of Roads Do People Drive (and How Are Roads and Traffic Described)?

Most driving, at least in the developed world, occurs on roads. Roads are classified hierarchically [Transportation Research Board (TRB), 2010]. At the lowest level are local roads, the roads that go to individual homes or businesses. Above them are collectors, into which local roads intersect. Above them are minor arterials and major arterials, each of which aggregates traffic from lower subclasses. Above them are limited-access highways, which have only a few entrances and exits, and are generally divided. Expressways and interstate highways in the United States, TransCanada roads in Canada, and motorways in other parts of the word are typically limited-access highways.

In the United States, local roads comprise about two-thirds of the rural and urban lane miles (not road miles) but carry only about 13% of the traffic (measured in vehicle miles traveled). In contrast, interstate highways, representing 3% of all lane miles, carry about 25% of the traffic. (See http://www.fhwa.dot.gov/ policyinformation/statistics/2008/hm260.cfm.)

Data on the number of lane miles and vehicle miles traveled by road type are difficult to obtain for other than the United States and may be inaccurate. The best source of data on miles of public roads, the International Road Federation (IRF, 2009) report, lists a number of countries (e.g., Australia, Brazil) as having zero miles of "motorways," which seems counterintuitive.

Roads are also classified by the entity responsible for their construction and maintenance. In the United States, major roads that cross state boundaries (interstate highways, U.S. routes) are the responsibility of the federal government. Below them are state or provincial roads, for which those governmental units are responsible. Finally, there are usually county and sometimes city or town roads for which those governmental units are responsible.

Driving is also affected by a road's physical characteristics—the number of lanes, lane width (typically 12 feet on U.S. Interstates), shoulder width and surface material, presence of barriers and medians (and their size), clear zones, curve radii, slopes, and so forth. Urban roads will have gutters and curbs. Many of these characteristics are described in detail in the *Highway Capacity Manual* (TRB, 2010) and the American Association of State and Highway Transportation Officials (AASHTO, 2004) "Green Book." Except for studies in which an author with knowledge of civil engineering is involved, driving studies often provide insufficient detail in describing test roads (e.g., a two-lane rural road).

Driving is also influenced by traffic and should be reported in a quantitative manner in all studies of driving. Traffic is generally described as the number of vehicles per lane per hour and the percentage of the traffic that is composed of trucks, though the percentage of the vehicles that are bicycles or motorcycles is also of interest. The amount of traffic is often categorized using the level of service (LOS), an indicator of delay. LOS is analogous to grades in school ranging from A to F, with A being excellent/free flow and F being failing/breakdown flow/traffic jam (AASHTO, 2004; TRB, 2010; http://en.wikipedia.org/wiki/Level_of_service).

Signs and signals also influence traffic flow. Similar to the LOS scheme for road segments, there is also an LOS grading scheme for delays at traffic signals. (A is less than or equal to 10 s for a traffic signal, and F is greater than or equal to 80 s.)

Finally, and also important, is how roads are marked and signs are designed and placed. In the United States, road markings and signs are specified in the *Manual of Uniform Traffic Control Devices*, commonly referred to as the MUTCD (FHWA, 2009b) and the AASHTO (2004) "Green Book." In Europe and elsewhere, signs and markings follow the Vienna Convention on Road Traffic [United Nations Economic Commission for Europe (UNECE), 1968].

Most driving studies have people drive without purpose, yet the purpose often provides a context as to how people drive and influences the road driven, time constraints, number of passengers, and cargo. That information, in turn, has implications for design requirements for vehicle handling and ride, stowage, and other characteristics. Vehicle travel is one of the few topics in this chapter for which there is good worldwide data (from national travel surveys), data that are publically available on the Internet. As of the date of this chapter, the Wikipedia entry on travel surveys has an extensive listing. The best-known travel studies are from the United States [Research and Innovative Technology Administration (RITA), 2003] and the U.K. Department of Transport (DfT, 2010). Travel studies contain data on the purpose of trips (work, social, shopping, etc.),

the transportation mode used (car, bus, etc.), the trip duration, the time of day, the number of occupants if by car, driver characteristics, and so forth. As an example of the scope, the 2001–2002 U.S. travel survey had interviews from over 26,000 households and was supplemented later by another 40,000 interviews. Additional information can be obtained from Transportation Research Board Committee ABJ40 (Committee on Travel Survey Methods).

## 3 WHAT HAPPENS WHEN PEOPLE DO NOT DRIVE WELL?

### 3.1 Crash Databases and Statistics

According to the World Health Organization (WHO, 2009), more than 1.2 million people die in motor crashes each year, or approximately 3300 per day. Furthermore, somewhere between 20 and 50 million people suffer injuries each year. Motor vehicle crashes are the ninth leading cause of death and are the leading cause of death of adults ages 15–29 (WHO, 2009). If the current trends continue, by the year 2030, traffic crashes will become the fifth largest cause of death after heart attacks, stroke, pneumonia, and lung diseases of various types (WHO, 2009).

Shown in Table 2 is an example of selected national differences drawn from the WHO's *World Report on Road Traffic Injury Prevention* (Peden et al., 2004). Notice that the more developed countries have a lower percentage of pedestrian fatalities, but that is not always the case (e.g., Thailand). Also note that in the Netherlands the number of cyclist fatalities is relatively large, as bicycle use is very common there. To provide some additional context, the National Bureau of Statistics of China (http://www.stats.gov.cn/tjsj/ndsj/2009/html/W2213e.htm) reports that 73,848 people were killed in traffic crashes in China in 2008. Curiously, according to those data, 21% were killed using motorcycles, 3% were killed using tractors, 2% were pedestrians and others, and 1% were bicyclists.

Overall, 91% of the deaths occur in low- and middle-income countries that have only 48% of the registered vehicles. In fact, their fatality rates (per 100,000 population) are double those of high-income countries. The largest numbers of deaths occur in China and India, and those totals are expected to increase as those countries become more motorized.

In 2009, approximately 34,000 people died in the United States in motor vehicle crashes, a decline of almost 9% from the previous year, even though total vehicle miles traveled increased by 0.2% (to 1.16 deaths per 100 million miles) (NHTSA, 2010a). However, U.S. crashes account for only a few percent of the world total.

Additional details on motor vehicle crashes are difficult to obtain because the United States is the *only* country in the world for which anyone with access to the Internet can get to the raw crash data at any time for free. No permission is required. Everywhere else, only aggregate statistics are available, though they are more extensive for European countries (Louma and Sivak, 2007). This lack of crash data from other than the United States (and access to it) leads to an overreliance on data from the United States, which can be misleading.

Crash analyses rely on three U.S. databases, (1) the Fatality Analysis Reporting System (FARS), (2) the National Automotive Sampling System (NASS) General Estimates System (GES), and (3) the Crashworthiness Data System (CDS). FARS (http://www.nhtsa.gov/people/ncsa/fars.html, retrieved September 6, 2010) contains data for all fatal crashes in the United States. GES (www.nhtsa.gov/people/ncsa/nass_ges.html, retrieved September 6, 2010) is a nationally representative sample of police-reported crashes of all degrees of severity (death, injury, or property damage). CDS (http://www.nhtsa.gov/PEOPLE/ncsa/nass_cds.html, retrieved September 6, 2010) is an annual probability sample of approximately 5000 police-reported crashes involving at least one passenger vehicle that was towed from the scene (from a population of almost 3.4 million tow-away crashes). Minor crashes (involving property damage only) are not in CDS. CDS crashes are investigated by specially trained teams of professionals who provide far more detail than is given in police reports.

These databases are extremely useful for those developing systems to mitigate crashes. For example, suppose one was developing a system to monitor driver alcohol levels. Statistics on the age and gender of intoxicated drivers and passengers, when crashes

### Table 2 Percentage of Road Users Killed by Various Transportation Modes

| Country | | Pedestrians | Cyclists | Motorized Two-Wheelers | Motorized Four-Wheelers | Other |
|---|---|---|---|---|---|---|
| More Developed | United States | 12 | 2 | 6 | 79 | 1 |
| | Norway | 16 | 4 | 12 | 65 | 3 |
| | Netherlands | 10 | 22 | 12 | 56 | 0 |
| | Australia | 18 | 3 | 11 | 67 | 1 |
| | Japan | 28 | 9 | 20 | 42 | 1 |
| Less Developed | India (Delhi) | 42 | 12 | 21 | 10 | 5 |
| | Indonesia (Bandung) | 33 | 7 | 42 | 15 | 3 |
| | Sri Lanka (Columbo) | 38 | 8 | 34 | 13 | 7 |
| | Thailand | 10 | 3 | 73 | 10 | 4 |

*Source*: Estimated from Peden et al., 2004, p. 42.

**Table 3  Precrash Events**

| Critical Precrash Event | | Weighted Percent | Subtotal Percent |
|---|---|---|---|
| Vehicle traveling | Turning or crossing at intersection | 36.2 | 69.6 |
| | Off the edge of the road | 22.2 | |
| | Over the lane line | 10.8 | |
| | Other | 0.7 | |
| Other vehicle in lane | Stopped | 12.2 | 17.3 |
| | Traveling in same direction | 4.8 | |
| | Traveling in opposite direction | 0.1 | |
| | Other | 0.1 | |
| Vehicle control loss | Traveling too fast | 5.0 | 8.9 |
| | Poor road condition | 2.0 | |
| | Vehicle problem | 1.2 | |
| | Other | 0.7 | |
| Other vehicle encroachment | | | 1.3 |

NHTSA (2008).

occurred, the roads and speeds driven, the precrash maneuvers, and so forth, could provide insight into how to design the system to maximize its effectiveness.

Traffic injuries are classified using the abbreviated injury scale (AIS) [Association for the Advancement of Automotive Medicine (AAAM), 2008]. There are six categories: (1) minor, (2) moderate, (3) serious, (4) severe, (5) critical, and (6) unsurvivable. The injury scale may be applied by the body part, for example, an AIS of 4 for an arm but 3 for the chest. There are also scales for organ injury and fracture classification (Chawda et al., 2004; Copes et al., 1990; Petrucelli, 1981).

To prevent crashes, one needs to know where, when, why, and how crashes occur. The U.S. National Motor Vehicle Crash Causation Survey (NHTSA, 2008) examined a statistically representative sample of 5471 crashes from 2005 to 2007. This unique report aggregated crashes by road, vehicle, traffic, and driver factors for that purpose. As an example, Table 3 shows that the most common precrash event was turning or crossing at an intersection, which suggests that intersection warning systems could be extremely helpful in reducing crashes.

### 3.2  Passive Safety Systems

Crash data have influenced the design of active and passive crash protection systems. Although issues relating to injury mechanisms have been primarily the purview of bioengineers, human factors engineers have been involved in the assessment of restraint systems as well as in design and programs to encourage use. For seat belts, the recent research focus has been on assessing effectiveness and increasing wear rates, with the overwhelming majority of publications concerning the United States (Eby et al., 2005; Strine et al., 2010), though there has been some research in Europe (e.g., Gras et al., 2007) and China (Routley et al., 2010). There has also been some general research on increasing use (Young et al., 2008).

Similarly, for air bags, the focus has been on reducing unintended injuries, increasing effectiveness,

and determining where additional air bags are needed (Braver etal., 2010; Carter and Maker, 2010).

Booster seats (Reed et al., 2009) and child restraint systems (child safety seats) can be very effective in reducing crash injuries to children. However, well over half of the time child safety seats are installed incorrectly, with the major problems being loose harness adjustment and loose seat belt adjustment (Decina and Lococo, 2005; Decina et al., 2009; Tsai and Perel, 2009). Additional research is needed to determine how installation errors can be reduced.

## 4  HOW SHOULD VEHICLES BE DESIGNED?

### 4.1  What Do Customers Want? (Clinics and JD Power Surveys)

Customer responses to vehicles can be gauged using tests in driving simulators and instrumented test vehicles, in naturalistic driving studies, and via other methods described elsewhere in this chapter (Green, 1993). Two particularly influential sources of information are clinics and JD Power data.

In a clinic, a manufacturer rents a facility to display a future product along with current model competitors. Many subjects, sometimes more than 100, are escorted from vehicle to vehicle and respond to a wide range of questions about each vehicle relating to fit and finish, appearance, comfort, ease of use, and so on. To get appropriate demographic data, clinics involving the same vehicles may be conducted in multiple cities (Curtis, 1996). Details about clinics are quite limited in the literature as manufacturers wish to keep their methods and findings secret. Sometimes clinics are combined with focus groups (Jalopnik, 2005).

JD Power conducts several annual surveys, summarized in confidential (and very expensive) reports. Their best-known automotive surveys are conducted in the United States, but they conduct surveys in other countries as well (Dance, 2010).

The JD Power Initial Quality Study (IQS) covers the first 90 days of new vehicle ownership and is conducted each year soon after the new models are produced. In the 2008/2009 U.S. survey, the target was 450 responses per vehicle model and 100 responses per manufacturing plant. There were more than 82,000 responses to the 288-question survey (http://www.autoblog.com/2010/06/17/jd-power-2010-initial-quality-study-domestics-lead-imports/, retrieved October 22, 2010).

The APEAL (Automotive Performance, Execution and Layout) survey concerns how satisfying a new vehicle is to own and drive based on 80 attributes. The 2010 results were based on 76,000 responses for 245 vehicle models (121 cars, 71 trucks, 53 crossover and utility vehicles). Customer ratings were obtained for 10 categories: vehicle interior, vehicle exterior, storage and space, audio/entertainment/navigation, seats, climate controls, driving dynamics, engine/transmission, visibility/driving safety, and fuel economy (http://www.jdpower.com/autos/articles/2010-APEAL-Study-Results/, retrieved October 22, 2010).

The VDS (Vehicle Dependability Survey) is conducted after three years of ownership. The JD Power target is 225 responses per model, with subjects completing an eight-page survey concerning 200 problem areas. The return rate for the last sample was 19% (http://www.jdpower.com/autos/articles/2010-Vehicle-Dependability-Study-Results/, retrieved October 22, 2010).

Users of these reports often assign considerable importance to small differences in rankings that most likely are within the limits of statistical error. Furthermore, there are often halo effects that bias responses. For example, there have been instances where the same component will receive a higher rating on a Lexus than on a Toyota. Nonetheless, JD Power ratings are treated as the absolute truth, and often they are close to it.

## 4.2 Vehicle Handling

Ideally, a vehicle should be designed from the inside out to accommodate drivers, passengers, and cargo. Commonly, however, motor vehicles are designed from the outside in to fit market and cost constraints and then modified to fit packaging requirements and customer input. Thus, the initial steps are to determine the wheelbase, length, width, height, weight, and then the suspension system to provide desired handling. In brief, if a vehicle cannot be driven, it will not sell.

Procedures for testing vehicle handling have existed for some time (e.g., Dugoff et al., 1970). That protocol examined five maneuvers: (1) limit braking (no steering), (2) response to rapid, extreme steering (no braking), (3) braking in a turn (fixed, nonzero steering angle), (4) rapid lane change, and (5) combined drastic steering and braking. Variations of these measures are still used today. Test procedures require careful control of the vehicle tire pressures, the loading of the vehicle, and the road surface and, for that reason, are often conducted on "black lakes."

Handling characteristics vary among vehicles. For example, the 2010 Chevrolet Corvette ZR1 will stop in 101 ft from 60 mi/h and handle about 1.0 g on a skid pad. For a 2011 Infiniti QX-56 SUV those values are 123 ft and 0.69 g. (See the full test results on Insideline.com.) Further careful examination shows significant differences between vehicles of the same type, different models of the same vehicle (due to tires), and within models due to loading. For example, there was a 22-ft difference in the stopping distance of common American pickup trucks (http://special-reports.pickuptrucks.com/2008/11/braking.html). In a *Popular Mechanics* test, a 2008 Ford F-150 King Ranch stopped in 20 ft less when it was "loaded."

Over time, many of these procedures have been standardized [International Organization for Standardization (ISO) Standard 3888; ISO, 1999, 2002a; New Car Assessment Procedures (NCAPs) (Forkenbrock et al., 2005); and Society of Automotive Engineers (SAE) Standards J1060 (SAE, 2000) and J1441 (SAE, 2007c)]. The extent to which Consumers Union, trade and enthusiasts' magazines, and websites (e.g., *Popular Mechanics*, *Road and Track*, *Edmunds Inside Line*) follow the recognized standards as part of their testing is unknown.

As this topic has been the subject of considerable public attention because of problems related to rollover, maneuvers have received memorable names. The moose test, which became famous because several well-known vehicles failed to pass it, involves a sudden lane change and then a return to that lane. The test simulates a driver avoiding a large animal (actually an elk in Sweden). (See http://en.wikipedia.org/wiki/Moose_test.) Other maneuvers have names such as J-turn and fishhook. (See http://en.wikipedia.org/wiki/J-turn.)

Finally, when reading the literature on handling, readers will see several terms of which they should be aware. *Understeer/oversteer* describes what happens to a vehicle if it is turning and the driver releases the steering wheel. With understeer, the vehicle will tend to straighten itself out, to which a limited degree is desired. With oversteer, the turn radius will decrease, which leads to some degree of instability. *Bump steer* is when some variability in the road surface (a bump or pothole) interferes with the ability to control the vehicle. *Body roll* is when the vehicle leans to the outside of a turn, and increases with a softer suspension and a greater height of the vehicle center of gravity. The *static stability factor*, the track width divided by two times the height of the vehicle center of gravity, indicates the vehicle's propensity to roll over. The larger the value, the less likely rollover is to occur. Values are about 1.40 for passenger cars, 1.24 for minivans, 1.17 for SUVs and pickup trucks, and 1.12 for full-size vans. (See http://en.wikipedia.org/wiki/Automobile_handling, NAS, 2002; Walz, 2005.)

Although this section emphasizes empiric test procedures, most of the work on the design and evaluation of handling qualities is done using computer models of the vehicle (especially Car-Sim and Truck-Sim, www.carsim.com) and the driver (Jagacinski and Flack, 2003; Macadam, 2003). See also Zschocke and Albers (2008).

## 4.3 Ride Quality

Ride quality analyses have their theoretical origins in human response to vibration. Studies in the 1950s

(Griffin, 1990) established the resonant frequencies of various human body parts and therefore which resonant frequencies to avoid (5 Hz for the abdomen, 25 Hz for the head, 50–200 Hz for gripping the steering wheel) to reduce discomfort.

Analyses of vehicle ride typically begin with quarter-car models, which consider the springs and shocks, along with the sprung and unsprung weight for each wheel. The key determinant of how a vehicle will ride is the transfer function that relates the input signal (the road) to the output (what the customer feels). The focus of those analyses is on body bounce and axle hop. Body bounce occurs in the 1–3-Hz range, with a peak at about 1 Hz, and relates to the combined motion of the entire body, in particular when both front and rear axles are simultaneously excited. That motion is influenced by vehicle length and speed. Axle hop, which occurs in the frequency range of 10–20 Hz (with a peak at 10 Hz), has to do with the tire's ability to stay in contact with the road surface.

The input signal, the road roughness, is measured using the international roughness index (IRI), an objective measure of the road profile. IRI is determined by driving a test device fitted with a probe whose cumulative vertical travel over a road segment is recorded, measured in millimeters per meter or inches per mile. In metric units, values of about 2 are typical for runways and express-ways in good repair, 1.5–2.5 for most new pavement, 2.5–6.0 for older pavement, 2.5–10.0 for maintained unpaved roads, and 4.0–11.0 for damaged pavement (http://training.ce.washington.edu/wsdot/Modules/09_pavement_evaluation/09-2_body.htm). The proce-dure for measuring IRI is described in American Soci-ety for Testing and Materials (ASTM) E950/E950M-09 (ASTM, 2009), 1364-95 (ASTM, 2005), and E1926-08 (ASTM, 2008). See also Sayers and Karamihas (1998).

The subjective response to the input is determined by having adults sit in a vehicle and rate the ride on a scale where a rating between 0 and 1 is very poor, and between 4 and 5 is very good. Originally, this measure was referred to as the present serviceability rating (PSR). The relationship between PSR and IRI depends on the study, but PSR $= 5e^{-0.2\text{IRI}}$ should provide an estimate. Currently, a more complex procedure that uses groups of subjects but the same scale results in a mean panel rating which can be estimated using the ride number, a 0–5 value similar to the PSR. The ride number is a number that can change over time as the population of roads and vehicles changes. (See Loizos, 2008.)

In practice, each manufacturer has its own procedure to determine ride quality. Tests are typically done at the manufacturer's test track, where at great expense they have reproduced pavement sections that duplicate real roads that challenge vehicle suspensions. For example, the GM Milford Proving Ground has a section that dupli-cates 12 Mile Road in Detroit (http://en.wikipedia.org/wiki/General_Motors_Proving_Grounds). Evaluation of off-road vehicles is much more complex (Els, 2005).

## 4.4 Vehicle Packaging (Occupant Space, Reach, and Field of View)

Packaging refers to designing the vehicle to accom-modate the people and cargo to be carried. Important considerations include (1) the location of the windows so that drivers can see both outside (to other vehicles, road signs, and traffic signals) and inside (to vehicle controls and displays), (2) the position of the driver (and passengers) so that they can sit comfortably and reach controls, and (3) the location of openings, handles, door sills, and steps so that occupants can get in and out eas-ily and also access the engine compartment (for service) and cargo.

The key dimensions for occupant packaging are specified in SAE and ISO standards [e.g., SAE J1100 (SAE, 2008b)], and summarized in Macey and Wardle (2009). See Figure 2. The primary landmarks are the heel point (where the aft part of the heel contacts the floor or pedal) and the H point (essentially the hip pivot point). Dimensions are commonly referred to by number, not name.

For passenger vehicle design, what drivers can see and reach is of critical concern. Drivers need to see traffic lights mounted above the road when they are stopped at intersections, information on the instrument panel (often through the steering wheel), other vehicles in the mirrors and, when they turn their heads, objects to the rear, especially small children close to the vehicle when the vehicle is backing up. Eye position is determined by the SAE eyellipse as specified in SAE J941 (SAE, 2010b and SAE J1052 (SAE, 2002b). See also Devlin and Roe (1968) and Manary et al. (1998). The eyellipse, a pair of football-like objects represents the three-dimensional distribution of driver's eyes such that on one side of a tangent to the eyellipse is either 95 or 99% of the driver population.

SAE J1050 describes three methods to determine monocular and binocular field of view. The most impor-tant obstructions are the A pillars (that frame the wind-shield) and the steering wheel. In theory, a pillar that is less than 65 mm wide, the interocular separation, should not completely block the driver's view of the road. For the speedometer/tachometer cluster, engineers need to verify that the cluster is inside the "mustache," the area of the speedometer/tachometer cluster not blocked by the steering wheel, so known because of its shape.

Recently, special attention has been given to devel-oping methods to assess field of view from large trucks, buses, and other large vehicles and to relate field of view to crashes (Blower, 2007).

When drivers cannot see something directly, they must be able to see it using cameras and/or mirrors. Mir-ror requirements are specified by Federal Motor Vehicle Safety Standard (FMVSS) 111 (http://www.fmcsa.dot.gov/rules-regulations/administration/fmcsr/fmcsrruletext.aspx?reg=571.111). The 111 standard specifies the size and field of view for interior and exterior mirrors for cars, trucks, buses, and motorcycles and cross-view mirrors for school buses and provides an assessment procedure. A cross-view mirror allows the driver to see pedestrians, usually children obscured by the front of the bus. This is a particular problem for school buses with a "nose" (engine forward).

Recent research has also emphasized using cameras to replace exterior mirrors, mainly for reasons of aerody-namics. However, because the image can be processed,
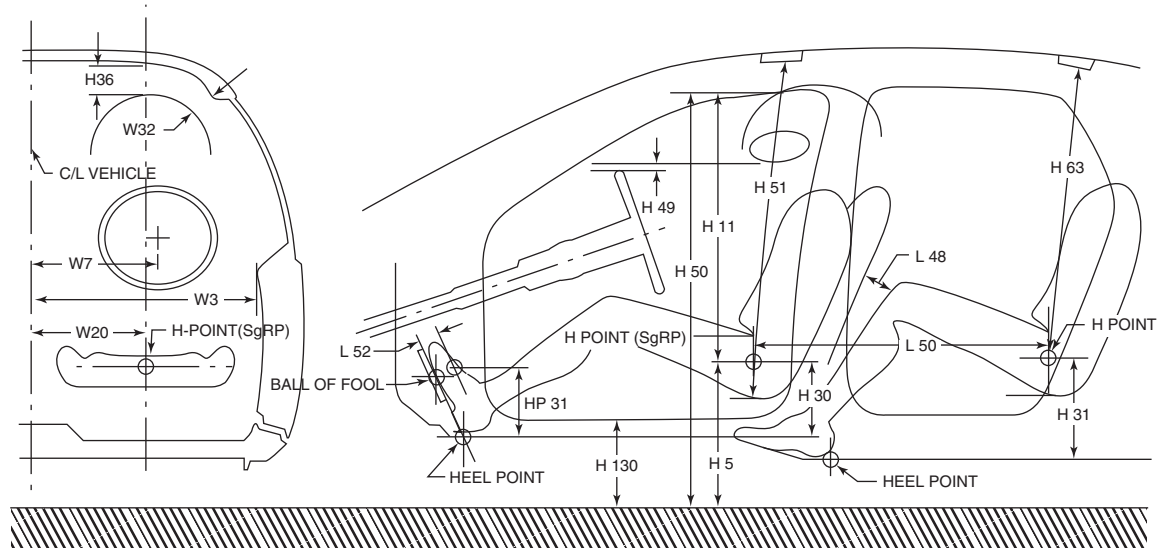
**Figure 2** Commonly used dimensions in SAE J1100 (http://www.instructables.com/image/FJK5LHFJZ3EQHO916N/Anthropometry-and-component-parts.jpg). Typical values from Macey and Wardle (2009, p. 95).

| # | Name | Typical Value (mm or deg) | # | Name | Typical Value (mm) |
|---|---|---|---|---|---|
| H5 | H point to ground | 475–550 | W3 | Shoulder room | 1350–1550 |
| H30 | Chair height | 250–275 | W5 | Hip room | 1325–1450 |
| H61 | Effective head room | 975 | W20 | Lateral location | 350–400 |
| A40 | Back angle | 22.0–24.0 | | | |
| A60 | Upward vision angle | 13.0–15.0 | | | |
| A61 | Downward vision angle | 5.0–7.0 | | | |

cameras can have advantages over mirrors, especially where there is glare, rain, fog, or snow. Systems that provide a bird's eye view around the vehicle are being developed as parking aids (Walls et al., 2004a, 2004b). At this point, published design standards for those systems do not exist.

Comfortable driver reach is determined per SAE standard J287 (SAE, 2007a; Figure 3). Although the adult population has changed since then, the primary changes have been in girth, not length. (See Parkinson and Reed, 2006.)

After handling, ride quality, and interior packaging are addressed, driver and passenger entry and exit are given attention (El Menceur et al., 2008) as well as reach to/service of items under the hood and in the trunk. The trend is to make increasing use of computerized biomechanical models such as Jack to predict reach, vision, and ingress/egress (Chateauroux et al., 2007; Dufour and Wang, 2005). Although these tools are quite powerful, learning how to position the simulated user and move about the environment is not easy (McInnes et al., 2009).

Those interested in developments in this topic should consult the proceedings from the latest SAE Digital Human Modeling for Design and Engineering Conference.

## 4.5 Vehicle Exterior Lighting

What drivers can see also depends on the amount of light present, and as demonstrated by the relative increase of crashes at night, lighting is critical. The research on headlamp design has been ongoing for decades, with the goal of optimizing the trade-off between road illumination for the driver and glare to oncoming drivers (Perel et al., 1983). Current research concerns the advantages of steerable headlamps and beam modifications, ideas that have been discussed for some time but have only recently become technically effective (Sivak et al., 2002). Another recent development, from naturalistic and other driving studies, is the collection of more complete data on the normal use of lighting systems (Buonarosa et al., 2008).

At this point, even after decades of research, there nonetheless is no single harmonized headlamp pattern. The United States and Canada comply with federal motor vehicle safety standard 108 [U.S. Department of Transportation (DOT), 2004] and Europe adheres to the UNECE (2005b) regulation with its sharper cutoff.

## 4.6 Basic Controls and Displays

Control and display selection and design occur after the basic package is designed. The design of basic vehicle controls and displays is relatively unconstrained, in
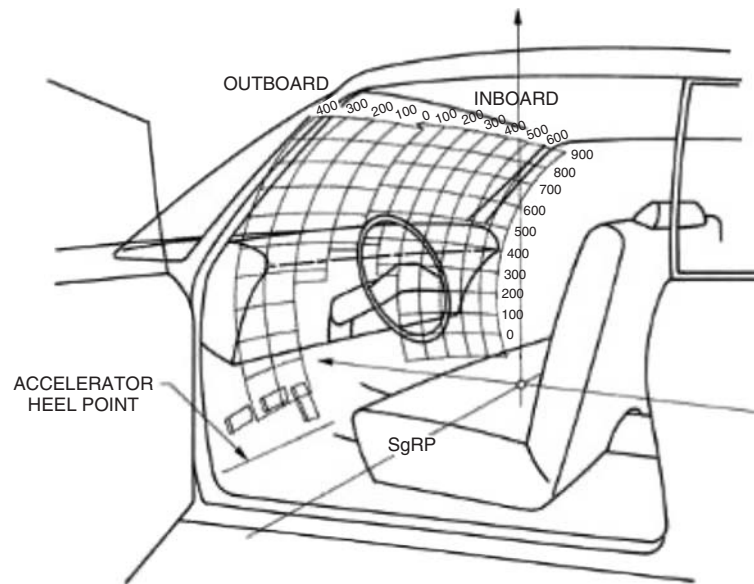
**Figure 3** SAE reach contours (http://www.sanatekhodro.com/CrThumb.aspx?Pic=sanatekh%5CImages%5C48%5C 102346223185900.jpg&X=348&Y=240).

particular with regard to control and display placement once basic vision and reach requirements, described earlier, are satisfied.

ISO 4040 (ISO, 2009) specifies where controls can be located. Generally, there are no recommendations for the details of stalk, knob, or button design, including their size. The companion SAE practice is J1138-2009 (SAE, 2009a).

One of the best-supported standards is ISO 12214, which provides recommended control-display movement relationships based on direction of motion stereotypes. The companion SAE practice is J1139 (SAE, 2010a).

The topic of symbols receives considerable attention and, in fact, there is an ISO working group (Technical Committee 22/Subcommittee 13/Working Group 5) solely concerned with symbols for road vehicles. The applicable standard, ISO 2575 (ISO, 2010a), has several hundred symbols, and that collection is growing. For some purposes, such as navigation functions, for which hundreds of symbols are needed, there are no standard symbols, so each manufacturer or supplier creates its own symbols.

### 4.7  Driving Performance Measures and Statistics

In the last 10 years, two topics have received considerable attention in the human factors literature and popular press: (1) driver assistance and warning systems and (2) driver distraction/overload and workload. During that time period, considerable research on these topics has been conducted and systems have been produced, though their deployment is not yet universal. For these two topics, there is considerable emphasis on

evaluation, which requires a clear definition of what is to be measured.

There have been many attempts to define driving performance measures and statistics (Green, 1993; Ostlund et al., 2005). Unfortunately, for many of the common measures and statistics, consistent naming is not used, and statistics and measures are not rigorously defined (Savino, 2009). For that reason, SAE recommended practice J2944 (Operational Definitions of Driving Performance Measures and Statistics) is being developed, some of whose measures and statistics are summarized in the section that follows.

One indication that a driver has responded to an event (a vehicle movement or a warning) is some movement of the foot pedals. However, a careful analysis of pedal movements, especially the accelerator, shows the driver is constantly making small adjustments to the pedal position, so distinguishing an overt change over some small time frame from normal variation is difficult. Table 4 shows the proposed definitions for responses to events, measured using response time measures. The definitions from J2944 have been abridged to save space. The names could change. The brake activation time is also usually the total response time. Note that for several of the measures there is a variable that is part of the name. For braking, it is undetermined at this point if a 10%, 50%, 90%, or some other percentage of maximum brake pressure is the critical measure.

Steering responses to events are analogous to accelerator response time measures in that responding involves detecting some amplitude change that is greater than normal variability over some time window. The difference is that there quite commonly is an accelerator/braking response to an event, but steering occurs only sometimes.

**Table 4 Candidate Definitions for Driving Response Time Measures**

| Measure Name | Description |
| --- | --- |
| Accelerator response time | From when an event/stimulus begins until some movement of the accelerator occurs in response to that event, either option A (the response ends when the accelerator position signal changes by more than a specified amount over a specified time window) or B (when the accelerator position signal reaches zero) |
| Accelerator to brake time | From accelerator release to brake application, either option A (when the driver's shoe first contacts the brake pedal) or B (when the brake lights first illuminate) |
| Initial brake response time | Equals accelerator response time + accelerator to brake time |
| Brake movement time ($x$%) | From when the brake is first contacted (version A) or the brake lamps are illuminated (version B) until some desired $x$ percent of maximum pressure is reached |
| Brake activation time ($x$%) | Equals initial brake response time + brake movement time ($x$%) |

There are no agreed-upon values in the literature for steering changes that signify an overt response.

In addition to showing the effects of a single event, steering changes in response to ongoing demand, such as workload from an in-vehicle task. Statistics of interest include the number of steering wheel reversals over some time period or the steering wheel reversal rate, with drivers making fewer and larger corrections when occupied by a secondary task (MacDonald and Hoffman, 1980).

A more contemporary measure of distraction and workload is steering entropy (Boer et al., 2005). In its simplest form, steering entropy (randomness) is a prediction of how well previous steering wheel positions predict the future. The more attentive to driving, the more stable the computation. The computations are not easy to understand.

The third category of statistics is related to lateral position. In brief, the less attention the driver is paying to staying in the lane, the more likely he or she is to strike another moving vehicle or roadside object such as a tree or parked car. Thus, when drivers are not fully engaged in the primary task of driving, one would expect the variability of lane position to increase and there to be more extreme lane positions, specifically lane departures.

As per SAE J2944, a lane departure may be considered to occur when (definition A) a tire, usually a front tire, touches the inside edge of a lane boundary or when (definition B) the widest part of the vehicle (either a mirror or the curvature of the body) is over the centerline of the lane boundary or when (definition C) a tire contacts the outside edge of the lane boundary. As most lane boundary stripes in the United States are 4 in. wide, the difference between definitions A and C is 4 in., a difference of practical significance. However, there are cases of a larger difference, for example, for definition B, where the vehicle is a pickup truck with large extended mirrors. There is no truly best definition, as the definition that is preferred depends upon the purpose for which it is used and the technology available to collect the data. From a safety perspective, definition B is preferred, as it best represents the situation in which two identical vehicles in adjacent lanes would just contact each other. Definition A is sometimes easier to measure. The author does not recommend definition C.

Also of interest is the vehicle's lateral position, with the standard deviation of lane position being the most commonly reported driving performance statistic. Investigators need to be extremely careful in determining lane position. In some older simulators, the road edges were constructed of chords, not true curves. Depending on the geometry, the vehicle–chord and vehicle–curve lateral distances could differ by several inches.

Another concern is the actual placement of the painted stripes on real roads. Road crews try to follow the pavement edge and expansion joints when painting markings, but the paint stripes may not be exactly at the lane edge. Thus, lane widths may vary even if the pavement does not. This can be particularly problematic when lane position is determined using a single painted stripe, not two, so apparent lateral movement could be random variation of a single paint stripe.

When a lane departure is expected, the time to line crossing, commonly referred to as TLC, is of interest (Godthelp, 1988; Godthelp and Kaeppler, 1988). Time to line crossing is the time required for some part of the vehicle to reach the edge marking (usually for a tire to touch the inside edge of the line) if the driver keeps to the same course. TLC can be computed three ways: definition A, trigonometrically; definition B, using lateral distance and velocity only; and definition C, using lateral distance, velocity, and acceleration (Van Winsum et al., 2000). The trigonometric calculation uses lateral velocity, lateral distance, lateral acceleration, the radius of curvature of the road, the radius of curvature of the path, and so forth, to compute an exact and error-free time. Surprisingly, when comparing definitions B and C, sometimes the estimate improves when lateral acceleration is omitted. Until now, on-road studies have not used definition A because the global positioning system (GPS) and road geometry data available in real time have not been able to provide all of the data needed. Similar data are available in a driving simulator.

Longitudinally of interest is the safety margin from the front of the subject vehicle to the rear of a lead vehicle, referred to as the gap (Figure 4). A gap can be measured either by distance or time, hence the names time gap and distance gap. Gaps can also be to the rear.

The distance or time between the front bumper of a vehicle to the same location on a following vehicle is
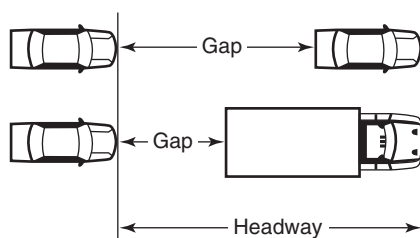
**Figure 4** Headway and gap.

called headway. Headway is important to civil engineers as an indicator of traffic flow. However, in many human factors studies, the term headway is used when gap is intended.

When a gap reaches zero, vehicles or objects collide. The measure that indicates the amount of safe clearance is time to collision (TTC). Similar to TLC, TTC is the time after which two objects will collide if the driver or drivers continue to do what they are doing (Godthelp et al., 1984; van der Horst and Hogema, 1994). Before being used in statistical summaries, TTC data need to be filtered, as TTC values can be infinite. Unfortunately, filtering rules are often not given when TTC statistics are provided.

### 4.8 Distraction/Overload and Workload Quantification, Assessment, and Specifications

A very prominent use of these and other measures and statistics is in studies of driver distraction/overload and workload, popular topics in the last few years (Regan et al., 2008; Rupp, 2010). The focus has been on the interference of cell phone conversations, whether there are differences in interference between hand-held and hands-free phones, the problems associated with texting, and so forth (Caird et al., 2006; Horrey and Wickens, 2006; McCartt et al., 2006; Collet et al., 2010a,b). Soon the use of websites and video calls while driving will become major issues.

A few key points from that literature deserve emphasis (Green, 2010):

1. Distraction and overload are not the same. Distraction has to do with attracting a person's attention and causing them to remain engaged in the task even though it may be unwise to do so (such as answering a ringing phone under almost any circumstance). Whereas distraction is a task-switching problem, overload is where the aggregate load of multiple tasks is just too much for a person to handle. Think of this as the classic spinning plates act in show business, where a performer tries to keep multiple plates spinning on poles and madly dashes between them.

2. Many of the driving workload studies are flawed because the workload of the primary task is specified qualitatively (light or heavy), not quantitatively, so test conditions cannot be rigorously compared. To solve that problem,

Schweitzer and Green (2007) developed a rating method that provides an absolute indication of the workload. Workload is rated relative to two specific anchor clips (having values of 2 for light traffic and 6 for heavy traffic, both exactly determined by those clips). In addition, workload can also be estimated using the equation that follows based on road geometry and traffic.

$$\text{workload} = 8.07 - 2.72(\text{LogMeanRange125})$$
$$+ 0.48(\text{MeanTrafficCount})$$
$$+ 2.17(\text{MeanAxFiltered})$$
$$- 0.34[\text{MinimumVpDot}$$
$$(0\text{removed})]$$

where

| | |
|---|---|
| LogMeanRange125 = | logarithm mean distances (m) to the same-lane lead vehicles. If there is no lead vehicle, mean distance = 125 m |
| MeanTrafficCount = | mean number of vehicles detected [15° field of view (FOV)] |
| MeanAxFiltered = | mean longitudinal acceleration (m/s$^2$) of the subject vehicle |
| MinimumVpDot (0 removed) = | minimum acceleration of lead vehicle (m/s$^2$), excluding the case of no lead vehicle |

3. The workload of secondary tasks needs to be specified quantitatively, not qualitatively (e.g., "the cognitive load was light"). One approach is to rate the workload of each subtask using the visual, auditory, cognitive, and psychomotor (VACP) scales from the U.S. Army IMPRINT model. Table 5 shows the UMTRI enhancement of the visual scale used in the SAVE-IT project as an example. For all four scales and further details of their use, see Yee et al. (2007).

There are a number of standards and guidelines pertaining to the design of driver interfaces for telematics. Noteworthy are the several-hundred-page-long guideline documents from Battelle (Campbell et al., 1997), the Harmonization of ATT Roadside and Driver Information in Europe (HARDIE) project (Ross et al. (1996), and UMTRI (Green et al., 1993). The more recent Transport Research Laboratory (TRL) guidelines are also worthy of note (Stevens et al., 2002). All of these guidelines have some use, but their coverage of hierarchical menu systems is limited.

Contemporary interface assessment follows the procedures described in the Alliance of Automobile Manufacturers principles (AAM, 2006; especially principles 2.1 and 2.2), the Japan Automobile Manufacturers

**Table 5  Visual Scale of the UMTRI-Modified IMPRINT VACP Scales**

| Rating | Definition | Example |
|---|---|---|
| 0.00 | No visual activity | Self-explanatory |
| 1.00 | Register/detect image | Observe a warning light turn on |
| 3.70 | Discriminate (detect visual difference) | Determine which traffic light is on |
| 4.00 | Inspect/check (static inspection) | Check side-mirror position while parked |
| 5.00 | Locate/align (selective orientation) | Change focus to a car |
| 5.40 | Track/follow (maintain orientation) | Watch a moving car |
| 5.90 | Read (symbol) | Read a native language |
| 7.00 | Scan/search/monitor (continuous) | Look through glove compartment |

(JAMA, 2004) guidelines, and SAE recommended practices J2364 (the 15-s task time rule; SAE, 2004) and J2365 (the calculation procedure for J2364; SAE 2002). The AAM principles are an expansion of the European Union guidelines (UNECE, 1999) that later became the European Statement of Principles (UNECE, 2005a).

Table 6 shows the primary ISO telematics guidelines and standards. Very few of these documents contain test criteria. See Green (2008a,b) for additional information.

### 4.9 Driver Assistance and Warning Systems

Driver warning and assistance systems are designed to reduce the workload of driving, stabilize traffic flow, and most importantly reduce opportunities for crashes. Included are adaptive cruise control systems (Ervin et al., 2005), lane departure warning and assistance systems (LeBlanc et al., 2006), fatigue warning systems

**Table 6  Primary ISO Standards and Related Documents for Telematics**

| Document | Short Title |
|---|---|
| Std 15005:2002(E) | Dialogue management principles and compliance procedures |
| Std 15006:2004 | Specifications and compliance procedures for auditory information |
| Std 15007-1:2002(E) | Visual behavior measurement 1: Definitions and parameters |
| Trial Std 15007-2:2001 | Visual behavior measurement 2: Equipment and procedures |
| Std 15008:2003 | Legibility (visual presentation of information) |
| Technical Report 16352:2005 | Warnings literature review |
| Trial Std 16951:2004 | Message priority |
| Std 17287:2002 | Suitability of interfaces while driving |
| Std 16673:2007 | Occlusion method to assess distraction |
| Committee Draft Std 26022 | Lane change test to assess distraction |

(Grace and Stewart, 2001; Wierwille et al., 1994), lane change/merge warning systems (Olsen, 2004; Van Winsum et al., 1999), blind-spot detection systems (Kiefer, and Hankey, 2008), forward crash warning systems (Kiefer et al., 2005; LeBlanc et al., 2001), curve speed warning systems, and intelligent speed assistance, to a name a few. In addition to assessing specific warning systems, there has been some more general research on developing test methods (Ference et al., 2007).

Based on this research, a large number of SAE, ISO, and DOT New Car Assessment Procedure (NCAP) procedures have been developed (Table 7). There is significant overlap and in some cases duplication between those three sets of procedures for particular warning systems, such as adaptive cruise control (ACC).

The next step is to extend the operational range of these standards, for example, to convert the ACC from high speed only to stop and go and to create others for scenarios not covered, such as intersections or to fill gaps, such as the SAE Lane Keeping Assistance System document. At this point, motor vehicles are transitioning from being controlled primarily by the driver to having semiautonomous control. That topic is in need of additional research.

## 5  SOURCES OF FURTHER INFORMATION

Although there are a substantial number of books on this topic, there are very few that have a design perspective, the emphasis of this chapter. Probably the best single reference is Peacock and Karwowski (1993), a dated book that is more a collection of chapters rather than a single integrated text. In terms of background, the author would also recommend Dewar and Olson (2007), which is in the process of being revised (and now being edited by Alison Smiley). Lawyers and Judges Publishing, the publisher of the Dewar and Olson volume, has a number of other books that consider driving from a forensic perspective.

Other books relating to human factors and driving include several written by Leonard Evans, the most recent of which is *Traffic Safety* (Evans, 2004). Evans's background in physics is reflected in his thoughtful and rigorous analyses of crash data and statistics.

Also worthy of note is a recent book by David Shinar, *Traffic Safety and Human Behavior* (Shinar, 2007), which focuses on factors that affect driving

**Table 7 Assessment Procedures for Driver Assistance and Warning Systems**

| Document | Short Title |
|---|---|
| ISO Std 3888-1:1999 | Test track test for a severe double lane change |
| ISO Std 3888-2:2002 | Test track test for obstacle avoidance |
| ISO Std 15622:2010 | Adaptive cruise control performance requirements and tests |
| ISO Std 15623:2002 | Forward vehicle collision warning performance requirements and tests |
| ISO Std 17361:2007 | Lane departure warning systems performance requirements and tests |
| ISO Draft Std 17387:2008 | Lane change decision aid systems (LCDAS) performance requirements and tests |
| ISO Std 22178:2009 | Low speed following (LSF) systems performance requirements and tests |
| ISO Std 22179:2009 | Full speed range adaptive cruise control (FSRA) systems performance requirements and tests |
| ISO Advanced Work Item 22839 | Forward vehicle collision mitigation systems — Operation and performance, requirements |
| ISO Std 22840:2010 | Devices to aid reverse manoeuvres — Extended-range backing aid systems (ERBA) |
| ISO/NP 26684 | Cooperative intersection signal information and violation warning systems (CISIVWS) |
| SAE Std J2399 | Adaptive cruise control (ACC) operating characteristics and user interface |
| SAE Information Report J2400 | Forward collision warning systems: Operating characteristics and user interface |
| SAE Recommended Practice 2802 | Blind spot system operating characteristics and user interface |
| SAE Information Report J2808 | Road/lane departure warning systems human interface |
| US DOT FCW NCAP | Forward crash warning system confirmation test |
| US DOT LDW NCAP | Lane departure warning system confirmation test |
| US DOT ESC NCAP | Electronic stability control confirmation test |

performance (e.g., driver capacities related to vision and age, driving style, alcohol, drugs, distraction, fatigue). The book is very thorough.

A useful complement to these books is the *H-Point* book (Macey and Wardle, 2009), which is concerned with vehicle packaging and is much more design oriented. It is not a human factors text per se but does contain useful background information.

The driving context is specified by three primary references used by civil and traffic engineers, of which all those doing automotive human factors work should be aware—the *Transportation Research Board (TRB) Highway Capacity Manual (* TRB, 2010), the *AASHTO "Green Book" (* AASHTO, 2004), so named because of the color of its cover, and the Manual of Uniform Traffic Control Devices (*MUTCD) Handbook (* FHWA, 2009b). These books were all described in Section 2.3.

Although books provide useful reference information, most professionals keep current by attending relevant technical conferences. Many automotive human factors engineers attend the annual Society of Automotive Engineers World Congress, for which there are sessions on human factors, vehicle lighting, and biomechanics. The Detroit location is convenient for those with offices in southeastern Michigan but draws attendees from all over the world. The meeting is held in mid-April. The SAE World Congress tends to have a very high acceptance rate, and paper quality is mixed.

Also well attended is the Transportation Research Board annual meeting, held in Washington, DC, in mid-January. The meeting draws a large number of government officials and has a strong highway focus, though there are sessions on other transportation modes as well. The meeting is quite large, and sessions are held at several hotels. The human factors workshops and the meetings of TRB safety and human factors committees are quite informative (http://www.trb.org/Safety HumanFactors/TRBCommittees.aspx).

For those interested in the cognitive aspects of driving, the best conference is Driving Assessment (drivingassessment.uiowa.edu/). This small, biennial conference is held in June at a resort. The papers are of high quality. The conference is single track and there are many opportunities for networking.

For those interested in the physical aspects of vehicle design, especially issues related to crash injuries, the primary conference is the Stapp Car Crash Conference (www.stapp.org/). This conference is usually held in late October or early November in a southern location in the United States. The paper quality is excellent.

Technical journals are the archival repository for engineering and scientific information. There are numerous journals of potential relevance to this topic, all of which have been referenced earlier. The journal that is most useful depends on one's interests—crash data analysis and cognitive aspects of driving or crash biomechanics. For research on the former, the primary journals are *Accident Analysis and Prevention* and *Transportation Research, Part F (Traffic Psychology and Behavior)* and to a lesser degree *Human Factors*. For the second category, the primary journal is the *Journal of Traffic Medicine*.

For many, the Internet is a popular source of information, generally searched using Google. In general, the quality of scientific and engineering information in journal articles and conferences is much better than websites because only the former are vetted. Using unsubstantiated information is risky. For that reason, scholar.google.com is preferred over google.com for searches.

Finally, for the purposes of this chapter, readers will find Cacciabue and Martinetto (2004) to be quite helpful. Many reports, along with many of the design guidelines, have been aggregated on the author's website (www.umich.edu/~driving).

## 6   KEY POINTS

1. Although there is reasonably good information on U.S. drivers, much less is known about drivers in emerging and critical markets such as China and India, making it difficult to design for customers in those countries. To a large extent, data on the kinds of vehicles they drive, the types of roads on which they drive, and their travel patterns are also limited. Additional information is needed.

2. There are several schemes for classifying vehicles that are well known. Similarly, how roads should be described is well established. This information should be more widely included in human factors reports to improve the replicability of human factors evaluations.

3. Except for the United States, no countries offer free, unconstrained, online access to crash databases. As the predominant types and causes of crashes vary between nations, relying on only data from the United States can be misleading.

4. Motor vehicle design is strongly influenced by the results from the JD Power IQS, APEAL, and VDS surveys.

5. Vehicle characteristics that determine handling and ride quality are established early in design. Handling and ride quality are assessed using accepted objective and subjective methods and statistics (e.g., moose test, ride number).

6. SAE J287, J941, J1100, and J1052 specify the vehicle packaging requirements.

7. ISO 4040, ISO 12214, SAE J1138, and SAE J1139 specify basic control design. ISO 2575 specifies symbols.

8. Driving performance measures and statistics are often undefined, and when they are defined, they are not defined consistently. Use of SAE J2944 should solve that problem.

9. Test conditions in distraction/overload and workload studies must be specified. The UMTRI workload equation and UMTRI-modified VACP scales could be used for that purpose.

10. There is a long list of SAE, ISO, and NCAP standards that pertain to the design of driver

assistance and warning systems and telematics. Most important are those that provide performance criteria (e.g., SAE J2364, AAM guidelines).

## REFERENCES

Alliance of Automobile Manufacturers (AAM) (2006), "Statement of Principles on Human-Machine Interfaces (HMI) for In-Vehicle Information and Communication Systems," Version 3.0, AAM, Washington, DC, available: www.umich.edu/~driving/guidelines/AAM_Statement2003.pdf, accessed July 15, 2007.

American Academy of Ophthalmology (AAO) (2006), *Policy Statement: Vision Requirements for Driving*, AAO, San Francisco, CA, available: http://www.aao.org/about/policy/upload/AAODrivingPolicyWebcopy.pdf.

American Association of State Highway and Transportation Officials (AASHTO) (2004), *A Policy on Geometric Design of Highways and Streets* ["*Green Book*"], 5th ed., AASHTO, Washington, DC.

American Society for Testing and Materials (ASTM) (2005), "Standard Test Method for Measuring Road Roughness by Static Level Method," Report ASTM E1364-95, ASTM, West Conshohocken, PA.

American Society for Testing and Materials (ASTM) (2008), "*Standard Practice for Computing International Roughness Index of Roads from Longitudinal Profile Measurements*," Paper ASTM E1926-08, ASTM, West Conshohocken, PA.

American Society for Testing and Materials (ASTM) (2009), "Standard Test Method for the Longitudinal Profile of Traveled Surfaces with an Accelerometer Established Inertial Profiling Reference," Report ASTM E950/E950M-09, ASTM, West Conshohocken, PA.

Association for the Advancement of Automotive Medicine (AAAM) (2008), *The Abbreviated Injury Scale (AIS) 2005—Update 2008*, AAAM, Committee on Injury Scaling, Des Plains, IL, available: http://www.carcrash.org/, accessed September 6, 2010.

Bertrand, M. Djankov, S., Hanna, R., and Mullainathan, S. (2008), "Corruption in Driving Licensing Process in Delhi," *Economic and Policy Weekly*, February 2, 71–76.

Blower, D. (2007), "Truck Mirrors, Fields of View, and Serious Truck Crashes," Report UMTRI-2007-25, University of Michigan Transportation Research Institute, Ann Arbor, MI.

Boer, E. R., Rakauskas, M. E., Ward, N. J., and Goodrich, M. A. (2005), "Steering Entropy Revisited," in *Proceedings of the 3rd International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, University of Iowa, Iowa City, IA, pp. 25–32.

Braver, E. R., Shardell, M., and Teoha. E. R. (2010), "How Have Changes in Air Bag Designs Affected Frontal Crash Mortality?" *Annals of Epidemiology*, Vol. 20, No. 7, pp. 499–510.

Buonarosa, M. L., Flannagan, M. J., and Bogard, S. (2008), "Real-World Frequency of Use of Automotive Lighting Equipment," *Leukos*, Vol. 5, No. 2, pp. 139–146.

Cacciabue, P. C., and Martinetto, M. (2004), "Report on the Review of the Available Guidelines and Standards," Report AIDE Deliverable 4.3.1, European Union, Brussels, Belgium, available: http://www.aide-eu.org/pdf/aide_d4-3-1_draft.pdf.

Campbell, J. L., Carney, C., and Kantowitz, B. (1997), "Human Factors Design Guidelines for Advanced Traveler Information Systems (ATIS) and Commercial Vehicle Operations (CVO)," Technical Report FHWA-RD-98-057, Washington, DC, U.S. Department of Transportation, Federal Highway Administration.

Caird, J. K., Scialfa, C. T., Ho, G., and Smiley, A. (2006), "A Meta-analysis of Driving Performance and Crash Risk Associated with the Use of Cellular Telephones while Driving," in *Proceedings of the Third International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, University of Iowa, Iowa City, IA, pp. 478–485.

Carter, P. R., and Maker, V. K. (2010), "Changing Paradigms of Seat Belt and Air Bag Injuries: What We Have Learned in the Past 3 Decades," *Journal of the American College of Surgeons*, Vol. 210, No. 2, pp. 240–252.

Chaffin, D. (2008), "Digital Human Modeling for Workspace Design," *Reviews of Human Factors and Ergonomics*, Vol. 4, pp. 41–74.

Chateauroux, E., Wang, X., and Trasbot, J. (2007), "A Database of Ingress/Egress Motions of Elderly People," SAE Paper 2007-01-2493, Society of Automotive Engineers, Warrendale, PA.

Chawda, M. N., Hildebrand, F., Pape, H. C., and Giannoudis, P. V. (2004), "Predicting Outcome after Multiple Trauma: Which Scoring System?" *Injury*, Vol. 25, No. 4, pp. 347–358.

Collet, C., Guillot, A., and Petit, C. (2010a), "Phoning While Driving I: A Review of Epidemiological, Psychological, Behavioural and Physiological Studies," *Ergonomics*, Vol. 53, No. 5, pp. 589–601.

Collet, C., Guillot, A., and Petit, C. (2010b), "Phoning While Driving II: A Review of Driving Conditions Influence," *Ergonomics*, Vol. 53, No. 5, pp. 602–616.

Copes, W. S., Champion, H. R., Sacco, W. J., Lawnick, M.M., Gann, D. S., Gennarelli, T., Mackienzie, E., and Schwaitzberg, S. (1990), "Progress in Characterizing Anatomic Injury," *Journal of Trauma*, Vol. 30, No. 10, pp. 1200–1207.

Curtis, M. (1996), "Putting Handheld CASI to the Test in an Automotive Consumer Clinic," paper presented at the International Conference on Computer-Assisted Survey Information Collection (CASIC), San Antonio, TX, available: http://www.researchinfo.com/docs/library/casi_auto.cfm.

Dance, D. (2010), "J.D. Power and Associates," presentation in *Industrial and Operations Engineering 437, Automotive Human Factors*, University of Michigan, Ann Arbor, MI.

Decina, L. E., and Lococo, K. H. (2005), "Child Restraint System Use and Misuse in Six States," *Accident Analysis and Prevention*, Vol. 37, No. 3, pp. 583–590.

Decina, L. E., Lococo, K. H., Ashburn, W., and Rose, J. (2009), "Identifying Strategies to Reduce the Percentage of Unrestrained Young Children," Report DOT HS 811 076, U.S. Department of Transportation, National Highway Traffic Safety Administration, Office of Behavioral Safety Research, Washington, DC.

Devlin W. A., and Roe R. W., (1968), "The Eyellipse and Considerations in the Driver's Forward Field of View," SAE Paper 680105, Society of Automotive Engineers, Warrendale, PA.

Dewar, R., and Olson, P. L. (2007), *Human Factors in Traffic Safety*, 2nd ed., Lawyers and Judges Publishing, Tucson, AZ.

Dufour, F., and Wang, X. (2005), "Discomfort Assessment of Car Ingress/Egress Motions Using the Concept of Neutral Movement," SAE Paper 2005-01-2706, Society of Automotive Engineers, Warrendale, PA.

Dugoff, H., Ervin, R. D., and Segel, L. (1970), "Vehicle Handling Test Procedures," Report HSRI PF-100a, University of Michigan, Highway Safety Research Institute, Ann Arbor, MI.

Eby, D. W., and Molnar, L. J. (2008), *2008 North American License Policies Workshop Proceedings*, AAA Foundation for Traffic Safety, Washington, DC.

Eby, D. W., Molnar, L. J., Kostyniuk, L. P., and Shope, J. T. (2005), "Developing an Effective and Acceptable Safety Belt Reminder System," paper presented at the 19th International Technical Conference on the Enhanced Safety of Vehicles, Washington, DC.

El Menceur, M. O. A., Pudlo, P., Gorce, P., Thevenon, A., and Lepoutre, F-X. (2008), "Alternative Movement Identification in the Automobile Ingress and Egress for Young and Elderly Population with or without Prostheses," *International Journal of Industrial Ergonomics*, Vol. 38, No. 11, pp. 1078–1087.

Els, P. S. (2005), "The Applicability of Ride Comfort Standards to Off-Road Vehicles," *Journal of Terramechanics*, Vol. 42, No. 1, pp. 47–64.

Ervin, R., Sayer, J., LeBlanc, D., Bogard, S., Mefford, M., Hagan, M., Bareket, Z., and Winkler, C. (2005), "Automotive Collision Avoidance System (ACAS) Field Operational Test Methodology and Results," Report US DOT HS 809 901, U.S. Department of Transportation, Washington, DC.

Evans, L. (2004), *Traffic Safety*, Science Serving Society, Bloomfield Hills, MI.

Federal Highway Administration (FHWA) (2009a), *Highway Statistics 2008*, U.S. Department of Transportation, FHWA, Washington, DC, available: http://www.fhwa.dot.gov/policyinformation/statistics/2008/, accessed October 16, 2010.

Federal Highway Administration (FHWA) (2009b), *Manual of Uniform Traffic Control Devices*, U.S. Department of Transportation, FHWA, Washington, DC.

Ference, J. J., Szabo, S., and Najm, W. G. (2007), "Objective Test Scenarios for Integrated Vehicle-Based Safety Systems," Paper No. 07-0183, in *Proceedings of the 20th International Technical Conference on the Enhanced Safety Vehicles*, National Highway Traffic Safety Administration, Washington, DC.

Forkenbrock, G. J., Elsasser, D., and O'Harra, B. O. (2005), "NHTSA's Light Vehicle Handling and ESC Effectiveness Research Program," Paper No. 05-0221, in *Proceedings of the 19th International Technical Conference on the Enhanced Safety Vehicles*, National Highway Traffic Safety Administration, Washington DC.

Godthelp, J. (1988), "The Limits of Path Error Neglecting in Straight Line Driving," *Ergonomics*, Vol. 31, No. 4, pp. 609–619.

Godthelp, J., and Kaeppler, W. D. (1988), "Effects of Vehicle Handling Characteristics on Driving Strategy," *Human Factors*, Vol. 30, No. 2, pp. 219–229.

Godthelp, J., Milgram, P., and Blaauw, G. J. (1984), "The Development of a Time-Related Measure to Describe Driving Strategy," *Human Factors*, Vol. 26, No. 3, pp. 257–268.

Gould, J. D., and Lewis, C. (1985), "Designing for Usability: Key Principles and What Designers Think," *Communications of the ACM*, Vol. 28, No. 3, pp. 300–311.

Grace, R., and Stewart, S. (2001), "Drowsy Driver Monitor and Warning System," paper presented at the International Driving Symposium on Human Factors in Driving Assessment, Training and Vehicle Design, University of Iowa, Iowa City, IA.

Gras, M. E., Cunill, M., Sullman, M. J. M., Planes, M., and Font-Mayolas, S. (2007), "Predictors of Seat Belt Use amongst Spanish Drivers," *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 10, No. 3, pp. 263–269.

Green, P. (1993), "Measures and Methods Used to Assess the Safety and Usability of Driver Information Systems," Report UMTRI-93-12, University of Michigan Transportation Research Institute, Ann Arbor, MI (also published as technical report FHWA-RD-94-088, U.S. Department of Transportation, Federal Highway Administration, McLean, VA).

Green, P. (1995), "Automotive Techniques," in *Research Techniques in Human Engineering*, Weimer, J., Ed. Prentice-Hall, Englewood Cliffs, NJ, pp. 165–208.

Green, P., Levison, W., Paelke, G., and Serafin, C. (1993), Preliminary Human Factors Guidelines for Driver Information Systems (Technical Report UMTRI-93-21), Ann Arbor, MI: The University of Michigan Transportation Research Institute (also published as FHWA-RD-94-087, McLean, VA: U.S. Department of Transportation, Federal Highway Administration).

Green, P. (2004), "Driver Distraction, Telematics Design, and Workload Managers: Safety Issues and Solutions," SAE Paper 2004-21-0022, in *Proceedings of the 2004 International Congress on Transportation Electronics (Convergence 2004 Conference Proceedings)*, Society of Automotive Engineers, Warrendale, PA.

Green, P. (2006), "Driver Status and Implications for Crash Safety," SAE Paper 2006-21-0028, in *Convergence 2006 Conference Proceedings*, Detroit, MI, Society of Automotive Engineers, Warrendale, PA.

Green, P. (2007), "Why Driving Performance Measures Are Sometimes Not Accurate (and Methods to Check Accuracy)," in *4th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, University of Iowa, Iowa City, IA.

Green, P. (2008a), "Driver Interface/HMI Standards to Minimize Driver Distraction/Overload," SAE Paper 2008-21-2002, in *Convergence 2008 Conference Proceedings*, Detroit, MI, Society of Automotive Engineers, Warrendale, PA.

Green, P. (2008b), "Driver Interface Safety and Usability Standards: An Overview," in *Driver Distraction: Theory, Effects, and Mitigation*, M. A. Regan, J. D. Lee, and K. L. Young, Eds., CRC Press, Boca Raton, FL, pp. 445–461.

Green, P. (2010), "Driver Distraction/Overload Research and Engineering: Problems and Solutions," SAE Paper 2010-01-2331, Society of Automotive Engineers, Warrendale, PA.

Griffin, M. J. (1990), *Handbook of Human Vibration*, Elsevier, San Diego, CA.

Horrey, W. J., and Wickens, C. D. (2006), "Examining the Impact of Cell Phone Conversations on Driving Using Meta-Analytic Techniques," *Human Factors*, Vol. 48, No. 1, pp. 196–205.

International Council of Ophthalmology (ICOPH) (2002), *Visual Standards: Aspects and Ranges of Vision Loss with Emphasis on Population Surveys*, ICOPH, Pacific Vision Foundation, San Francisco, CA, available: http://www.icoph.org/pdf/visualstandardsreport.pdf.

International Council of Ophthalmology (ICOPH) (2006), *Vision Requirements for Driving Safety with Emphasis on Individual Assessment*, ICOPH, San Francisco, CA.

International Organization for Standardization (ISO) (1999), "Passenger Cars—Test Track for a Severe Lane-Change Manoeuvre—Part 1: Double Lane-change," ISO standard 3888-1:1999, ISO, Geneva.

International Organization for Standardization (ISO) (2002a), "Passenger Cars—Test Track for a Severe Lane-Change Manoeuvre—Part 2: Obstacle Avoidance," ISO standard 3888-21:2002, ISO, Geneva.

International Organization for Standardization (ISO) (2002b), "Transport Information and Control Systems—Adaptive Cruise Control Systems—Performance Requirements and Test Procedures," ISO standard 15622:2002, ISO, Geneva.

International Organization for Standardization (ISO) (2004), "Intelligent Transport Systems—Lane Departure Warning Systems—Performance Requirements and Test Procedures," ISO standard 17361:2004, ISO, Geneva.

International Organization for Standardization (ISO) (2007), "Intelligent Transport Systems—Lane Departure Warning Systems—Performance Requirements and Test Procedures," ISO standard 17361:2007, ISO, Geneva.

International Organization for Standardization (ISO) (2009), "Road Vehicles—Location of Hand Controls, Indicators, and Tell-Tales in Motor Vehicles," ISO standard 4040:2009, ISO, Geneva.

International Organization for Standardization (ISO) (2010a), "Intelligent Transport Systems—Lane Keeping Assistance Systems—Performance Requirements and Test Procedures," ISO document TC 204/SC N470.12: 2010, ISO, Geneva.

International Organization for Standardization (ISO) (2010b), "Road Vehicles—Symbols for Controls, Indicators and Tell-Tales," ISO standard 2575:2004/Amendment 5:2019, ISO, Geneva.

International Organization for Standardization (ISO) (2010c), "Road Vehicles—Direction of Stereotypes for Automotive Hand Controls," ISO standard 12214:2010, ISO, Geneva.

International Road Federation (IRF) (2009), *IRF World Road Statistics 2009 (Data 2002-2007)*, IRF, Geneva.

Jagacinski, R. J., and Flack, J. (2003), *Control Theory for Humans: Quantitative Approaches to Modeling*, Lawrence Erlbaum Associates, Mahwah, NJ.

Jalopnik (2005), "You Are There: Inside a Ford Edge Focus Group," available: http://www.jalopnik.com/cars/industry-news/you-are-there-inside-a-ford-edge-focus-group-143847.php, accessed September 25, 2006.

Japan Automobile Manufacturers Association (JAMA) (2004), "JAMA Guideline for In-Vehicle Display Systems," Version 3.0, JAMA, Tokyo, available: www.jama.or.jp/safe/guideline/pdf/jama_guideline_v30_en.pdf, accessed July 15, 2007.

Kiefer, R., LeBlanc, D., and Flannagan, C. (2005), "Developing an Inverse Time to-Collision Crash Alert Timing Approach Based on Driver's Last-Second Braking and Steering Judgments," *Accident Analysis and Prevention*, Vol. 37, No. 2, pp. 295–303.

Kiefer, R. J., and Hankey, J.M. (2008), "Lane Change Behavior with a Side Blind Zone Alert System," *Accident Analysis and Prevention*, Vol. 40, No. 2, pp. 683–690.

Kozak, K., Pohl, J., Birk, W., Greenberg, J., Artz, B., Blommer, M., Cathey, L., and Curry, R. (2006), "Evaluation of Lane Departure Warnings for Drowsy Drivers," in *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 2400–2404.

LeBlanc, D., Kiefer, R., Deering, R., Shulman, M., Palmer, M., and Salinger, J., (2001), "Forward Collision Warning: Preliminary Requirements for Crash Alert Timing," SAE Paper 2001-01-0462, Society of Automotive Engineers, Warrendale, PA.

LeBlanc, D., Sayer, J., Winkler, C., Ervin, R., Bogard, S., Devonshire, J., Mefford, M., Hagan, M., Bareket, Z., Goodsell, R., and Gordon, T. (2006), "Road Departure Crash Warning System Field Operational Test: Methodology and Results," Report UMTRI-2006-9-2, U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, DC.

Leonard, R. (2002), *Statistics on Vision Impairment: A Resource Manual*, 5th ed., Arlene R. Gordon Research Institute of Lighthouse International, New York, available: http://www.gesta.org/estudos/statistics0402.pdf.

Loizos, A. (2008), "An Alternative Approach to Pavement Roughness Evaluation," *International Journal of Pavement Engineering*, Vol. 9, No. 1, pp. 69–78.

Louma, J., and Sivak, M (2007), "Charateristics and Availability of Fatal Road-Crash Databases in 20 Countries Worldwide," *Journal of Safety Research*, Vol. 38, pp. 323–327.

Macadam, C. (2003), "Understanding and Modeling the Human Driver," *Vehicle System Dynamics*, Vol. 40, Nos. 1-3, pp. 101–134.

MacDonald, W. A., and Hoffman, E.R. (1980), "Review of the Relationships between Steering Wheel Reversal and Driving Task Demand," *Human Factors*, Vol. 22, No. 6, pp. 733–739.

Macey, S., and Wardle, G. (2009), *H-Point: The Fundamentals of Car Design and Packaging*, Art Center College of Design, Pasadena, CA.

Manary, M. A., Flannagan, C. A. C, Reed, M. P., and Schneider, L. W. (1998), "Development of an Improved Driver Eye Position Model," SAE Paper 980012, Society of Automotive Engineers, Warrendale, PA.

McCartt, A. T., Hellinga, L. A., and Bratiman, K. A. (2006), "Cell Phones and Driving: Review of Research," *Traffic Injury Prevention*, Vol. 7, No. 2, pp. 89–106.

McInnes, B., Stephens, A. and Potvin, J. (2009), "Within and between-Subject Reliability Using Classic Jack for Ergonomic Assessments," in *Digital Human Modeling, HCII 2009, Lecture Notes In Computer Science*, Vol. 5620, V. G. Duffy, Eds., Springer, New York, pp. 653–660.

Monnier, G., Chateauroux, E., and Wang, X. (2009), "Inverse Dynamic Reconstruction of Truck Cabin Ingress/Egress Motions," *SAE International Journal of Passenger Cars-Mechanical Systems*, Vol. 2, No. 1, pp. 1593–1599.

National Academy of Sciences (NAS) (2002), "An Assessment of the National Highway Traffic Safety Administration's Rating System for Rollover Resistance," TRB Special Report 265, NAS, Transportation Research Board, Washington, DC.

National Highway Traffic Safety Administration (NHTSA) (2007), "Rear Detection System for Single-Unit Trucks," Federal Motor Vehicle Safety Standard 111, U.S. Department of Transportation, NHTSA, Washington, DC.

National Highway Traffic Safety Administration (NHTSA) (2008), *National Motor Vehicle Crash Causation Survey*, Report DOT HS 811059, U.S. Department of Transportation, NHTSA, Washington, DC.

National Highway Traffic Safety Administration (NHTSA) (2010a), *Early Estimate of Motor Vehicle Traffic Fatalities in 2009*, Report DOT HS 811 291, National Center for Statistics and Analysis, U.S. Department of Transportation, NHTSA, Washington, DC.

National Highway Traffic Safety Administration (NHTSA) (2010b), "Forward Crash Warning System NCAP Confirmation Test March 2010," Docket NHTSA-2006-26555-0128, U.S. Department of Transportation, NHTSA, Office of Crash Avoidance Standards, Washington, DC.

National Highway Traffic Safety Administration (NHTSA) (2010c), "Lane Departure Warning System NCAP Confirmation Test March 2010," Docket NHTSA-2006-26555-0125, U.S. Department of Transportation, NHTSA, Office of Crash Avoidance Standards, Washington, DC.

Olsen, E. (2004), "Lane Change Warning Design Guidelines," in *Proceedings of the 48th Annual Meeting of the Human Factors and Ergonomics Society*, Human Factors and Ergonomics Society, Santa Monica, CA.

Ostlund, J., Peters, B., Thorslund, B., Engstrom, J., Markkla, G., Keinath, A., Horst, D., Juch, S., Mattes, S., and Foehl, U. (2005), "Driving Performance Assessment—Methods and Metrics," Report AIDE Deliverable D2.2.5, European Union, Brussels, available: http://www.aide-eu.org/pdf/sp2_deliv_new/aide_d2_2_5.pdf.

Parkinson, M. B., and Reed, M. P. (2006), "Optimizing Vehicle Occupant Packaging," SAE Paper 2006-01-0961, Society of Automotive Engineers, Warrendale, PA.

Peacock, B., and Karwowski, W. (1993), *Automotive Ergonomics*, CRC Press, Boca Raton, FL.

Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A. A., Jarawan, E., and Mathers, C. (2004), *World Report on Road Traffic Injury Prevention*, World Health Organization, Geneva.

Perel, M., Olson, P. L., Sivak, M., and Medlin, J. W. (1983), "Motor Vehicle Forward Lighting," SAE Paper 830567, Society of Automotive Engineers, Warrendale, PA.

Petrucelli, E. (1981), "The Abbreviated Injury Scale (AIS)—Ten Years of Progress," SAE PAPer 810212, Society of Automotive Engineers, Warrendale, PA.

Reed, M. P., Ebert, S. M., Sherwood, C. P., Klinich, K. D., and Manary, M. A. (2009), "Evaluation of the Static Fit Provided by Belt Positioning Booster Seats," *Accident Analysis and Prevention*, Vol. 41, No. 3, pp. 598–607.

Regan, M. A., Lee, J. D., and Young, K. L., Eds. (2008), *Driver Distraction: Theory, Effects, and Mitigation*, CRC Press, Boca Raton, FL.

Research and Innovative Technology Administration (RITA) (2003), *Highlights of the 2001 National Household Travel Survey*, Report BTS03-05, U.S. Department of Transportation, RITA, Bureau of Transportation Statistics, Washington, DC, available: http://www.bts.gov/publications/highlights_of_the_2001_national_household_travel_survey/pdf/entire.pdf.

Ross, T., Midtland, K., Fuchs, M., Pauzie, A., Engert, A., Duncan, B., Vaughn, G., Vernet, M., Peters, H., Burnett, G., and May, A. (1996), "HARDIE Design Guidelines Handbook: Human Factors Guidelines for Information Presentation by ATT Systems," Commission of the European Communities, Luxembourg.

Routley, V., Ozanne-Smith, J., Yu, M., Wang, J., Wu, M., Zhang, J., Qin, Y., and Zhao, M. (2010), "Focus on Seat Belt Use in China," *Traffic Injury Prevention*, Vol. 11, No. 6, pp. 578–586.

Rupp, G., Eds. (2010), *Performance Metrics for Assessing Driver Distraction: The Quest for Improved Road Safety*, Society of Automotive Engineers, Warrendale, PA.

Savino, M. R. (2009), "Standardized Names and Definitions for Driving Performance Measures," unpublished masters thesis, Tufts University, Medford, MA.

Sayers, M. W., and Karamihas, S. M. (1998), *The Little Book of Profiling*, University of Michigan Transportation Research Institute, Ann Arbor, MI.

Schweitzer, J., and Green, P. A. (2007), "Task Acceptability and Workload of Driving Urban Roads, Highways, and Expressway: Ratings from Video Clips," Report UMTRI-2006-6, University of Michigan Transportation Research Institute, Ann Arbor, MI.

Shinar, D. (2007), *Traffic Safety and Human Behavior*, Elsevier, Amsterdam, The Netherlands.

Sivak, M., Schoettle, B., Flannagan, M., and Nakata, Y. (2002), "Benefits of Applying Adaptive Lighting to the U.S. and European Low-Beam Patterns," SAE Paper 2002-01-0524 (also in *Advanced Lighting Technology for Vehicles*, Report SP-1668), Society of Automotive Engineers, Warrendale, PA.

Society of Automotive Engineers (SAE) (2000), "Subjective Rating Scale for Evaluation of Noise and Ride Comfort Characteristics Related to Motor Vehicle Tires," SAE Recommended Practice J1060, SAE, Warrendale, PA.

Society of Automotive Engineers (SAE) (2002a), "Calculation of the Time to Complete In-Vehicle Navigation and Route Guidance Tasks," SAE Recommended Practice J2365, SAE, Warrendale, PA.

Society of Automotive Engineers (SAE) (2002b), "Motor Vehicle Driver and Passenger Head Position," SAE Standard J1052, SAE, Warrendale, PA.

Society of Automotive Engineers (SAE) (2003a), "Adaptive Cruise Control Operating Characteristics and User Interface," SAE Standard J2399, SAE, Warrendale, PA.

Society of Automotive Engineers (SAE) (2003b), "Describing and Measuring the Driver's Field of View," SAE Recommended Practice J1050, SAE, Warrendale, PA.

Society of Automotive Engineers (SAE) (2003c), "Human Factors in Forward Collision Warning Systems: Operating Characteristics and User Interface Requirements," SAE Information Report J2400, SAE, Warrendale, PA.

Society of Automotive Engineers (SAE) (2004), "Navigation and Route Guidance Function Accessibility While Driving," SAE Recommended Practice 2364, SAE, Warrendale, PA.

Society of Automotive Engineers (SAE) (2007a), "Driver Hand Control Reach," SAE Standard J287, SAE, Warrendale, PA.

Society of Automotive Engineers (SAE) (2007b), "Road/Lane Departure Warning Systems: Information for the Human Interface," SAE Information Report J2802, SAE, Warrendale, PA.

Society of Automotive Engineers (SAE) (2007c), "Subjective Rating Scale for Vehicle Handling," SAE Standard J1441, SAE, Warrendale, PA.

Society of Automotive Engineers (SAE) (2008 draft), "Blind Spot Monitoring System (BSMS): Operating Characteristics and User Interface," SAE Recommended Practice J2802, SAE, Warrendale, PA.

Society of Automotive Engineers (SAE) (2009a), "Design Criteria—Driver Hand Controls Location for Passenger Cars, Multipurpose Passenger Vehicles, and Trucks (10 000 GVW and Under)," SAE Standard J1138, SAE, Warrendale, PA.

Society of Automotive Engineers (SAE) (2009b), "Motor Vehicle Dimensions," SAE Standard J1100, SAE, Warrendale, PA.

Society of Automotive Engineers (SAE) (2010a), "Direction of Motion Stereotypes for Automotive Hand Controls," SAE Recommended Practice J1139, SAE, Warrendale, PA.

Society of Automotive Engineers (SAE) (2010b), "Motor Vehicle Driver's Eye Locations," SAE Standard J941-2010, SAE, Warrendale, PA.

Stevens A., Quimby, A., Board, A., Kersloot, T., and Burns, P. (2002), "Design Guidelines for Safety of In-Vehicle Information Systems," Project Report PA3721/01, Transport Research Laboratory, Wokingham, Berkshire, United Kingdom.

Strine, T. W., Beck, L. F., Bolen, J., Okoro, C., Dhingra, S., and Balluz, L. (2010), "Geographic and Sociodemographic Variation in Self-Reported Seat Belt Use in the United States," *Accident Analysis and Prevention*, Vol. 42, No. 4, pp. 1066–1071.

Transportation Research Board (TRB) (2010), *Highway Capacity Manual*, National Academy of Sciences, TRB, Washington, DC.

Tsai, Y-. D., and Perel, M. (2009), "Drivers' Mistakes When Installing Child Seats," Technical Report, U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, DC.

U.K. Department for Transport (DfT) (2010), *National Travel Survey Statistical Release*, DfT, London, available: http://www.dft.gov.uk/pgr/statistics/datatablespublications/nts/latest/nts2009-01.pdf.

United Nations Economic Commission for Europe (UNECE) (1968), *Convention on Road Traffic*, UNECE, Brussels, available: http://www.unece.org/trans/conventn/crt1968e.pdf.

United Nations Economic Commission for Europe (UNECE) (1999), "Statement of Principles on Human Machine Interface (HMI) for In-Vehicle Information and Communication Systems," Annex 1 to Commission Recommendation of 21 December 1999 on Safe and Efficient In-Vehicle Information and Communication Systems: A European Statement of Principles on Human Machine Interface, UNECE, Brussels.

United Nations Economic Commission for Europe (UNECE) (2005a), "European Statement of Principles on the Design of Human Machine Interaction" (ESoP 2005) draft, UNECE, Brussels, available: ec.europa.eu/information_society/activities/esafety/doc/esafety_library/esop_hmi_statement.pdf, accessed August 15, 2007.

United Nations Economic Commission for Europe (UNECE) (2005b), "Uniform Provisions Concerning the Approval of Motor Vehicle Headlamps Equipped with Gas-Discharge Light Sources," Regulation 98, UNECE, Geneva.

U.S. Department of Transportation (DOT) (2004), "Lamps, Reflective Devices, and Associated Equipment— Passenger Cars, Multipurpose Passenger Vehicles, Trucks, Buses, Trailers, (except pole trailers and trailer converter dollies), and Motorcycles," Federal Motor Vehicle Safety Standard 108, DOT, Washington, DC.

Van der Horst, A. R. A., and Hogema, J. H. (1994), "Time-to-Collision and Collision Avoidance Systems," in *Proceedings 6th ICTCT Workshop*, Kuratorium fur Verkehrssicherheit, Salzburg, Austria.

Van Winsum, W., de Waard, D., and Brookius, K. A. (1999), "Lane Change Manoeuvers and Safety Margins," *Transportation Research, Part: F*, Vol. 2, No. 3, pp. 139–149.

Van Winsum, W., Brookhuis, K. A., and deWaard, D. (2000), "A Comparison of Different Ways to Approximate Time-to-Line Crossing (TLC) during Car Driving," *Accident Analysis and Prevention*, Vol. 32, No. 1, pp. 47–56.

Walls, S. M., Amann, J., Cullinane, B., Green, P., Gadgil, S., and Rubin, R. (2004a), "Alternative Images for Perpendicular Parking: A Usability Test of a Multi-Camera Assistance System," Report UMTRI 2004-17, University of Michigan Transportation Research Institute, Ann Arbor, MI.

Walls, S. M., Green, P., Gadgil, S., Amann, J., and Cullinane, B. (2004b), "Alternative Images for Parallel Parking: A Usability Test of a Multi-Camera Assistance System," Report UMTRI 2004-31, University of Michigan Transportation Research Institute, Ann Arbor, MI.

Walz, M. C. (2005), "Trends in the Static Stability Factor of Passenger Cars, Light Trucks, and Vans," Report DOT HS 809 868, U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, DC.

Wierwille, W. W., Ellsworth, L. A., Wreggit, S. S., Fairbanks, R. J., and Kirn, C. L. (1994), "Research on Vehicle-Based Driver Status/performance Monitoring: Development, Validation, and Refinement of Algorithms for Detection of Driver Drowsiness," Report DOT HS 808 247, U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, DC.

World Health Organization (WHO) (2009), *Global Status Report on Road Traffic Safety*, WHO, Geneva, available: http://www.who.int/entity/violence_injury_prevention/publications/road_traffic/world_report/en/index.html, accessed September 8, 2010.

Yee, S., Nguyen, L., Green, P. A., Oberholtzer, J., and Miller, B. (2007), "Visual, Auditory, Cognitive, and Psychomotor Demands of Real In-Vehicle Tasks," Report UMTRI-2006-20, University of Michigan Transportation Research Institute, Ann Arbor, MI.

Young, K. L., Regan, M. A., Triggs, T. J., Stephan, K., Mitsopoulos-Rubens, and Tomasevic, N. (2008), "Field Operational Test of a Seatbelt Reminder System: Effects on Driver Behaviour and Acceptance," *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 11, No. 6, pp. 434–444.

Zschocke, A. K., and Albers, A. (2008), "Links between Subjective and Objective Evaluations Regarding the Steering Character of Automobiles," *International Journal of Automotive Technology*, Vol. 9, No. 4, pp. 473–481.

# CHAPTER 59

# HUMAN FACTORS AND ERGONOMICS IN AUTOMATION DESIGN

**John D. Lee and Bobbie D. Seppelt**
**University of Wisconsin—Madison**
**Madison, Wisconsin**

## 1 INTRODUCTION

Automation has a long history marked by many successes and equally notable failures. In the early nineteenth century the Luddites in northern England protested against the introduction of automation in the weaving industry by sabotaging the machines. Although the term *luddite* now refers to technophobes, these people correctly foresaw some of the highly damaging changes that automation would bring to their lives. Automation and the industrial revolution radically changed the craft-centered culture of the time. More recently, information technology has had an equally important effect on industries as diverse as process control, aviation, and ship navigation.

The Luddites foresaw the threat to their lifestyle. Of greater concern are situations in which people fail to recognize the risks of adopting automation and are surprised by unanticipated effects (Sarter et al., 1997). Automation frequently surprises designers, operators, and managers with unforeseen mishaps. As an example, the cruise ship *Royal Majesty* ran aground because the global positioning system (GPS) signal was lost and the position estimation reverted to position extrapolation based on speed and heading (dead reckoning). For over 24 h, the crew followed the compelling electronic chart display and did not notice that the GPS signal had been lost or that the position error had been accumulating. The crew failed to heed indications from boats in the area, lights on the shore, and even salient changes in water color that signal shoals. The surprise of the GPS failure was discovered only when the ship ran aground [National Transportation Safety Board (NTSB), 1997; Lutzhoft and Dekker, 2002]. This mishap demonstrates the power of technology to either make us smart or surprisingly stupid (Norman, 1993). Automation exemplifies this power.

*Automation* has been defined as a device or system that performs a function previously performed by a human operator (Parasuraman et al., 2000). However, automation does not simply supplant the person, but enables new activities, creates new roles for the person, and changes existing activities in unexpected ways (Woods, 1994). As a result, automation often produces surprises at many levels, from the societal, as with the

Luddites, to the individual, as with the *Royal Majesty*. For automation to achieve its intended benefits, its design must anticipate these changes. The need to anticipate and avert automation-related surprises is more difficult now than ever. One of the ironies in automation design is that as automation increasingly supplants human control, it becomes increasingly important for designers to consider the contribution of the human operator (Bainbridge, 1983). This chapter draws upon over 30 years of research to identify general automation-related failures and to identify strategies for improving automation design.

## 2   AUTOMATION PROMISES AND PITFALLS

Automation has many clear benefits. In the case of the control of cargo ships and oil tankers, it has made it possible to operate a vessel with as few as 8–12 crew members compared to the 30–40 that were required 40 years ago (Grabowski and Hendrick, 1993). In the case of aviation, automation has reduced flight times and increased fuel efficiency (Nagel, 1988). Similarly, automation in the form of decision support systems has been credited with saving millions of dollars in guiding policy and production decisions (Singh and Singh, 1997). Automation promises greater efficiency, lower workload, and fewer human errors; however, these promises are not always fulfilled.

Many pitfalls plague the introduction of automation. Well-documented failures of information technology show that it seldom provides the promised economic benefits (Landauer, 1995) and often fails to provide promised safety benefits (Perrow, 1984). When automation is introduced to eliminate human error, the result is sometimes new and often more catastrophic errors (Sarter and Woods, 1995). Automation often fails to provide expected benefits because it does not simply replace the human in performing a task but also transforms the job and introduces new tasks. These new tasks are not always recognized and so designers fail to provide operators with adequate feedback and support. Automation also fails because the role of the person performing the task is often underestimated, particularly the ability to compensate for the unexpected, and that role is not supported. Although any list of automation-related problems and surprises will be incomplete, changes in feedback, task structure, and relationships represent critical challenges of automation design (Lee and Seppelt, 2009).

### Feedback changes:
- Out-of-the-loop unfamiliarity
- Surprising mode transitions
- Inadequate training and skill loss

### Task structure changes:
- Clumsy automation
- Automation task errors
- Behavioral adaptation

### Relationships change:
- Mismatched expectations and eutactic behavior
- Inappropriate trust (misuse, disuse, and complacency)
- Job satisfaction and health

### 2.1   Feedback Changes

Automation often fails because it dramatically changes the feedback the operator receives. Diminished or eliminated feedback is a common occurrence with automation and it can leave people less prepared to detect automation failures or to intervene.

*Out-of-the-loop unfamiliarity* refers to the diminished ability of people to detect automation failures and to resume manual control (Endsley and Kiris, 1995). Several factors underlie this problem. First, automation might reduce feedback, and the remaining feedback may be qualitatively different than that received when operating under manual control (McFadden et al., 2003). With manual control operators often have both proprioceptive and visual cues, whereas under automatic control they may have only visual cues (Wickens and Kessel, 1981). Automation also reduces feedback because it distances operators from the process. Introducing automation into papermaking plants eliminated the informal feedback associated with vibrations, sounds, and smells that many operators relied upon (Zuboff, 1988). Second, monitoring the performance of automation involves passive observation of changes in system state, which is qualitatively different than the active monitoring associated with manual control (Gibson, 1962; Eprath and Young, 1981). In manual control, perception actively supports control, and control actions guide perception (Flach and Jagacinski, 2002). Monitoring automatic control disrupts this process. Third, automatic control can induce the operator to disengage and direct attention to other activities, further compromising the feedback from the system. The tendency to rely complacently on automation, particularly during multitask situations, may reflect this tendency to disengage from the monitoring task (Parasuraman et al., 1993, 1994; Metzger and Parasuraman, 2001). Finally, the operator's mental model may be inadequate to guide expectations and control. In particular, the automation may use control algorithms that are at odds with the control strategies and mental model of the person, making it difficult to anticipate the actions and limits of the automation (Goodrich and Boer, 2003). Operators with substantial previous experience and well-developed mental models detect disturbances more rapidly than operators without this experience, but extended periods of monitoring automatic control may undermine this skill and diminish operators' ability to generate expectations of correct behavior (Wickens and Kessel, 1981). This skill loss may also undermine operators' self-confidence, which can make them less inclined to intervene (Lee and Moray, 1994). Overall, out-of-the-loop unfamiliarity stems from disrupted feedback that diminishes the ability to form correct expectations, detect anomalies, and control the system manually.

An example of out-of-the-loop unfamiliarity occurs in driving. Adaptive cruise control (ACC) has the

potential to induce out-of-the-loop unfamiliarity, leading to delayed and less effective braking responses in situations in which the ACC is not able to respond to a braking lead vehicle (Stanton and Young, 1998). ACC uses sensors to maintain not only a set speed, as with conventional cruise control, but also a set distance to cars ahead. When drivers engage ACC, they no longer receive the haptic feedback that conveys the degree of braking needed to slow the vehicle in response to the braking behavior of vehicles ahead. Drivers also revert to passive monitoring of other vehicles rather than directing their attention toward the active control of their headway. Most important, ACC may induce drivers to direct their attention to nondriving activities such as cell phone conversations or reading the newspaper (Ward, 2000). Such distractions clearly delay driver response (Lee and Strayer, 2004). More subtly, drivers may have a poor mental model of the ACC control algorithms and so may not be able to anticipate situations that lie beyond the capabilities of the automation.

The difficulty of anticipating the behavior of the automation has also been seen in aviation, in which verbal protocol data show that pilots have problems with automation because of poor feedback and difficulties developing expectations regarding the behavior of the automation (Olson and Sarter, 2001). Analysis of eye movements, mental model probes, and pilot behavior showed that inadequate feedback and the associated poor monitoring are major contributors to automation mishaps (Sarter et al., 2007).

*Mode errors* often result from poor monitoring associated with poor feedback (Woods, 1994; Sarter and Woods, 1995). These arise when operators fail to detect the mode or recognize the consequence of mode transitions in complex automation. Substantial research with cockpit automation demonstrates that flight management systems often surprise pilots with unexpected mode transitions. These complex systems use a combination of the pilots' commands and system coupling to transition between modes. Mode transitions are often not commanded explicitly by the pilot and sometimes go unnoticed (Sarter and Woods, 1995).

Electronic charts in maritime navigation also offer the potential for mode errors. Such charts have several modes for determining a ship's position. One uses GPS data, another uses position extrapolation based on speed and heading (dead reckoning) to estimate the ship's position. If the GPS signal is lost, the electronic chart system changes automatically to the dead reckoning mode. This mode transition is signaled by a short alarm. If the alarm is not detected, however, the mariner may not notice that the GPS signal is no longer the basis for position estimates. Furthermore, many electronic charts do not maintain a continuous visual record of the vessel track. A track line is shown as long as the same chart or scale is used, but if the scale is changed, the track line is lost. The lack of track line continuity further undermines the ability of mariners to detect a transition from GPS to dead reckoning position estimates. If the mariner does not notice this mode transition, the ship can drift many miles from the intended course while the electronic chart continues to display the position

as if the vessel were following that course precisely. This is exactly what happened in the grounding of the cruise ship *Royal Majesty*, where the GPS signal was lost and the position estimation transitioned to the dead reckoning mode. The mode transition was noticed only when the ship ran aground (NTSB, 1997).

*Skill loss* refers to automation that leaves the operator without the appropriate skills to accommodate the demands of the job. In situations in which the automation takes on the tasks previously assigned to the operator, the skills of the operator may atrophy as they go unexercised (Endsley and Kiris, 1995). Part of this skill loss reflects diminished feedback. This is a particular concern in aviation, where pilots' aircraft handling skills may degrade when they rely on the autopilot. In response, some pilots disengage the autopilot and fly the aircraft manually to maintain their skills (Billings, 1997).

## 2.2 Task Structure Changes

*Clumsy automation* refers to the situation in which automation makes easy tasks easier and hard tasks harder (Wiener, 1989). As Bainbridge (1983) notes, designers often leave the operator with the most difficult tasks—those designers are unable to automate. Because the easy tasks have been automated, the operator has less experience and an impoverished context for responding to the difficult tasks, as a result of the out-of-the-loop problem mentioned above. In this situation, automation has the effect of both reducing workload during already low-workload periods and increasing it during high-workload periods. For example, a flight management system tends to make the low-workload phases of flight (such as straight and level flight or a routine climb) easier but high-workload phases (such as the maneuvers in preparation for landing) more difficult, as pilots have to share their time between landing procedures, communication, and programming the flight management system. Such effects are seen not only in aviation but also in the operating room (Cook et al., 1990b; Woods et al., 1991). The unfortunate tendency of operators to more willingly delegate tasks to automation during periods of low workload, compared to situations of high workload (Bainbridge, 1983), increases the prevalence of clumsy automation. This observation demonstrates that clumsy automation is not simply a technical problem, but one that depends on operator attitudes such as trust (Lee and See, 2004; Madhavan and Wiegmann, 2007).

The burden of clumsy automation is more prevalent than reported because operators adapt to clumsy automation, either tailoring their tasks or configuring the automation to adapt to poorly designed automation (Cook et al., 1990a). These strategies can mask the effects of clumsy automation in routine situations and make it appear more effective than it really is. When operators encounter abnormal situations, the problems of clumsy automation may emerge unexpectedly.

An example of potentially clumsy automation in maritime navigation occurs when the GPS is integrated with digital charts to create electronic chart display and information systems (ECDISs). When combined

with existing advanced maritime navigation systems (e.g., automatic radar plotting aid), these technological innovations tend to reduce repetitive physical activity while potentially increasing the mental demands made on the crew. The reduction in physical demands implies the possibility of reducing the number of personnel required on the bridge from as many as four people (captain, watch officer, helmsman, and lookout) to one. Recent studies suggest that under proper conditions workload declines and performance rises with one-person operations (Schuffel et al., 1988); however, this research has addressed only routine performance and has not considered more stressful conditions. Software failures and dense traffic situations combine to increase the workload substantially relative to the traditional system (Lee and Sanquist, 1996). This finding is consistent with poorly designed automation in the aviation and operating room, which reduces workload under routine conditions but increases it during stressful conditions (Wiener, 1989; Woods, 1991).

Clumsy automation can occur at the individual and organizational levels. Automation promises to reduce the need for human labor; during routine circumstances, fewer people are able to control the system effectively. The dramatic reduction in crew members needed to operate large ships testifies to this fact. However, clumsy automation at the macrolevel can occur when abnormal situations or high-tempo operations challenge the resources of the diminished crew (Lee and Morgan, 1994). Frequently, the wider effects of automation on training and recruitment go unexamined (Strain and Eason, 2000). Clumsy automation at the microlevel of the operator and the macrolevel of the organization represent critical challenges in anticipating the effect of automation.

*Automation–task errors* refer to the new forms of human error associated with new tasks generated by the introduction of automation. Managers and system designers often introduce automation to eliminate human error. Ironically, new and more disastrous errors can sometimes result. Automation often extends the scope of human actions and delays feedback associated with those actions. As a consequence, human errors may be more likely to go undetected and do more damage.

New automation-related tasks imply new skills are needed. Sophisticated automation eliminates many physical tasks and leaves complex cognitive tasks that may appear superficially easy, leading to less emphasis on training and a poor understanding of the automation. On ships, misunderstanding of new radar and collision avoidance systems has contributed to accidents (NTSB, 1990). One contribution to such accidents is training and certification that fail to reflect the demands of the automation. An analysis of the exam used by the U.S. Coast Guard to certify radar operators indicated that 75% of the items assess skills that have been automated and are not required by the new technology (Lee and Sanquist, 2000). The new technology makes it possible to monitor a greater number of ships, enhancing the need for interpretive skills such as understanding the rules of the road and the automation. These are the very skills that are underrepresented on the Coast Guard exam. While increasing automation might relieve the operator

of some tasks, they are likely to create new and more complex tasks that require more, not less, training.

*Brittle failures* are typical of human–automation interactions in which novel problems arise or even simple data entry mistakes are made with systems that completely automate the decision process and leave operators to assess the automation's decision (Roth et al., 1987; Roth and Woods, 1988). Such failures contrast with *graceful degradation*, a common characteristic of time-tested manual processes. Brittle failures are characterized by a sudden and dramatic decline in system performance, whereas graceful degradation is characterized by a more gradual and predictable decline. For example, a flight-planning system for pilots can induce dramatically poor decisions because it assumes that weather forecasts represent reality and lacks the flexibility to consider situations in which the actual weather might deviate from the forecast (Smith et al., 1997).

In maritime navigation, electronic charts introduce the potential for brittle failures in position estimation. Electronic charts distance mariners from the process of recording vessel position, leaving them with little insight into the factors that might lead to erroneous position estimates. The manual process of recording a position on a paper chart superimposes at least two position estimates, one based on extrapolation of the previous position and one based on visual bearings or other position information. These complementary position estimates help identify errors in determining position (Hutchins, 1995). Unlike the manual position recording on paper charts, electronic charts show the quality of the position estimation only indirectly, in terms of GPS signal quality; however, many mariners have little understanding of the relevance of these numbers, and gross errors in position can result (Lee and Sanquist, 2000).

Automation-related tasks also introduce the opportunity for *configuration errors*. Many forms of automation involve complex configurations or setups, and mistakes made during this process can later prove disastrous. For example, with electronic charts that aid maritime navigation, it is possible to configure the system to test the actual position automatically against the intended track using a feature in which an acceptable safety margin can be specified. If the ship deviates beyond this distance, an alarm sounds (provided that the feature was engaged and the GPS is functioning normally). Failing to engage this feature could jeopardize ship safety if mariners have come to rely on the automated warning. Also, because any one of several mariners can configure the system, system configuration and behavior can change in unanticipated ways as different mariners enter different safety margins. The danger of an inappropriate or unanticipated chart configuration is not a failure mode associated with paper charts but represents an automation-induced error that can threaten ship safety.

Brittle failures and configuration errors tend to undermine individual reliability and may have even greater detrimental effects on team performance (Skitka et al., 2000b). These automation-related errors may be particularly troublesome if the automation also undermines effective error-correcting strategies such as feedback and redundancies in the multiperson position-fixing process

(Hutchins, 1995). For example, because poor position fixes are visible to all team members, crew members who generate these fixes are corrected quickly. Through configuration errors, automation also creates the potential for one team member to induce errors in other team members. These factors all point toward the need to consider automation-induced errors at the level of the team and at the level of the individual.

*Behavioral adaptation* refers to the tendency of operators to adapt to the new capabilities of the automation, particularly to change behavior and tasks so that the potential safety benefits of the technology are not realized. Automation intended by designers to enhance safety may instead lead operators to reduce effort and leave safety unaffected or even diminished. Behavioral adaptation occurs at the individual (Wilde, 1988, 1989; Evans, 1991), organizational (Perrow, 1984), and societal levels (Tenner, 1996).

Antilock brake systems (ABSs) for cars show behavioral adaptation. The ABS modulates brake pressure automatically to maintain maximum brake force without skidding. This automation makes it possible for drivers to maintain control in extreme crash avoidance maneuvers, which should enhance safety. However, ABSs have not produced the intended safety benefits, in part because with them drivers tend to drive less conservatively, adopting higher speeds and shorter following distances (Sagberg et al., 1997). Similarly, vision enhancement systems make it possible for drivers to see more at night, potentially enhancing safety; however, drivers tend to adapt to the systems by increasing their speed (Stanton and Pinto, 2000).

Behavioral adaptation can also undermine the benefits of automation if the automation causes a diffusion of responsibility and a tendency to exert less effort when automation is available (Mosier et al., 1998; Skitka et al., 2000a). As a result, people tend to commit more omission errors (failing to detect events not detected by the automation) and more commission errors (concurring incorrectly with erroneous detection of events by the automation) when they work with automation. Automation can lead people to conserve cognitive effort rather than increase detection performance. This effect parallels the adaptation of people when they work in groups. Diffusion of responsibility leads people to perform more poorly when they are part of a group compared to individually (Skitka et al., 1999). A similar phenomenon is seen with decision support automation. People often use decision support systems to reduce effort rather than to enhance decision quality (Todd and Benbasat, 1999, 2000). The strong tendency of people to minimize effort and adapt their behavior to the most salient feedback they receive merits careful consideration in the design and implementation of automation. The effects of behavioral adaptation, particularly the diffusion of responsibililiy, suggest automation can change relationships between people, a topic we turn to next.

## 2.3 Relationships Change

Automation redefines not just tasks but also relationships between co-workers, with management, and between designers and users. Critical to defining these relationships is the degree to which people rely and comply with automation. Often designers and managers expect operators to rely on automation in a way that diverges from how people actually use the automation, or how they need to use the automation to maintain system safety and performance.

*Eutactic behavior* is behavior that approximates an optimal or satisficing response to the automation (Moray, 2003). As a consequence, eutactic behavior is not an instance of inappropriate reliance on automation but an instance of appropriate reliance that may be inconsistent with the expectations of the designers or managers. Misuse and disuse may sometimes reflect poorly calibrated trust, automation bias, or complacency. However, misuse and disuse may also reflect eutactic behavior and appropriate reliance once the costs and benefits are assessed completely. Automation that is generally reliable should be relied upon, even if it fails periodically, if the costs of a failure are modest, and if it relieves the operator of substantial mental effort. Careful monitoring to catch the periodic failure might not be worth the effort in such a case. What may appear to be complacent behavior may actually be appropriate given the costs of monitoring.

Discriminating between complacency and eutactic behavior requires optimizing a cost function that includes the cost of failing to detect failures *and* the cost of monitoring (Moray, 2003). Overreliance may be appropriate given the cost of monitoring. Similarly, disuse may also be appropriate. The new tasks associated with programming, engaging, monitoring, and disengaging automation can make the burden of managing the automation outweigh its benefit (Kirlik, 1993). In this situation, the aid will go unused by a well-adapted operator (Kirlik, 1993). Such behavior is eutactic and should be expected, but designers might be surprised if they fail to consider the burden of automation-related tasks.

Not all over- and underreliance is appropriate. Often operators respond to automation inappropriately, exhibiting a tendency toward misuse and disuse. *Misuse* refers to the failures that occur when people inadvertently violate critical assumptions and rely on automation inappropriately, whereas *disuse* signifies failures that occur when people reject the capabilities of automation (Parasuraman and Riley, 1997).

Another useful distinction in how operators use automation is that of reliance and compliance (Meyer, 2001). *Reliance* refers to the situation in which the operator does not act because the automation has not issued a warning or seems to be performing adequately. In contrast, *compliance* refers to the situation in which the operator acts in response to a warning or command from the automation. Overreliance results in errors of omission (failing to detect events not detected by the automation), and overcompliance results in errors of commission (concurring incorrectly with erroneous detection of events by the automation; Skitka et al., 2000a). Underreliance and compliance are differentially affected by false alarms and misses. Automation prone to false alarms affects compliance and reliance, but miss-prone automation tends to affect only reliance (Dixon et al., 2007).

The influence of false alarms on reliance and compliance is complicated. Although a high rate of false alarms often induces a cry wolf effect and an associated disuse of automation (Bliss et al., 1995), in some cases, such as air traffic conflict alerting systems, the relatively high rate of false alarms did not lead to disuse (Wickens et al., 2009). One reason for this effect is that false alarms are not a homogeneous class of warnings. Users may view some false alarms as useful and other false alarms might help them understand the system and so do not undermine compliance and reliance (Lees and Lee, 2007).

Misuse and disuse of automation may depend on certain attitudes of users, such as trust and self-confidence (Lee and Moray, 1994; Dzindolet et al., 2001). As an example, the difference in operators' trust in a route planning aid and their self-confidence in their own ability was highly predictive of reliance on the aid (de Vries et al., 2003). Many studies have demonstrated that trust is a meaningful concept to describe human–automation interaction, in both naturalistic settings (Zuboff, 1988) and laboratory settings (Halprin et al., 1973; Lee and Moray, 1992; Muir and Moray, 1996; Lewandowsky et al., 2000). People tend to rely on automation they trust and to reject automation they do not trust. In the context of operator reliance on automation, trust has been defined as an attitude that the automation will help achieve an operator's goals in a situation characterized by uncertainty and vulnerability (Lee and See, 2004).

Inappropriate reliance associated with misuse and disuse depends in part on how well trust matches the true capabilities of the automation. Calibration, resolution, and specificity of trust describe the match between trust and the capabilities of automation. *Calibration* refers to the correspondence between a person's trust in automation and the automation's capabilities (Lee and Moray, 1994; Lee and See, 2004). Definitions of the appropriate calibration of trust parallel those of misuse and disuse in describing appropriate reliance. Overtrust is poor calibration in which trust exceeds system capabilities; with distrust, trust falls short of automation capabilities. Figure 1 shows good calibration as the diagonal line where the level of trust matches automation capabilities. Above this line is overtrust and below is distrust. Overreliance on automation has sometimes been termed *complacency* and can result from trusting the automation more than is warranted.

*Resolution* refers to how precisely a judgment of trust differentiates levels of automation capability (Cohen et al., 1999). Figure 1 shows that poor resolution occurs when a large range of automation capability maps onto a small range of trust. With low resolution, large changes in automation capability are reflected in small changes in trust. *Specificity* refers to the degree to which trust is associated with a particular component or aspect of the trustee. Functional specificity describes the differentiation of functions, subfunctions, and modes of automation. With high functional specificity, a person's trust reflects capabilities of specific subfunctions and modes. Low functional specificity means that the person's trust reflects the capabilities of the entire system. Specificity can also describe changes in trust as a function of the situation over time. High temporal specificity means that a person's trust reflects moment-to-moment fluctuations in automation capability, whereas low temporal specificity means that the trust reflects only long-term changes in automation capability. Although temporal specificity implies a generic change over time as the person's trust adjusts to failures in the automation,
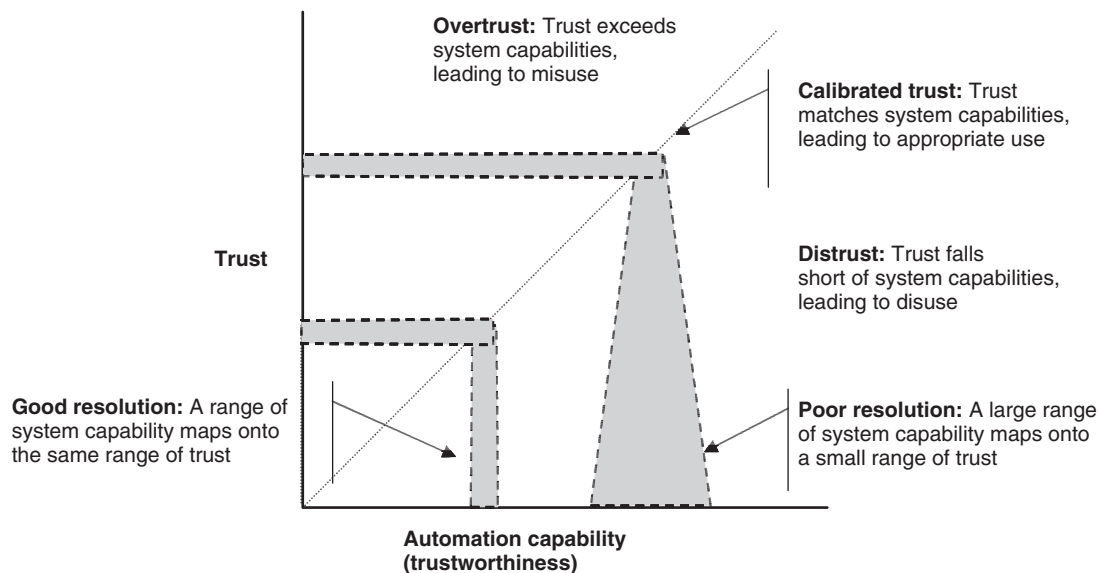


**Figure 1** Calibration, resolution, and automation capability define appropriate trust in automation. Overtrust may lead to misuse, and distrust may lead to disuse. (Reprinted with permission from Lee and See, 2004. Copyright 2004 by the Human Factors and Ergonomics Society.)

temporal specificity also addresses adjustments that should occur when the situation or context changes and affects the capability of the automation. High functional and temporal specificity increase the likelihood that the level of trust will match the capabilities of a particular element of the automation at a particular time. Good calibration, high resolution, and high specificity of trust can mitigate misuse and disuse of automation.

The information required to support appropriate trust can be considered in terms of attributional abstraction, which varies from the demonstrations of competence to the intentions of the automation (Lee and See, 2004). A recent review of trust literature concluded that three general levels summarize the bases of trust: ability, integrity, and benevolence (Mayer et al., 1995). Lee and Moray (1992) made similar distinctions in defining the factors that influence trust in automation and identified performance, process, and purpose as the general bases of trust.

*Performance* refers to the current and historical performance and reliability of the automation. Performance information describes *what* the automation does. More specifically, performance refers to the competency or expertise of the system, as demonstrated by its ability to achieve the operator's goals. Because performance is linked to the ability to achieve specific goals, it demonstrates the task- and situation-dependent nature of trust. This is similar to Sheridan's (1992) concept of robustness. The operator will tend to trust automation that performs in a manner that reliably achieves his or her goals.

*Process* is the degree to which the algorithms of the automation are appropriate for the situation and able to achieve the operator's goals. Process information describes *how* the automation operates. In interpersonal relationships, this corresponds to the consistency of actions associated with adherence to a set of acceptable principles (Mayer et al., 1995). Process as a basis for trust reflects a shift away from focus on specific behaviors and toward qualities and traits attributed to the automation. With the process dimension, trust is in the automation and not in the specific actions of the automation. As an example, knowing why automation failed increased trust even when it was not warranted (Dzindolet et al., 2003). In contrast, trust tends to drop with any sign of incompetence of the automation, even if the overall system performance is unaffected (Muir and Moray, 1996). Thus, the process basis of trust relies on dispositional attributions and inferences and is similar to Sheridan's (1992) concept of understandability. The operator will tend to trust the automation if its algorithms can be understood and it seems capable of achieving the operator's goals in the current situation.

*Purpose* refers to the degree to which the automation is being used within the realm of the designer's intent. It addresses the question of *why* the automation was developed. With interpersonal relationships, this depends on the intentions and motives of the trustee. This can take the form of abstract, generalized value congruence (Sitkin and Roth, 1993), which can be described as whether and to what extent the trustee has a motive to lie (Hovland et al., 1953). The purpose basis of trust reflects the attribution of these characteristics to the automation. Frequently, whether or not this attribution takes place will depend on whether the designer's intent has been communicated to the operator. If so, the operator will tend to trust the automation to achieve the goals it was designed to achieve. Often, the complexity, authority, and autonomy of the automation lead to a perceived animacy, in which the automation seems capable of independent and willful action independent of the operator (Sarter and Woods, 1994). In this situation, the intents that the operator infers may have little relationship to the purpose of the design, leading to a serious miscalibration of trust.

Although trust depends heavily on the interactions between an operator and the automation, the team and organizational structure within which they function may have an important effect on the diffusion of trust among co-workers. Communication with co-workers augments direct interaction with the automation and may have a strong influence on trust in the automation. A model of trust in automation and evolution of trust in multiperson groups that share responsibility for managing automation showed substantial influence of sharing automation-related information on trust and cooperation of other team members (Gao and Lee, 2006). In this situation, sharing information regarding the performance of the automation not only develops appropriate trust in the automation but also develops appropriate trust in team members who also manage the automation.

One of the ironies of automation is that operators often express a desire for simple and reliable automation, but want the automation to aid them with their most complex tasks (Tenney et al., 1998). Similarly, a highly sensitive warning system that results in many warnings can undermine trust because operators feel that the warnings fail to reflect the danger of the situation accurately (Gupta et al., 2002). These results suggest that understandable and reliable performance on easy tasks may not leave operators willing to rely on the automation to handle more difficult situations. Poor performance of automation on easy tasks severely undermines trust and reliance (Madhavan et al., 2006). Designing automation to promote appropriate trust may help resolve these conflicts. Ideally, trust in automation guides reliance when the complexity of a system makes complete understanding impractical and when the situation demands adaptive behavior (Lee and See, 2004). However, how to design automation to promote appropriate trust, particularly for complex automation that cannot be fully understood by the operator, is a substantial challenge.

One challenge to designing for appropriate trust is that trust has a strong emotional component and may respond to influences that would not be considered in the traditional information processing model that often underlies automation design. As an example, passenger attitudes towards an automated pilot depended on ticket price, suggesting price is used to infer quality. More importantly, inducing positive affect led to higher ratings, providing evidence that ratings of trust are strongly influenced by feelings (Hughes, et al., 2009). These results confirm a more general finding that emotion strongly influences attention, judgment and

decision making with respect to automation interactions (Lee, 2006).

*Job satisfaction and health* often depend on automation in unexpected ways, in part because automation changes the relationship between operators and managers and between operators and work. The issues noted above have addressed primarily the direct performance problems associated with automation. The issue of job satisfaction goes well beyond performance to consider the morale and moral implications of a worker whose job is being changed by automation. Automation that is introduced merely because it increases the profit of the company may not necessarily be well received. Automation often has the effect of de-skilling a job, suddenly making obsolete skills that operators worked for years to perfect. Properly implemented, automation should re-skill workers and enable them to leverage their old skills into new ones that are extended by the automation. Many operators are highly skilled and proud of their craft; automation can thus either empower or demoralize them (Zuboff, 1988). Unhappy operators may fail to capitalize on the potential of an automated system or may even actively sabotage the automation, similar to what the Luddites did.

Automation can also change the relationship to work, increasing demands and decreasing decision latitude. Such an environment can undermine worker health, leading to problems ranging from increased heart disease to increased incidents of depression (Vicente, 1999). However, if automation extends the capability of the operator, it can enhance both satisfaction and health if operators are given sufficient decision latitude. As an example, night shift operators had greater decision latitude than that of day shift operators who worked under the eye of the managers. The night shift operators used this latitude to learn how to manage the automation more effectively (Zuboff, 1988). These effects demonstrate the need to consider the management and implementation of the automation.

### 2.4 Interaction between Automation Problems

Although described independently, the problems of automation often reflect an interacting and dynamic process. One problem may lead to another. Figure 2 summarizes the general problems with automation and identifies some of the important interactions. In many of these relationships, positive feedback reinforces the problem, creating vicious cycles that exacerbate the difficulty. As an example, inadequate training and skill loss may lead the operator to disengage from the monitoring task. This, in turn, will exacerbate the out-of-the-loop unfamiliarity, which will further undermine the operator's skills, and so on. A similar dynamic exists between clumsy automation and automation-induced errors. Clumsy automation produces workload peaks, which increase the chance of mode and configuration errors. Recovering from these errors can further increase workload, and so on. Designing and implementing automation without regard for human capabilities and defining the human role as a by-product has been referred to as automation abuse (Parasuraman and Riley,
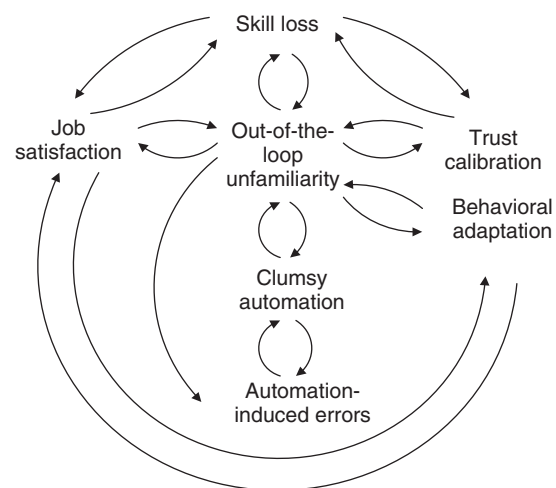


**Figure 2** Interactions among the problems with automation.

1997) and is likely to initiate the negative dynamics shown in Figure 2.

## 3 TYPES OF AUTOMATION

The first step in minimizing the problems and maximizing the benefits of automation is to clarify what is meant by the term *automation*. Automation is not a homogeneous technology. Instead, there are many types of automation and each poses different design challenges. Automation can highlight, alert, filter, interpret, decide, and act for the operator. It can assume different degrees of control and can operate over time scales that range from milliseconds to months. The type of automation, its limits, the operating environment, and human characteristics interact to produce the problems just discussed. Descriptions of automation from different perspectives can reveal the implications of automation for system performance. One such description considers automation in terms of the four stages of human information processing and levels of automation (Parasuraman et al., 2000). Another description considers popular metaphors for automation: tools, prostheses, and agents. Finally, automation can be considered in terms of the scope of the tasks it supports: strategic, tactical, and operational. Any such low-dimensional description of a high-dimensional space will certainly fail to capture important distinctions; nevertheless, these perspectives can make meaningful distinctions that can support design decisions.

### 3.1 Information-Processing Stages and Levels of Automation

If automation is considered as technology that replaces the human in performing a function, it is then reasonable to describe automation in terms of the information-processing functions of the person. Although imperfect,

the information process model of human cognition provides a useful engineering approximation that has been widely applied to system design (Broadbent, 1958; Rasmussen, 1986). The basic information-processing functions—information acquisition, information analysis, action selection, and action implementation—provide simple distinctions that can describe human and automation functions in a common language. A different type of automation corresponds to each stage of information processing. For each of these four functions, different degrees of automation are possible, ranging from full automation to manual control (Sheridan and Verplank, 1978). Information-processing stages and the degree of automation combine to describe a wide array of automation in a way that can guide automation design (Parasuraman et al., 2000).

*Information acquisition automation* refers to technology that complements the process of human attention. Such automation highlights targets (Yeh and Wickens, 2001; Dzindolet et al., 2002), provides alerts and warnings (Bliss, 1997; Bliss and Acton, 2003), and organizes, prioritizes, and filters information. Highlighting targets exemplifies a low degree of information acquisition automation because it preserves the underlying data and allows operators to guide their attention to the information they believe to be most critical. Filtering exemplifies a high degree of automation, and operators are forced to attend to the information the automation deems relevant. *Information analysis* refers to technology that supplants perception and working memory in the interpretation of a situation. Such automation supports situation assessment and diagnosis. As an example, critiquing a diagnosis generated by the operator represents a low degree of automation, whereas automation that provides a single diagnosis represents a high degree of automation. *Action selection automation* refers to technology that combines information in order to make decisions for the operator. Unlike information acquisition and analysis, action selection automation suggests or decides on actions using assumptions about the state of the world and the costs and values of the possible options (Parasuraman et al., 2000). Providing the operator with a list of suggested options represents a relatively low level of action selection automation. In contrast, automation that commands the operator to respond, as in the verbal "pull up, pull up" command of the ground proximity warning system, represents a high level of action selection automation. *Action implementation automation* supplants the operators' activity in executing a response. Olson and Sarter (2001) describe two degrees of action implementation automation *management by consent*, in which the automation acts only with the consent of the operator, and a greater degree of automation *management by exception*, in which automation initiates activities autonomously.

Each of these four stages of automation combines with the degree of automation to describe how the technology supplants the operator's role in perceiving and responding to the environment. Figure 3 shows two hypothetical systems. System B replaces the operator to a relatively high degree for all information-processing stages. In contrast, system A represents a generally
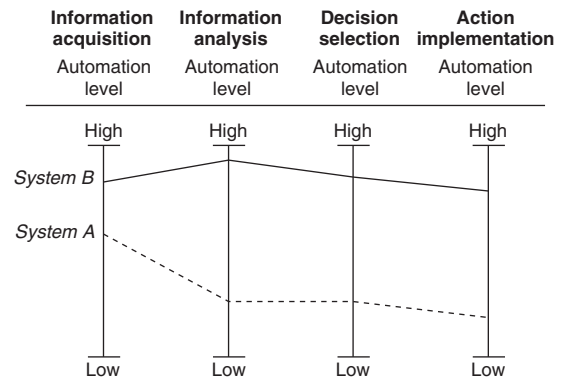


**Figure 3** Two examples of automation defined by a profile of the degree of automation over the four information-processing stages. (From Parasuraman et al., 2000. Copyright © IEEE 2000.)

lower level of automation, with only a moderate degree of automation in the information acquisition stage (Parasuraman et al., 2000).

## 3.2 Tool, Prostheses, and Agents

As automation becomes more complex, considering it simply as a replacement for the information-processing functions of the operator may not differentiate adequately between important types of automation. In many situations, automation is not merely a system that operators engage and disengage. Often, automation consists of a complex array of modes and levels that operators must manage. Interacting with the automation involves coordinating multiple goals and strategies to select a mode of operation that fits the situation (Olson and Sarter, 2000). The simple distinction of engaging manual and automatic control does not capture the complexity of many types of automation. Important design issues emerge as automation evolves from a tool the operator uses to act on the environment to a prosthesis that replaces a human ability to an agent that acts on behalf of the operator. Increasing capacity of automation makes the metaphor of automation as a team member increasingly apt (Klein et al., 2004). The metaphors of automation as a tool, prosthesis, and agent provide complementary perspectives to the information-processing metaphor of automation.

Automation, considered as a *cognitive tool,* extends and complements human capabilities. According to the tool metaphor of automation, operators work directly on the environment, but automation augments their interactions. Just as a hammer augments human action in physical tasks, automation can augment operators in cognitive tasks (Woods, 1987). The benefit of automation as a tool is that its influence is clear and its failures are obvious. An example of automation as a tool that augments human capabilities is a haptic gas pedal, which increases the resistance as the driver approaches a car ahead. This contrasts with adaptive cruise control that automates car following or collision warnings that

only alert the driver when he or she gets too close to the car in front. The continuous feedback of the haptic pedal provides the driver with a useful tool that improves car-following performance (Mulder et al., 2008)

Automation, considered as a *cognitive prosthesis,* acts to replace human function with a more capable computer version. Often, designers adopt this approach in an attempt to enhance system performance or safety by eliminating human error. A cognitive prosthesis eliminates a variable or error-prone aspect of human behavior and replaces it with a consistent computer-based process. The cost of this approach is lost flexibility and reduced ability to adapt to unforeseen situations (Roth et al., 1988). For these reasons, the cognitive prosthesis approach is most appropriate for routine, low-risk situations where decision consistency is more important than adapting to unusual situations. Automation that must accommodate unusual circumstances should adapt a cognitive tool perspective that complements rather than replaces human decision making.

Automation considered as an *agent* acts as a semiautonomous partner with the operator. According to the agent metaphor, the operator no longer acts directly on the environment but acts through an intermediary agent (Lewis, 1998) or intelligent associate (Jones and Jacobs, 2000). As an agent, automation initiates actions that are not in direct response to operators' commands. The authority, autonomy, and complexity of many advanced automated systems make them seem like intentional agents to operators, even if the designers had not intended to adopt this metaphor (Sarter and Woods, 1997). This autonomy and authority can lead to instances of poor coupling and coordination breakdowns because the agents fail to communicate their intentions (Sarter and Woods, 2000; Hoc, 2001). One of the greatest challenges with automated agents is that of mutual intelligibility. Instructing the agent to perform even simple tasks can be onerous, but agents that try to infer operators' intent and act autonomously can surprise operators, who might lack mental models of agent behavior. One approach to improve operator–agent cooperation is for the agents to learn and adapt to the characteristics of the operator through a process of remembering what they have been told to do in similar situations (Bocionek, 1995). After the agent completes a task, it can be equally challenging to make the results meaningful to the operator (Lewis, 1998). Because of these characteristics, agents are most useful for highly repetitive and simple activities, where the cost of failure is limited. In high-risk situations, constructing effective management strategies and providing feedback to clarify agent intent and communicate behavior become critical (Olson and Sarter, 2000; Sarter, 2000).

The differences between automation as a tool, prosthesis, and agent reflect a shift in the locus of control. With a tool, the operator firmly maintains control, but with an agent, the locus of control is more ambiguous and may pass back and forth between the operator and the automation. Ambiguity in the locus of control introduces important considerations regarding inferred intent and the dynamic coordination of actions (Woods, 1994).

The metaphors of tool, prosthesis, and agent complement the information-processing description of automation in important ways. The information-processing metaphor emphasizes the idea that automation replaces the person in performing a function but that function and system remain unchanged. Other metaphors, such as that of automation as an agent, emphasize the far-reaching changes that automation may induce. Rarely is automation a simple replacement of the human, rather, as Woods (1994, p. 4) describes, "technology change produces a complex set of effects. In other words, automation is a wrapped package—a package that consists of changes on many different dimensions bundled together as a hardware/software system." Just as the information-processing metaphor of automation leverages a long history of experimental psychology research, the agent metaphor may leverage recent developments in distributed cognition and team effectiveness (Seifert and Hutchins, 1992; Hutchins, 1995). Such a shift may lead to a change in the boundaries that define the unit of analysis, from one centered on a single operator and a single element of automation to one that considers multioperator, multiautomation interactions (Hollan et al., 2000; Gao and Lee, 2004).

### 3.3 Multilevel Control

The scope of automation varies dramatically, from decision support systems that guide corporate strategies over months and years to antilock brake systems that modulate brake pressure over milliseconds. Substantially different human limitations govern operator interaction with automation at these extremes. A three-level structure that has been used to describe driver behavior seems appropriate for discussing the more general issue of human–automation coordination (Michon, 1985; Ranney, 1994). Figure 4 shows three levels of control that provide a framework for considering issues of coordination and communication of intent. Each level of the figure defines a different level of control that could be supported by a different type of automation. *Strategic automation* concerns balancing values and costs as well as defining goals; *tactical automation*, on the other hand, involves priorities and coordination. Finally, *operational automation* has to do with perceptual cues and motor response.

The bottom of Figure 4 shows operational automation, which governs system behavior over the span of approximately 0.5–5 s. Automation at this level concerns the moment-to-moment control of dynamic processes. An example in the driving domain is ACC, which controls the speed of the car and its distance to the vehicle ahead. The middle of Figure 4 shows tactical automation, which governs system response over a time span of seconds to minutes. In driving, this automation would include route guidance systems that notify drivers of upcoming turns. At the top of the figure is strategic automation, which governs behavior from minutes to days. In driving, this automation helps drivers to select routes and plan trips.

The multilevel control perspective shown in Figure 4 identifies design considerations for the different types of automation and the interaction between different
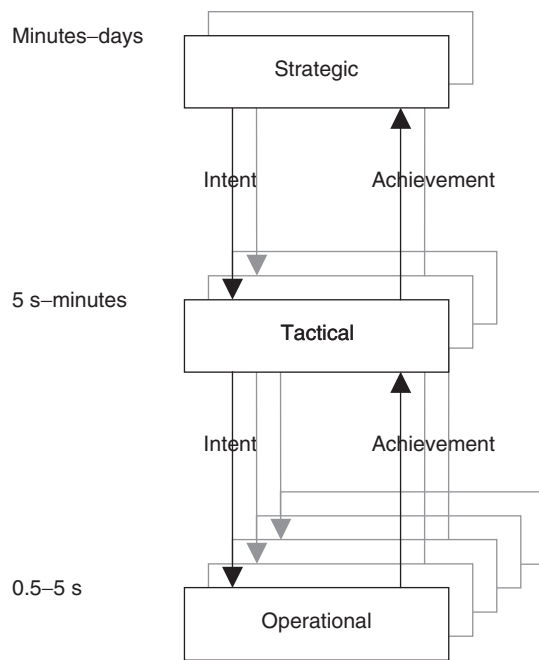
**Figure 4** Strategic, tactical, and operational automation describe types of automation that reveal important coordination and feedback requirements.

elements of automation. First, automation at one level may have unanticipated effects on behavior at another level. For example, automatic control at the lower level might lead people to adopt different behaviors at a higher level, such as when ACC reduces the attention needed in routine car following at the operational level and influences decisions at the tactical level, such as deciding to engage in a cell phone conversation. Second, time constants have a critical effect on monitoring and control behavior. Detection of low-frequency events requires sustained monitoring best suited to time scales at the operational level (0.5–5 s), but such events often occur on a time scale that is orders of magnitude greater. At the other extreme are systems that demand responses on a time scale so short that it exceeds human capabilities. For these systems, automation may need to assume final authority for actions (Moray et al., 2000; Inagaki, 2003). The three-level structure highlights qualitative differences in the time constants of system control that automation design should consider (Hoc, 1993). Third, this perspective highlights the critical issue of communicating intent and the achievement of intent. Figure 4 highlights this requirement because automation at the operational level must be coordinated to achieve the intent developed at the tactical level. Adequate performance of an element of automation at the operational level does not guarantee success at the tactical level unless it is coordinated properly. Automation at one level of control must be managed to minimize interference between agents that might otherwise jeopardize

achieving common tasks at another level of control (Hoc, 2001). Finally, Figure 4 points to the need to consider what some have termed *macrocognition* (Klein et al., 2003). Macrocognitive processes include situation assessment, planning, and coordination. Typical laboratory studies of automation have focused on *microcognition*, associated with operational and tactical levels; however, many critical problems with automation lie at the strategic level and in the interaction between the strategic and tactical levels.

The term *automation* represents a broad array of technology, and no one dimension or framework will capture the many factors that contribute to the problems often encountered with its implementation. Metaphors of information-processing systems, tools, prostheses, agents, and multilevel control all provide complementary perspectives on the nature of automation and how it influences operator performance. These perspectives are not mutually exclusive. A given instance of automation could be described using any or all of the three perspectives. Each provides a different description to guide automation design. Similarly, each perspective provides a partial and distorted description of the true complexities of automation. Although each perspective is limited, each can enhance our understanding of human–automation interaction.

## 4 STRATEGIES TO ENHANCE HUMAN–AUTOMATION INTERACTION

Defining the problems encountered with automation should instill caution in those who believe that automation can enhance system performance and safety by replacing the human operator. The perspectives on the nature and types of automation reveal the complexity of automation. Neither caution nor perspective, however, is sufficient to develop successful automation. In this section we describe specific strategies for designing effective automation, which include:

- Fitts's list and function allocation
- Dynamic function allocation (adaptable and adaptive automation)
- Matching automation to human performance characteristics
- Representation aiding and multimodal feedback
- Matching automation to mental models
- Formal automation analysis techniques

### 4.1 Fitts's List and Function Allocation

One approach to automation is to assess each function and determine whether a human or automation would perform that function better (Kantowitz and Sorkin, 1987; Sharit, 2003). Functions better performed by automation are automated and the operator remains responsible for the rest and for recovering during the periodic failures of the automation. Fitts's list provides a heuristic basis for determining the relative performance of humans and automation for each function (Fitts,

**Table 1  Fitts's List: Relative Strengths of Automation and Humans for the Four Information-Processing Stages**

| Information-Processing Stage | Humans Are Better At: | Automation Is Better At: |
| --- | --- | --- |
| Information acquisition | Detecting small amounts of visual, auditory, or chemical signals | Monitoring processes |
| | Detecting a wide range of stimuli | Detecting signals beyond human capability |
| Information analysis | Perceiving patterns and making generalizations | Ignoring extraneous factors and making quantitative assessments |
| | Exercising judgment | Consistently applying precise criteria |
| | Recalling related information and developing innovative associations between items | Storing information for long periods and recalling specific parts and exact reproduction |
| Action selection | Improvising and using flexible procedures | Repeating the same procedure in precisely the same manner many times |
| | Reasoning inductively and correcting errors | Reasoning deductively |
| Action implementation | Switching between actions as demanded by the situation | Performing many complex operations at once |
| | Adjusting dynamically to a wide range of conditions | Responding quickly and precisely |

1951). Table 1 shows a revised Fitts list for the stages of automation identified earlier. The relative capability of the automation and human depend on the stage of automation (Sheridan, 2000).

Using the heuristics in Table 1 to determine which functions should be automated mitigates skill loss and lack of training by clearly identifying the human role in a system. This approach also enhances job satisfaction by designing a role for the operator that is compatible with human capabilities. Ideally, the function allocation process should not focus on what functions should be allocated to the automation or to the human but should identify how the human and the automation can complement each other in jointly satisfying the functions required for system success (Hollnagel and Bye, 2000).

Applying the information in Table 1 to determine an appropriate allocation of function has, however, substantial weaknesses. One weakness is that there are many interconnections between functions. Any description of functions is a somewhat arbitrary decomposition of activities that masks complex interdependencies. As a consequence, automating functions as if they were independent has the tendency to fractionate the operator's role, leaving the operator with only those tasks too difficult to automate (Bainbridge, 1983). Automation must be designed to support the job of the operator as an integrated whole. Another weakness with this approach is the situation dependence of the automation and human performance. The same function may require improvisation in some circumstances and precise application of a fixed response in others. Another weakness is that the work and the automation coevolve, with the automation making unanticipated work practices possible and the work leading to unanticipated applications of the automation (Dearden et al., 2000). A final weakness with function allocation using Fitts's list is the diminishing list of situations in which human abilities exceed those of the automation. Strict adherence to the application

of Fitts's list to allocate functions between people and machines has been widely recognized as problematic (Parasuraman et al., 2000; Sheridan, 2000).

Although imperfect, Table 1 contains some general considerations that can improve design. People tend to be effective with complete patterns and less so with highly precise repetition. Human memory tends to organize large amounts of related information in a network of associations that can support effective judgments requiring the consideration of many factors. People also adapt, improvise, and accommodate unexpected variability. For these reasons it is important to leave the "big picture" to the human and the details to the automation (Sheridan, 2002).

## 4.2  Dynamic Function Allocation: Adaptable and Adaptive Automation

Using Fitts's list or some other method to allocate functions between humans and automation results in static function allocation in which the division of labor is fixed by the designer. Functions once performed by the human are now performed by automation. Static allocation of function contrasts with dynamic allocation of function, in which adaptable and adaptive automation makes it possible to adjust the division of labor between the human and the automation over time (Scerbo, 1996; Sarter and Woods, 1997). Dynamic allocation of function addresses the need to adjust the degree and type of automation according to individual differences, the state of the operator, and the state of the system. Adaptable and adaptive automation is often preferable to automation that is fixed and rigid.

*Adaptable automation* is that which the operator can engage or disengage as needed. The operator adapts the level and type of automation to the situation. Giving operators the option of manual or automatic control can be more effective than making available only automatic or only manual control (Harris et al., 1995).

More generally, adaptable automation gives operators additional degrees of freedom needed to accommodate unanticipated events (Hoc, 2000). The decision to rely on the automation or to intervene with manual control depends on many factors, including perceived risk, workload, trust, and self-confidence (Riley, 1989, 1994; Lee and Moray, 1994). To the extent that operators trust the automation appropriately and have appropriate self-confidence, they tend to rely on the automation appropriately and avoid some of the out-of-the-loop unfamiliarity problems. Allowing operators to transition easily between automatic and manual control can also mitigate clumsy automation. On the other hand, one of the critical deficiencies of adaptable automation is that it gives the operator the additional tasks of engaging and disengaging the automation. If the effort associated with these tasks is great, adaptable automation can increase the workload of demanding situations and thus become an example of clumsy automation.

*Adaptive automation* goes a step further than adaptable automation by automatically adjusting the level of automation based on the operator's performance, the operator's state, or the task situation (Rouse, 1988; Byrne and Parasuraman, 1996). Often, adaptive automation focuses on increasing the level of automation when either the operator's workload increases or the operator's capacity decreases. One way to estimate operator workload is through physiological measures such as heart rate and electroencephalography (EEG) signals (Byrne and Parasuraman, 1996). For example, it is possible to moderate an operator's workload by using closed-loop control algorithms to adjust the level of automation according to the operator's EEG signal (Prinzel et al., 2000). Other estimates of workload depend on models that relate the task situation to expected cognitive load and operator performance. For example, by combining operator performance and task variables it is possible to engage automation and mitigate predictable workload increases (Scallen and Hancock, 2001). Most promising is an approach that combines data from all three sources along with model-based predictions of workload. By engaging higher levels of automation during periods of high workload, adaptive automation promises to solve some of the problems of clumsy automation.

Alleviating overload is often the motive behind the development of adaptive automation. It may be equally important, however, to consider how it can mitigate problems of underload. Both underload and overload stress an operator's ability to respond (Hancock and Warm, 1989), and automation that returns tasks to the operator during underload situations may place operators in a less stressful situation. Similarly, operators who monitor reliable automation for long periods become surprisingly inefficient at detecting automation failures. Adaptive automation can mitigate this automation-induced complacency by returning manual control periodically to the operator (Parasuraman et al., 1996). Adaptive automation that used EEG signals led to higher levels of situation awareness and lower levels of workload compared to adaptable automation that

required people to manage the users to engage the automation (Bailey et al., 2006).

Adaptive automation is a sort of meta-automation that can suffer from some of the same problems of automation if implemented improperly. Adaptive automation relieves the operator of the task of engaging and disengaging the automation, but it imposes the additional task of monitoring the adaptive automation, which can also increase workload (Kaber et al., 2001). In addition, adaptive automation faces challenging measurement and control problems. Adaptive automation depends on a precise measure of operator state, which can include physiological variables. If the time constant of these variables is longer than the time constant of the demands of the environment, automation will not adapt quickly enough. Even if operator state can be measured in a precise and timely manner, developing control algorithms that relate the operator state to an appropriate level of automation is difficult. Many of the limits of applying the Fitts list to static allocation of function also make dynamic allocation of function a challenge. Finally, even if an appropriate algorithm for adjusting the automation dynamically can be defined, the operator might respond in unexpected ways. For example, operators may manipulate their physiological state to influence the automation (Byrne and Parasuraman, 1996). Most important, operators may not understand the adaptive automation and so will view the system as behaving erratically. Such dynamic changes also introduce interface inconsistencies and increase the potential for mode errors.

## 4.3 Matching Automation to Human Performance Characteristics

Another approach to automation design considers how operators respond to different types of imperfect automation (Parasuraman et al., 2000). How well an operator is able to recognize and recover from automation failures often governs overall system performance. As a consequence, an important approach to automation design is to consider how human performance characteristics interact with the type of automation. The objective of this design approach is to minimize the tension that arises from mismatches between human performance characteristics and the type of automation (Sharit, 2003). A specific example of this approach considers the levels of automation and types of automation as defined by the stages of information processing. Primary considerations for automation design include workload, situation awareness, complacency, and skill maintenance (Parasuraman et al., 2000). These considerations do not specify a universally applicable degree of automation for each information-processing stage. Instead, appropriate automation design depends on the reliability of the automation and the consequences of failure as well as on technical and economic considerations (Parasuraman et al., 2000). In the context of air traffic control, human performance characteristics argue for the following upper bounds on the level of automation: information acquisition (high), information interpretation (high), action selection (medium), and action implementation (medium).

As an example, displays that indicate the status of the system (information interpretation automation) are preferable to those that advise the operator on how to respond (action selection automation) (Crocoll and Coury, 1990). Specifically, alerts regarding hazardous road conditions presented as a command (e.g., merge left) led to more dangerous lane changes compared to the same information presented as a notification (e.g., road construction in right lane) (Lee et al., 1999). Similar findings for a decision aid to help pilots make decisions regarding the dangers of aircraft icing suggest that status displays are preferable to command displays in high-risk domains where the automation is imperfect (e.g., space flight, medicine, and process control) (Sarter and Schroeder, 2001). Action implementation automation can be helpful when reliable but dangerously compelling when unreliable. Operators benefit more from action implementation automation than from action selection automation, but only when the automation performs reliably (Endsley and Kaber, 1999). Although a greater degree of automation enhances performance and reduces workload during routine situation, it can also reduce situation awareness and undermine the ability to respond—when the automation fails, operators perform better with lower levels of automation (Kaber et al., 2000). In addition to the reliability of the automation, time pressure influences the benefit of a greater degree of automation. Pilots preferred management by consent, a relatively low level of automation; however, during periods of high time pressure and high workload, they preferred management by exception, a higher level of automation (Olson and Sarter, 2000).

Expert systems represent a high degree of decision automation that has frequently failed to meet expectations. Typically, an expert system acts as a prosthesis, supposedly replacing flawed and inconsistent human reasoning with more precise computer algorithms. Unfortunately, the level of automation associated with such an approach often conflicts with the range of situations the automation must face: The system gives the wrong answer when confronted with cases for which the automation is not fully competent. In addition, the operator typically plays a passive role such as entering data or assessing automation decisions, which leads to brittle failures (Roth et al., 1988).

A lower degree of automation, which places the automation in the role of critiquing the operator, has met with much more success. In critiquing, the computer presents alternative interpretations, hypotheses, or choices that complement those of the operator (Guerlain et al., 1999; Sniezek et al., 2002). A specific example is a decision support system for blood typing (Guerlain et al., 1999). Rather than using the expert system as a cognitive prosthesis to identify blood types, the critiquing approach suggests alternative hypotheses regarding possible interpretations of the data. In cases where the automation was fully competent, the operators made correct diagnoses 100% of the time, compared to 33–63% for those without the critiquing system. In cases where the critiquing system was not fully competent, performance degraded gracefully and operators still correctly diagnosed 32% more cases than those without

the critiquing system. In situations where the automation is imperfect or the cost of failure is high, a lower level of automation, such as that used in the critiquing approach, is less likely to induce errors. Although much of the benefit of a critiquing system stems from the lower degree of automation and the greater involvement of the operator in the decision process, representation aiding plays an important role in supporting efficient operator–automation interaction.

## 4.4  Representation Aiding and Multimodal Feedback

Even if the type of automation is well matched to the task situation and human capabilities, inadequate feedback can undermine human–automation interaction. Inadequate feedback underlies many of the problems with automation from developing appropriate trust and clumsy automation to the out-of-the-loop phenomenon (Norman, 1990). However, providing sufficient feedback without overwhelming the operator is a critical design challenge. Poorly presented or excessive feedback can increase operator workload and undermine the benefits of the automation (Entin et al., 1996). In addition, without the proper context, abstraction, and integration, information regarding the behavior of complex automation may not be understandable. Representation aiding and multimodal feedback are two approaches that can help people understand how the automation works and how it is performing.

*Representation aiding* capitalizes on the power of visual perception to convey complex dynamic relationships. For example, graphical representations for pilots can augment the traditional airspeed indicator with target airspeeds and acceleration indicators. Integrating this information into a traditional flight instrument allows pilots to assimilate automation-related information with little extra effort (Hollan et al., 2000). Using a display that combines pitch, roll, altitude, airspeed, and heading can directly specify task-relevant information such as what is "too low" (Flach, 1999). Integrating automation-related information with traditional displays and combining low-level data into meaningful information are two important ways to enhance feedback without overwhelming the operator.

In regard to process control, Guerlain et al. (2002) identified three specific strategies for visual representation of complex process control algorithms. First, create visual forms whose emergent features correspond to higher order relationships. Emergent features are salient symmetries or patterns that depend on the interaction of the individual data elements. A simple emergent feature is *parallelism*, which can occur with a pair of lines. Higher order relationships are combinations of the individual data elements that govern system behavior. The boiling point of water is a higher order relationship that depends on temperature and pressure. Second, use appropriate visual features to represent the dimensional properties of the data. For example, magnitude is a dimensional property that should be displayed using position or size on a visual display, not color or texture. Third, place data in a meaningful context. The meaningful context for any variable depends on what

| CV DETAIL | | | | | |
|---|---|---|---|---|---|
| RX/REGEN CTL | ON | OFF | WARM | | OPTIMIZING |

TAG          25ATCV01
DESC         DCO YIELD
SOURCE       25ATCVO1.PV                                LINEAR OBJ COEF        -1.00
                                                        QUAD OBJ COEF           0.00
PV VALUE     579.3     STATUS       GOOD                DESIRED CV VAL          0.00
PRED VAL     579.36                                     SCALING FACTOR          0.329
FUTURE       579.38    SP.LIM TRACKS PV    YES  NO
SS VALUE     581.36    UPDATE FREQUENCY     =    <      CV LO ERROR WEIGHT      1.00
                       CRITICAL CV         YES  NO      CV HI ERROR WEIGHT      1.00
SETPOINT               CONTROL THIS CV     YES  NO
                                                        PERFORMANCE RATIO       1.00
LO LIMIT     400.00    # OF BAD READS ALLOWED    5      CLS LOOP RESP INT      54.800
ACTIVE       400.00                                     FF TO FB PERF RATIO     0.50
                       LO LIMIT RAMP RATE   10.000
HI LIMIT     600.00    HI LIMIT RAMP RATE   10.000      SETPOINT GAP            0.00
ACTIVE       600.00    UNBIASED MODEL PV    379.35      NUMBER OF BLOCKS       10.0

| APPLCN MENU | PROCESS DISPLAY | CV DISPLAY | MV DISPLAY | DV DISPLAY | STATUS MESG | MV TUNING | CV TUNING | GAIN/ DELAY | TREND DISPLAY |
|---|---|---|---|---|---|---|---|---|---|

(a)
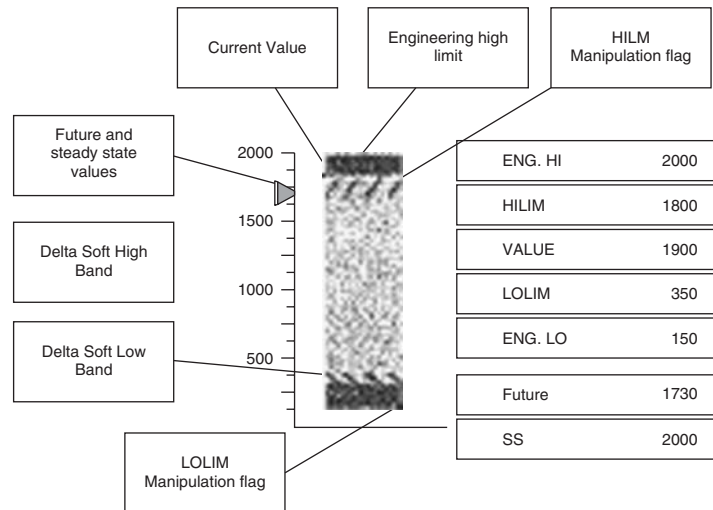


(b)

**Figure 5** (a) Comparison of a traditional interface for automation; (b) example of representation to support operator understanding of automation.

comparisons need to be made. For automation, this includes the allowable ranges relative to the current control variable setting and the output relative to the desired level. Figure 5 shows some of the principles of representation aiding—use analog rather than digital or text, provide meaningfully integrated rather than raw data, and provide a context to support visual rather than mental comparisons.

Representation aiding makes it more likely that operators will trust automation more appropriately. However, trust also depends on more subtle elements of the interface (Lee and See, 2004). In many cases, trust and credibility depend on surface features of the interface that have no obvious link to the true capabilities of the system (Briggs et al., 1998; Tseng and Fogg, 1999). For example, in an online survey of over 1400 people, Fogg et al. (2001b) found that for websites credibility depends heavily on "real-world feel," which is defined by factors such as response speed, a physical address, and photos of the organization. Similarly, a formal photograph of the author enhanced trustworthiness of a research article, whereas an informal photograph decreased trust (Fogg et al., 2001a). These results show that trust tends to increase when information is displayed in a way that provides concrete details that are consistent and clearly organized.

A similar pattern of results appears in studies of automation for target detection. Increasing image

realism increased trust and led to greater reliance of the cueing information (Yeh and Wickens, 2001). Similarly, the tendency of pilots to follow the advice of the system blindly increased when the aid included detailed pictures (Ockerman, 1999). Just as highly realistic images can increase trust, degraded imagery can decrease trust, as was shown in a target cueing situation (MacMillan et al., 1994). Adjusting image quality and adding information to the interface regarding the capability of the automation can promote appropriate trust. In a signal detection task, the reliability of the sources was coded with different levels of luminance, leading participants to weigh reliable sources more than unreliable ones (Montgomery and Sorkin, 1996). These results suggest that the particular interface form can increase the level of trust, particularly the emphasis on concrete realistic representations.

Trust and reliance can also be enhanced with information that conveys the performance and expected value of automation. Such information can address appraisal errors—failures to properly judge the benefit of the automation. In one study, performance feedback reduced disuse rates from 84 to 55% (Beck et al., 2007). This result suggests that even when operators understand the expected value of the automation, they persist in disuse, indicating a John Henry effect in the form of intent errors. Intent errors were mitigated with scenario training that conveyed the appropriate thought process for interpreting automation suggestions. When combined with feedback, scenario training reduced disuse to 29% (Beck et al., 2007). The degree of personal investment operators have in performing the task has a strong influence on the prevalence of intent errors and, consequently, the importance of scenario training (Beck et al. 2009). Such

Representation aiding tends to focus on interfaces that require focal as opposed to peripheral vision. Operators already face substantial demands on focal vision, and presenting automation-related information in that channel may overwhelm the operator. *Multimodal feedback* provides operators with information through haptic, tactile, auditory, and peripheral vision to avoid overwhelming the operator. Haptic feedback has proved more effective in alerting pilots to mode changes in cockpit automation compared to visual cues (Sklar and Sarter, 1999). Pilots receiving visual alerts detected 83% of the mode changes; those with haptic warnings detected 100% of the mode changes. Importantly, the haptic warnings did not interfere with the performance of concurrent visual tasks. Similarly, peripheral visual cues also helped pilots detect uncommanded mode transitions and did not interfere with concurrent visual tasks any more than did currently available automation feedback (Nikolic and Sarter, 2001). Haptic warnings may also be less annoying and acceptable compared to auditory warnings (Lee et al., 2004). Although promising, multimodal interfaces lack the resolution of visual interfaces, making it difficult to convey complex relationships and detailed information.

### 4.5 Matching Automation to Mental Models

The complexity of automation sometimes makes it difficult to convey its behavior using representation

aiding or multiple-modal feedback. Sometimes a more effective strategy is to simplify the automation (Riley, 2001) or to match its algorithms to the operators' mental model (Goodrich and Boer, 2003). This is particularly true when a technology-centered approach to automation design has created an overly complex array of modes and features. The out-of-the-loop unfamiliarity problems result partially from the difficulties that operators have in generating correct expectations for the counterintuitive behavior of complex automation. Automation designed to perform in a manner consistent with operators' preferences and expectations can make it easier for operators to recognize failures and intervene.

Adaptive cruise control is a specific example of where matching the mental model of the operator may be quite effective. Because drivers must focus their attention on the roadway, representation aiding could be distracting. Because ACC can apply only moderate levels of braking, drivers must intervene if the car ahead brakes heavily. If drivers must intervene, they must quickly enter the control loop because fractions of a second matter. If the automation behaves in a manner consistent with that of the driver, he or she will be more likely to detect and respond to the operational limits of the automation (Goodrich and Boer, 2003). To design an ACC algorithm consistent with drivers' mental models, driver behavior was partitioned according to perceptually relevant variables of inverse time to collision and time headway. Inverse time to collision ($T_c^{-1}$) is the relative velocity divided by the distance between the vehicles. Time headway ($T_h$) is the distance between vehicles divided by the velocity of the driver's vehicle. These variables define the boundary that separates speed regulation and headway maintenance from active braking associated with collision avoidance. Figure 6 shows this boundary in the space defined by time headway and inverse time
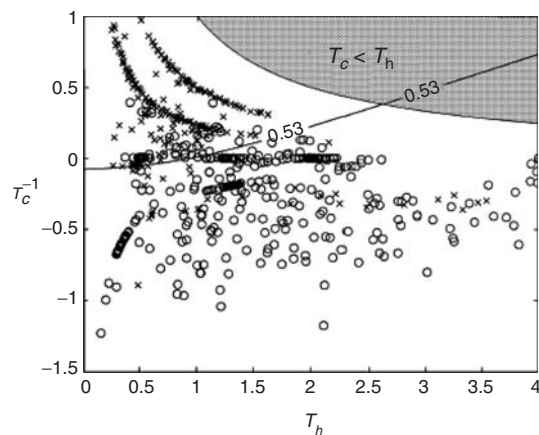


**Figure 6** Driver braking behavior, showing a clear boundary between headway maintenance (○) and collision avoidance (×) that could be used to define operational limits of ACC. (From Goodrich and Boer, 2003. Copyright © IEEE 2002.)

to collision. This boundary provides a template for designing ACC—the ACC should signal the driver to intervene as the driving situation crosses the boundary.

For situations in which the metaphor for automation is an agent, the mental model that people may adopt to understand the automation is that of a human collaborator. If the template for understandable automation is the operator's mental model, an agent should respond as would a human. Specifically, Miller (2002) suggests that computer etiquette may have an important influence on human–automation interaction. Etiquette may influence trust because category membership associated with adherence to a particular etiquette helps people to infer how the automation will perform. Specific rules for automation etiquette adapted from Miller and Funk (2001) include:

- Make many correct interactions for every erroneous interaction.
- Make it very easy to override the automation.
- Do not make the same mistake twice—stop a behavior if corrected by the operator.
- Do not enable interaction features just because they are possible.
- Explain what is being done and why.
- Be able to take instruction.
- Do not assume every operator is the same—be sensitive and adapt to individual, contextual, and cultural differences.
- Be aware of what the operator knows and do not repeat unnecessarily.
- Use multiple modalities to communicate.
- Try not to interrupt.
- Be cute only if it furthers specific interaction goals.

Developing automation etiquette could promote appropriate trust, but it could lead to inappropriate trust if people infer inappropriate category memberships and develop distorted expectations regarding the capability of the automation. Even in simple interactions with technology, people often respond as they would to another person (Reeves and Nass, 1996; Nass and Lee, 2001). If anticipated, this tendency could help operators develop appropriate expectations regarding the behavior of the automation; however, unanticipated anthropomorphism could lead to surprising misunderstandings of the automation.

An important prerequisite for designing automation according to the mental model of the operator is the existence of a consistent mental model. Individual differences may be difficult to accommodate. This is particularly true for automation that acts as an agent, in which a mental model–based design must conform to complex social and cultural expectations. In addition, the mental model must be consistent with the physical constraints of the system if the automation is to work properly (Vicente, 1990). Mental models often contain misconceptions, and transferring these to the automation could be counterproductive or deadly. Even if operators

have a single mental model that is consistent with the system constraints, automation based on a mental model may not achieve the same benefits as automation based on more sophisticated algorithms. In this case, designers must consider the trade-off between the benefits of a complex control algorithm and the costs of a poorly understood system. Representation aiding can mitigate this trade-off.

## 4.6 Formal Automation Analysis Techniques

Effective representation aiding depends on identifying the relevant information needed to understand the behavior of the automation. With complex automation, this can be a substantial challenge. One approach to meeting this challenge is to use formal verification techniques (Leveson, 1995; Degani and Heymann, 2002). Specifically, state machines can define the behavior of the automation and the operator's model. The state machine that defines the operator's model is constructed from the training materials and the information available on the interface. State machines provide a formal modeling language to define mismatches between the operator's model of the automation and the automation. These mismatches cause automation-related errors and surprises to occur.

State machines identify the legal and illegal states defined by the task constraints that the automation and operator must satisfy. When the automation model enters an illegal state and the operator's model does not, the analysis predicts that the associated ambiguity will surprise operators and lead to errors (Degani and Heymann, 2002). Such ambiguities have been discovered in actual aircraft autopilot systems (Degani and Heymann, 2002). Mismatches between the operator and automation models indicate deficiencies in the operator's mental model that should be addressed by changing the automation, training, or interface (Heymann and Degani, 2007). The state machine formalism makes it possible to generate training and interface requirements automatically.

Often, designers overestimate the benefit of automation because of the surprising interactions between the automation, environment, and operator. Formal analysis that considers these interactions in terms of expected-value calculations can reduce the surprise and guide design. In the example of a rear-end collision warning system for cars, a Bayesian approach combined with signal detection theory shows that the posterior probability of a collision situation given a warning is surprisingly low because the base rate of collision situation is so low (Parasuraman et al., 1997). This analysis shows that the selection of a detection threshold should consider the base rate; otherwise, the relatively high rate of false alarms could undermine driver acceptance.

More generally, calculating the expected value of manual and automatic control provides a rigorous means of selecting the best alternative (Sheridan and Parasuraman, 2000). In the simplest case this involves comparing the expected value of the operator and automation response to a binary failure state—a system is either operating normally or it has failed. The expected-value calculation combines the benefits and costs of four general responses to the system: a true

positive, a true negative, a false negative, and a false positive. The expected value of the automation and the expected value of the operator response depend on the costs of being wrong and the benefits of being correct, together with the prior probabilities of the failure and the probabilities of the automation and operator being wrong and correct. If the expected value for automatic control is greater than the expected value for manual control, automation should be implemented. A similar analysis shows that the time-dependent value of automation makes it reasonable to give the automation final authority in some situations, such as in guiding the pilot to make go/no go decisions in aborting a takeoff (Inagaki, 2003). A similar analysis might help designers balance the information-processing demands of feedback regarding automation behavior with the time demands of the situation. Experiments assessing human interaction with automation should consider this calculation in defining experimental conditions, defining the reward structure, and interpreting the participants' behavior (Bettman and Payne, 1990; Payne et al., 1992; Meyer, 2004); otherwise, it is impossible to differentiate automation bias from eutactic behavior.

An expected-value analysis provides a way to formalize the cost–benefit analysis that might otherwise be guided by the qualitative Fitts list heuristics. Although it promises to precisely quantify otherwise ambiguous decisions, estimating the numbers required to support the calculations can be a challenge. The costs and probabilities of rare, catastrophic events are notoriously difficult to estimate. More subtly, operator performance may affect the prior probabilities of events such that good operators experience fewer failures than do poor operators. In this situation, the automation will perform more poorly for better operators (Meyer and Bitan, 2002). Although precise probabilities and values may be difficult or impossible to estimate, such an approach is quite useful even if only relative benefits and costs of the automation and operator can be estimated (Sheridan and Parasuraman, 2000).

Simulation can also guide designers to consider the costs and benefits of automation more thoroughly. A simulation of a supervisory control situation shows that well-adapted operators are sensitive to the costs of engaging and disengaging automation (Kirlik, 1993). This simulation analysis identifies how the time costs of engaging the automation interact with the dynamics of the environment to undermine the value of the automation. A similar analysis argues that designers must make the normative strategy less effortful than competing strategies if operators are to use automation effectively (Todd and Benbasat, 2000). More generally, simulation models that capture the human performance consequences of different levels of automation reliability, and the environmental constraints are needed to support design. For example, a connectionist model of complacency provides a strong theoretical basis that accounts for empirical findings (Farrell and Lewandowsky, 2000). Cognitive architectures such as ACT-R also offer a promising approach to modeling human–automation interaction (Anderson and Libiere, 1998). Although ACT-R may not be able to capture the full complexity

of this interaction, it may provide a useful tool for approximating the costs and benefits of various automation alternatives (Byrne and Kirlik, 2005).

## 5 EMERGING CHALLENGES

Substantial progress has been made regarding how to design automation to support people effectively. However, continuous advances in software and hardware development combined with an ever-expanding range of applications make future problems with automation likely. The following section highlights some of these emerging challenges. The first is the demands of managing a new type of automation, *swarm automation*, in which many semiautonomous agents work together. The second is the implication of automation in large interconnected networks of people and other automated elements, where issues of coordination and competition become critical. Automation in this environment requires considerations beyond those of the typical single operator interacting with one or two elements of automation. The third is the introduction of automation into daily life: specifically, automation in the car. These three examples represent some of the challenges associated with new types of automation, new types of human–automation organizations, and new application domains.

### 5.1 Swarm Automation

*Swarm automation* is an alternative approach to automation that may make it possible to respond to environmental variability while reducing the chance of system failure. These capabilities have important applications in a wide range of domains, including planetary exploration, unmanned aerial vehicle reconnaissance, landmine neutralization, or even data exploration, where hundreds of simple agents might be more effective than a single complex agent. Biology-inspired roboticists provide a specific example of swarm automation. Instead of the traditional approach of relying on one or two larger robots, they employ swarms of insect robots as an alternative (Brooks et al., 1990; Johnson and Bay, 1995). The swarm robot concept assumes that small machines with simple reactive behaviors can perform important functions more reliably and with lower power and mass requirements than can larger robots (Beni and Wang, 1993; Brooks and Flynn, 1993; Fukuda et al., 1998). Typically, the simple programs running on an insect robot are designed to elicit desirable emergent behaviors in the swarm (Sugihara and Suzuki, 1990; Min and Yin, 1998). For example, a large group of small robots might be programmed to search for concentrations of particular mineral deposits by building on the foraging algorithms of honeybees or ants.

In addition to physical examples of swarm automation, swarm automation has potential in searching large complex data sets for useful information. For example, the pervasive issue of data overload and the difficulties associated with effective information retrieval suggest a particularly useful application of swarm automation. Current approaches to searching large complex data sources, such as the Internet, are limited. People are

likely to miss important documents, disregard data that represent a significant departure from initial assumptions, misinterpret data that conflict with an emerging understanding, and disregard more recent data that could revise interpretation (Patterson, 1999). These issues can be summarized as the need to broaden searches to enhance opportunity to discover highly relevant information, promote recognition of unexpected information to avoid premature fixation on a particular viewpoint or hypothesis, and manage data uncertainty to avoid misinterpretation of inaccurate or obsolete data (Woods et al., 1999). These represent important challenges that may require innovative design concepts and significant departures from current tools (Patterson, 1999). Just as swarm automation might help explore physical spaces, it might also help explore information spaces.

Managing swarm automation requires a qualitatively different approach than that of more traditional automation (Lee, 2001). Swarms of bees and ants, in which many simple individuals combine to behave as a single entity, provide some useful insights into the characteristics of swarm behavior and how they might be managed (Bonabeau et al., 1997). A defining characteristic of swarm behavior is that it emerges from parallel interaction between many agents. For example, swarms of bees adjust their foraging behavior to the environment dynamically in a way that does not depend on the performance of any individual. A colony of honeybees functions as a large, diffuse, amoeboid entity that can extend over great distances and simultaneously tap a vast array of food sources (Seeley, 1997). Direct control of this emergent behavior is not possible. Instead, mechanisms influencing individual elements of the swarm indirectly influence swarm behavior. Two particularly important mechanisms are positive feedback and random variation. Positive feedback reinforces existing activities, and random variation generates new activities and encourages adaptation (Resnick, 1991). One way that positive feedback and random variation combine to influence behavior is through *stimergy*, in which communication and control occur through a dynamically evolving structure. Through stimergy, social insects communicate directly through the products of their work (e.g., the bees' honeycomb and the termites' chambers). A specific example of stimergy is the pheromone trail that guides the self-organizing foraging behavior of ants. Stimergy in foraging behavior involves a trade-off of speed of trail establishment and search thoroughness; A trail that is more quickly established will sacrifice the thoroughness of the search. Stimergy represents a powerful alternative to a static set of instructions that specify a sequence of activity. Parallel interaction between many agents, positive feedback, random variation, and stimergy make it possible for many simple individuals to produce complex group behavior (Bonabeau et al., 1997). However, such control mechanisms may be difficult for operators to understand.

The concept of hortatory control describes some of the challenges of controlling swarm automation. Hortatory control applies in situations where the system being controlled retains a high degree of autonomy and operators must exert indirect rather than direct control (Murray and Liu, 1997). Interacting with swarm automation requires people to consider swarm dynamics independent of the individual agents. In these situations it is most useful for the operator to control parameters affecting group rather than individual agents and for the operators to receive feedback about group rather than individual behavior. Swarm automation has great potential to extend human capabilities, but only if a thorough empirical and analytic investigation identifies the display requirements, feasible control mechanisms, and range of swarm dynamics that can be comprehended and controlled by humans.

## 5.2 Management of Complex Networks of Operators and Automation

As automation becomes pervasive, it creates complex networks of increasingly tightly coupled elements. In this situation, the appropriate unit of analysis may shift from a single operator interacting with a single element of automation to that of multiple operators interacting with multiple elements of automation. Important dynamics can only be explained with this more complex unit of analysis. More so than single-operator situations, in these highly coupled systems, poor coordination between operators and inappropriate reliance on automation can degrade the decision-making performance and lead to catastrophes (Woods, 1994). As an example, the largest power grid failure in the nation's history occurred on August 14, 2003. In this failure, the flow of approximately 61,800 MW of electricity was disrupted, leaving 50 million customers from Ohio to New York and parts of Canada without power. An important contribution to this event was a lack of cooperation between two regional electrical grid operators that monitor the same region. These operators manage the flow of the electricity from suppliers to distributors. Poor communication and a failure to exchange detailed information on their operations prevented them from understanding and responding to changes in the power grid. Similar failures occur in supply chains as well as petrochemical processes, where people and automation sometimes fail to coordinate their activities.

Supply chains represent an increasingly important example of multioperator multiautomation. A supply chain is composed of a network of suppliers, transporters, and purchasers who work together, usually as a decentralized virtual company, to convert raw materials into products for end users. The growing popularity of supply chains reflects the general trend of companies to move away from vertical integration, where a single company converts raw materials into products for end users. Many manufacturers increasingly rely on supply chains; a typical U.S. company purchases 55% of the value of its products from other companies (Dyer and Singh, 1998). Efficient supply chains play a critical role in maintaining the economic health of the U.S. economy.

However, supply chains suffer from serious problems that erode their promised benefits. One is the *bullwhip effect*, in which small variations in end-item demand induce large-order oscillations, excess inventory, and backorders (Sterman, 1989). This effect can have enormous consequences on a company's efficiency and value. As an example, news reports of supply chain

glitches associated with the bullwhip effect resulted in abnormal declines of 10.28% in companies' stock price (Hendricks and Singhal, 2003). Automation that forecasts demands can moderate these oscillations (Lee and Whang, 2000; Zhao and Xie, 2002). However, people must trust and rely on that automation, and substantial cooperation between supply chain members must exist to share such information.

Another major problem facing supply chains is the breakdown in cooperation as relationships between members of a supply chain devolve through an escalating series of conflicts that has been termed a *vicious cycle* (Akkermans and van Helden, 2002). Such conflicts can have dramatic negative consequences for a supply chain. For example, a strategic alliance between Office Max and Ryder International Logistics devolved into a legal fight in which Office Max sued Ryder for $21.4 million and then Ryder sued Office Max for $75 million (Handfield and Bechtel, 2002). Beyond the legal costs, these breakdowns can threaten competitiveness and undermine the market value of the company (Dyer and Singh, 1998). Vicious cycles also undermine information sharing, which can exacerbate the bullwhip effect. Even with the substantial benefits of cooperation, supply chains frequently fall into a vicious cycle in which poor cooperation leads to further poor cooperation. Trust between people plays a critical role in developing and sustaining cooperative relationships. People must trust each other to share information, and this trust can be undermined if poorly managed automation of one supply chain member compromises the success of another.

The bullwhip effect and vicious cycles and other supply chain problems reflect the influence of inappropriate actions at the local level that drive dysfunctional network dynamics. These effects are unique to highly coupled networks and require a unit of analysis that goes beyond the single person interacting with a single element of automation. As exemplified by the bullwhip effect and vicious cycles, the problems of supply chain management reflect generic challenges in using decentralized control to achieve a central objective. Decentralized networks promise efficiency and the capacity to adapt to unexpected perturbations, but their complexity and inefficient information sharing can lead people to respond to local rather than global considerations. Automation can alleviate the tendency for attention to local goals to magnify a small disturbance into a widespread disruption, or properly designed, it may alleviate this tendency. However, too little or too much trust in automation leads to inappropriate reliance, which can induce dysfunctional dynamics, such as the bullwhip effect and vicious cycles (Lee and Gao, 2006).

Other domains share the general promise and pitfalls of modern supply chain management. For example, power grid management involves a decentralized network that makes it possible to supply the United States efficiently with power, but it can fail catastrophically when cooperation and information sharing break down (Zhou et al., 2003). Similarly, datalink-enabled air traffic control makes it possible for pilots to negotiate flight paths efficiently, but it can fail when pilots have trouble anticipating the complex dynamics of the system (Olson

and Sarter, 2001; Mulkerin, 2003). Also, grid computing makes its enormous computing power available for use by many independent agents, but it can fail if load balancing and job scheduling do not consider global considerations (Lorch and Kafura, 2002; Chervenak et al., 2003). Overall, technology is creating many highly interconnected networks that have great potential but that also raise important concerns. Resolving these concerns depends on designing effective multioperator, multiautomation interactions.

### 5.3  Driving and Roadway Safety

Much of the existing research on automation has focused on operators of large complex systems for which expensive automation has been practical to develop. As computer and sensor technology becomes more affordable, automation will become more common in systems encountered in day-to-day life. Automation for cars and trucks is an example of automation that will touch the day-to-day lives of many people. Vehicle automation may touch more peoples' lives and have a greater safety consequence than any other type of automation. In the United States alone, people drive over 2 trillion miles a year in cars and light trucks (Pickrell and Schimek, 1999). The safety consequence is equally impressive. Over 6 million crashes kill approximately 42,000 people each year and result in an economic cost of over $164 billion per year (Wang et al., 1999). Motor vehicle crashes are also the leading cause of workplace injuries, being responsible for 42% of work-related fatalities (Bureau of Labor Statistics, 2003). Automation in cars and trucks, like that of increasing automation in other parts of daily life, has the potential to influence the safety and comfort of many people.

Functions that vehicle automation might support range from routing and navigation to collision avoidance and vehicle control (Lee, 1997; Young and Stanton, 2007). Table 2 shows some of the many examples of current and potential types of vehicle automation. Currently, examples include navigation systems that use GPS data and electronic map databases that give drivers turn-by-turn directions. Also, adaptive cruise control uses sensors and new control algorithms to extend cruise control so that cars slow down automatically and maintain a safe distance from the car ahead. Many vehicles even have a system that uses sensor data (e.g., airbag deployment) to detect a crash, calls for emergency aid, and then transmits the crash location using the car's GPS. The potential of automation to enhance the safety and comfort of drivers is substantial.

Designing automation to support driving confronts many of the same challenges as those found with automation in other domains. Sensor imperfections and complexity of the driving environment make adaptive cruise control and collision warning systems fallible. Recent studies suggest that adaptive cruise control may induce complacency and the potential of overtrust. Specifically, many drivers intervene too slowly to prevent a collision when the adaptive cruise control fails to brake (Stanton et al., 1997). Behavioral adaptation also threatens to undermine the safety benefits of automation. Automation aimed to enhance safety, such as an ABS,

**Table 2 Automation for Driving and Other In-Vehicle Technology**

| General Functions | Specific Examples |
|---|---|
| Routing and navigation | Trip planning, multimode travel coordination and planning, predrive route and destination selection, dynamic route selection, route guidance, route navigation, automated toll collection, route scheduling, posttrip summary |
| Motorist services | Broadcast services/attractions, services/attractions directory, destination coordination, delivery-related information |
| Augmented signage | Guidance sign information, notification sign information, regulatory sign information |
| Safety and warning | Immediate hazard warning, road condition information, aid request, vehicle condition monitoring, driver monitoring, sensory augmentation |
| Collision avoidance and vehicle control | Forward object collision avoidance, road departure collision avoidance, lane change and merge collision avoidance, intersection collision avoidance, railroad crossing collision avoidance, backing aid, vehicle control |
| Driver comfort, communication, and convenience | Real-time communication; asynchronous communication; contact search and history; entertainment and general information; heating, ventilation, air conditioning, and noise |

*Source*: Adapted from Lee and Kantowitz (2005).

has not produced the expected safety benefits because drivers with an ABS tend to change their driving behavior and follow more closely (Sagberg et al., 1997). A similar response may occur with collision warning systems that aim to give drivers advance notice of impending collisions. Such systems may lead some drivers to think they can safely engage in distracting activities, such as reading or watching DVDs, while driving. Understanding how to develop vehicle automation to enhance safety such that behavioral adaptation does not erode its benefits is a critical challenge.

Another challenge that confronts the design of vehicle automation is the potential for driver confusion in the face of many poorly integrated systems. Similar problems of automation coordination and integration have occurred with maritime navigation aids (Lee and Sanquist, 2000), flight management systems (Sarter and Woods, 1995), and medical devices (Cook et al., 1990a). Already, early examples of vehicle automation show the substantial confusion and frustration associated with poorly integrated systems, such as the recent controversy

and confusion regarding the 700 features of the BMW iDrive (Norman, 2003). Forward object, road departure, lane change, and intersection collision warning systems may all populate the car of the future, and identifying which warning has been activated may be a challenge for drivers. To avoid such confusion requires a design approach that considers the overall driving ecology and the information needed to negotiate it rather than an approach focused on sensor technology and arbitrarily defined collision types.

Unlike operators of automation in domains such as aviation and process control, drivers do not receive specific training on how to operate particular features of their car. In addition, drivers belong to a very heterogeneous group that spans a wide range of age, experience, and goals for driving. The difficulty of providing systematic training for automotive automation and the diversity of drivers make it likely that many drivers will misunderstand and misuse vehicle automation. Drivers misunderstand even a simple system, such as an ABS, and benefit from training on how to use it (Mollenhauer et al., 1997). More complex systems such as adaptive cruise control may confuse drivers, particularly as they move from a vehicle they are accustomed to driving to one they are not (e.g., a rental car). Ensuring that all drivers are properly trained is much more difficult than ensuring that process control operators or pilots understand the automation they manage. Automation that affects day-to-day life, such as vehicle automation, faces the particular challenges of being understood and used appropriately by a highly diverse array of potential users.

## 6  AUTOMATION—DOES IT NEED US?

The Luddites faced the prospect of automation changing their lives, and we face a similar prospect today. Increasingly sophisticated automation makes it possible to replace the human in many situations, and the situations in which humans outperform automation are diminishing rapidly. Although the need for human adaptability, creativity, and flexibility makes complete automation of most systems infeasible, the increasing capability of automation may eliminate even these reasons to include human operators. Soon, automating based on the criterion of whether the human or machine is better suited to perform a task may be irrelevant. This situation requires a deeper consideration of the purpose of technology (Hancock, 1996). Although automation allows people to avoid dangerous and unpleasant situations, unrestrained automation may eliminate activities that provide intrinsic enjoyment and purpose to life (Nickerson, 1999). Ironically, automating everything that is technologically possible or even everything that enhances system efficiency and safety may have the unanticipated effect of diminishing the lives of the people that automation should ultimately serve. Like the Luddites, we may ultimately need to confront the issue of whether automation needs us. "At least we have it in our power to say no to new technology, or do we?"(Sheridan, 2000, p. 203).

# REFERENCES

Akkermans, H., and van Helden, K. (2002), "Vicious and Virtuous Cycles in ERP Implementation: A Case Study of Interrelations between Critical Success Factors," *European Journal of Information Systems*, Vol. 11, No. 1, pp. 35–46.

Anderson, J. R., and Libiere, C. (1998), *Atomic Components of Thought*, Lawrence Erlbaum Associates, Mahwah, NJ.

Bailey, N. R., Scerbo, M. W., Freeman, F. G., Mikulka, P. J., and Scott, L. A. (2006), "Comparison of a Brain-Based Adaptive System and a Manual Adaptable System for Invoking Automation," *Human Factors*, Vol. 48, No. 4, pp. 693–709.

Bainbridge, L. (1983), "Ironies of Automation," *Automatica*, Vol. 19, No. 6, pp. 775–779.

Beck, H. P., Dzindolet, M. T., and Pierce, L. G. (2007), "Automation Usage Decisions: Controlling Intent and Appraisal Errors in a Target Detection Task," *Human Factors*, Vol. 49, No. 3, pp. 429–437.

Beck, H. P., McKinney, J. B., Dzindolet, M. T., and Pierce, L. G. (2009). "Effects of Human-Machine Competition on Intent Errors in a Target Detection Task," *Human Factors*, Vol. 51, pp. 477–486.

Beni, G., and Wang, J. (1993), "Swarm Intelligence in Cellular Robotic Systems," in *Robots and Biological Systems: Towards a New Bionics*, P. Dario, G. Sansani, and P. Aebischer, Eds., Springer-Verlag, Berlin.

Bettman, J. R. J. E. J., and Payne, J. W. (1990), "A Computational Analysis of Cognitive Effort in Choice," *Organizational Behavior and Human Decision Processes*, Vol. 45, pp. 111–139.

Billings, C. E. (1997), *Aviation Automation: The Search for a Human-Centered Approach*, Lawrence Erlbaum Associates, Mahwah, NJ.

Bliss, J., and Acton, S. A. (2003) "Alarm Mistrust in Automobiles: How Collision Alarm Reliability Affects Driving," *Applied Ergonomics*, Vol. 34, pp. 499–509.

Bliss, J. P. (1997), "Alarm Reaction Patterns by Pilots as a Function of Reaction Modality," *International Journal of Aviation Psychology*, Vol. 7, No. 1, pp. 1–14.

Bliss, J., Dunn, M., and Fuller, B. S. (1995), "Reversals of the Cry-Wolf Effect: An Investigation of Two Methods to Increase Alarm Response Rates," *Perceptual and Motor Skills*, Vol. 80, pp. 1231–1242.

Bocionek, S. R. (1995), "Agent Systems That Negotiate and Learn," *International Journal of Human-Computer Studies*, Vol. 42, No. 3, pp. 265–288.

Bonabeau, E., Theraulaz, G., Deneubourg, J. L., Aron, S., and Camazine, S. (1997), "Self-Organization in Social Insects," *Trends in Ecology and Evolution*, Vol. 12, No. 5, pp. 188–193.

Briggs, P., Burford, B., and Dracup, C. (1998), "Modeling Self-Confidence in Users of a Computer-Based System Showing Unrepresentative Design," *International Journal of Human-Computer Studies*, Vol. 49, No. 5, pp. 717–742.

Broadbent, D. E. (1958), *Perception and Communication*, Pergamon, London.

Brooks, R. A., and Flynn, A. M. (1993), "A Robot Being," in *Robots and Biological Systems: Towards a New Bionics*, P. Dario, G. Sansani, and P. Aebischer, Eds., Springer-Verlag, Berlin.

Brooks, R. A., Maes, P., Mataric, M. J., and More, G. (1990), "Lunar Base Construction Robots," in *Proceedings of the 1990 International Workshop on Intelligent Robots and Systems*, Ibaraki, Japan, pp. 389–392.

Bureau of Labor Statistics (2003) *Census of Fatal Occupational Injuries*, U.S. Department of Labor, Washington, DC.

Byrne, E. A., and Parasuraman, R. (1996), "Psychophysiology and Adaptive Automation," *Biological Psychology*, Vol. 42, No. 3, pp. 249–268.

Byrne, M. D., and Kirlik, A. (2005) "Using Computational Cognitive Modeling to Diagnose Possible Sources of Aviation Error," *International Journal of Aviation Psychology*, Vol. 15, No. 2, pp. 135–155.

Chervenak, A., et al. (2003) "High-Performance Remote Access to Climate Simulation Data: A Challenge Problem for Data Grid Technologies," *Parallel Computing*, Vol. 29, No. 10, pp. 1335–1356.

Cohen, M. S., Parasuraman, R., and Freeman, J. (1999) "Trust in Decision Aids: A Model and Its Training Implications," Technical Report USAATCOM TR 97-D-4, Cognitive Technologies, Arlington, VA.

Cook, R. I., Woods, D. D., and Howie, M. B. (1990a), "The Natural History of Introducing New Information Technology into a Dynamic High-Risk Environment," in *Proceedings of the Human Factors Society 34th Annual Meeting*, Human Factors Society, Santa Monica, CA.

Cook, R. I., Woods, D. D., McColligan, E., and Howie, M. B. (1990b), "Cognitive Consequences of 'Clumsy' Automation on High Workload, High Consequence Human Performance," paper presented at the Space Operations, Applications and Research Symposium, SOAR '90, NASA Johnson Space Center, Houston, TX.

Crocoll, W. M., and Coury, B. G. (1990), "Status or Recommendation: Selecting the Type of Information for Decision Aiding," in *Proceedings of the Human Factors Society 34th Annual Meeting*, Vol. 2, Human Factors Society, Santa Monica, CA, pp. 1524–1528.

Dearden, A., Harrison, M., and Wright, P. (2000), "Allocation of Function: Scenarios, Context and the Economics of Effort," *International Journal of Human-Computer Studies*, Vol. 52, No. 2, pp. 289–318.

Degani, A., and Heymann, M. (2002), "Formal Verification of Human-Automation Interaction," *Human Factors*, Vol. 44, No. 1, pp. 28–43.

Deutsch, M. (1958), "Trust and Suspicion," *Journal of Conflict Resolution*, Vol. 2, No. 4, pp. 265–279.

Deutsch, M. (1960), "The Effect of Motivational Orientation upon Trust and Suspicion," *Human Relations*, Vol. 13, pp. 123–139.

de Vries, P., Midden, C., and Bouwhuis, D. (2003) "The Effects of Errors on System Trust, Self-Confidence, and the Allocation of Control in Route Planning," *International Journal of Human-Computer Studies*, Vol. 58, No. 6, pp. 719–735.

Dixon, S. R., Wickens, C. D., and McCarley, J. S. (2007). "On the Independence tf Compliance and Reliance: Are Automation False Alarms Worse Than Misses?," *Human Factors*, Vol. 49, No. 4, pp. 564–572.

Dyer, J. H., and Singh, H. (1998), "The Relational View: Cooperative Strategy and Sources of Interorganizational Competitive Advantage," *Academy of Management Review*, Vol. 23, No. 4, pp. 660–679.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., and Anderson, B. W. (2001), "Predicting Misuse and Disuse of Combat Identification Systems," *Military Psychology*, Vol. 13, No. 3, pp. 147–164.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., and Dawe, L. A. (2002), "The Perceived Utility of Human and Automated Aids in a Visual Detection Task," *Human Factors*, Vol. 44, No. 1, pp. 79–94.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003) "The Role of Trust in Automation Reliance," *International Journal of Human-Computer Studies*, Vol. 58, No. 6, pp. 697–718.

Endsley, M. R., and Kaber, D. B. (1999) "Level of Automation Effects on Performance, Situation Awareness and Workload in a Dynamic Control Task," *Ergonomics*, Vol. 42, No. 3, pp. 462–492.

Endsley, M. R., and Kiris, E. O. (1995), "The Out-of-the-Loop Performance Problem and Level of Control in Automation," *Human Factors*, Vol. 37, No. 2, pp. 381–394.

Entin, E. B., Entin, E. E., and Serfaty, D. (1996), "Optimizing Aided Target-Recognition Performance," in *Proceedings of the Human Factors and Ergonomics Society*, Vol. 1, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 233–237.

Eprath, A. R., and Young, L. R. (1981), "Monitoring vs. Man-in-the-Loop Detection of Aircraft Control Failures," in *Human Detection and Diagnosis of System Failures*, J. Rasmussen and W. B. Rouse, Eds., Plenum, New York, pp. 143–154.

Evans, L. (1991), *Traffic Safety and the Driver*, Van Nostrand Reinhold, New York.

Farrell, S., and Lewandowsky, S. (2000), "A Connectionist Model of Complacency and Adaptive Recovery under Automation," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 26, No. 2, pp. 395–410.

Fitts, P. M. (1951), *Human Engineering for an Effective Air Navigation and Traffic Control System*, National Research Council, Washington, DC.

Flach, J. M. (1999) "Ready, Fire, Aim: A 'Meaning-Processing' Approach to Display Design," in *Attention and Performance XVII, Cognitive Regulation of Performance: Interaction of Theory and Application*, D. Gophen and A. Koriat, Eds., MIT Press, Cambridge, MA, pp. 197–221.

Flach, J. M., and Jagacinski, R. J. (2002), *Control Theory for Humans*, Lawrence Erlbaum Associates, Mahwah, NJ.

Fogg, B., et al. (2001a), "Web Credibility Research: A Method for Online Experiments and Early Study Results," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ACM, New York, pp. 293–294.

Fogg, B., et al. (2001b), "What Makes Web Sites Credible? A Report on a Large Quantitative Study," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Seattle, WA, ACM, New York, pp. 293–294.

Fukuda, T., Funato, D., Sekiyama, K., and Arai, F. (1998), "Evaluation on Flexibility of a Swarm Intelligent System," in *Proceedings of the 1998 IEEE International Conference on Robotics and Automation*, Leuven, Belgium, pp. 3210–3215.

Gao, J., and Lee, J. D. (2004), "Information Sharing, Trust, and Reliance: A Dynamic Model of Multi-Operator Multi-Automation Interaction," in *Proceedings of the 5th Conference on Human Performance, Situation Awareness and Automation Technology*, Vol. 2, D. A. Vincenzi,

M. Mouloua, and P. A. Hancock, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 34–39.

Gao, J., and Lee, J. D. (2006). "A Dynamic Model of Interaction between Reliance on Automation and Cooperation in Multi-Operator Multi-Automation Situations," *International Journal of Industrial Ergonomics*, Vol. 36, No. 5, pp. 512–526.

Gibson, J. J. (1962), "Observations on Active Touch," *Psychological Review*, Vol. 69, pp. 477–491.

Goodrich, M. A., and Boer, E. R. (2003) "Model-Based Human-Centered Task Automation: A Case Study in ACC System Design," *IEEE Transactions on Systems, Man and Cybernetics, Part A, Systems and Humans*, Vol. 33, No. 3, pp. 325–336.

Grabowski, M. R., and Hendrick, H. (1993), "How Low Can We Go? Validation and Verification of a Decision Support System for Safe Shipboard Manning," *IEEE Transactions on Engineering Management*, Vol. 40, No. 1, pp. 41–53.

Guerlain, S. A., et al. (1999) "Interactive Critiquing as a Form of Decision Support: An Empirical Evaluation," *Human Factors*, Vol. 41, No. 1, pp. 72–89.

Guerlain, S., Jamieson, G. A., Bullemer, P., and Blair, R. (2002), "The MPC Elucidator: A Case Study in the Design for Human-Automation Interaction," *IEEE Transactions on Systems, Man and Cybernetics, Part A, Systems and Humans*, Vol. 32, No. 1, pp. 25–40.

Gupta, N., Bisantz, A. M., and Singh, T. (2002), "The Effects of Adverse Condition Warning System Characteristics on Driver Performance: An Investigation of Alarm Signal Type and Threshold Level," *Behaviour and Information Technology*, Vol. 21, No. 4, pp. 235–248.

Halprin, S., Johnson, E., and Thornbury, J. (1973), "Cognitive Reliability in Manned Systems," *IEEE Transactions on Reliability*, Vol. 22, pp. 165–169.

Hancock, P. A. (1996), "Teleology of Technology," in *Automation and Human Performance*, R. Parasuraman and M. Mouloua, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 461–498.

Hancock, P. A., and Warm, J. S. (1989), "A Dynamic Model of Stress and Sustained Attention," *Human Factors*, Vol. 31, No. 5, pp. 519–537.

Handfield, R. B., and Bechtel, C. (2002), "The Role of Trust and Relationship Structure in Improving Supply Chain Responsiveness," *Industrial Marketing Management*, Vol. 31, No. 4, pp. 367–382.

Harris, W. C., Hancock, P. A., Arthur, E. J., and Caird, J. K. (1995), "Performance, Workload, and Fatigue Changes Associated with Automation," *International Journal of Aviation Psychology*, Vol. 5, No. 2, pp. 169–185.

Hendricks, K. B., and Singhal, V. R. (2003) "The Effect of Supply Chain Glitches on Shareholder Wealth," *Journal of Operations Management*, Vol. 21, No. 5, pp. 501–522.

Heymann, M., and Degani, A. (2007), "Formal Analysis and Automatic Generation of User Interfaces: Approach, Methodology, and an Algorithm," *Human Factors*, Vol. 49, No. 2, pp. 311–330.

Hoc, J. M. (1993), "Some Dimensions of a Cognitive Typology of Process-Control Situations," *Ergonomics*, Vol. 36, No. 11, pp. 1445–1455.

Hoc, J. M. (2000), "From Human-Machine Interaction to Human-Machine Cooperation," *Ergonomics*, Vol. 43, No. 7, pp. 833–843.

Hoc, J. H. (2001), "Towards a Cognitive Approach to Human–Machine Cooperation in Dynamic Situations," *International Journal of Human–Computer Studies*, Vol. 54, No. 4, pp. 509–540.

Hollan, J., Hutchins, E., and Kirsh, D. (2000), "Distributed Cognition: Toward a New Foundation for Human-Computer Interaction Research," *ACM Transactions on Computer-Human Interaction*, Vol. 7, No. 2, pp. 174–196.

Hollnagel, E., and Bye, A. (2000), "Principles for Modelling Function Allocation," *International Journal of Human-Computer Studies*, Vol. 52, No. 2, pp. 253–265.

Hovland, C. I., Janis, I. L., and Kelly, H. H. (1953), *Communication and Persuasion: Psychological Studies of Opinion Change*, Yale University Press, New Haven, CT.

Hughes, J. S., Rice, S., Trafimow, D., and Clayton, K. (2009), "The Automated Cockpit: A Comparison of Attitudes towards Human And Automated Pilots," *Transportation Research Part F—Traffic Psychology and Behaviour*, Vol. 12, No. 5, pp. 428–439.

Hutchins, E. (1995), *Cognition in the Wild*, MIT Press, Cambridge, MA.

Inagaki, T. (2003) "Automation and the Cost of Authority," *International Journal of Industrial Ergonomics*, Vol. 31, No. 3, pp. 169–174.

Johnson, P. J., and Bay, J. S. (1995), "Distributed Control of Simulated Autonomous Mobile Robot Collectives in Payload Transportation," *Autonomous Robots*, Vol. 2, No. 1, pp. 43–63.

Jones, P. M., and Jacobs, J. L. (2000), "Cooperative Problem Solving in Human-Machine Systems: Theory, Models, and Intelligent Associate Systems," *IEEE Transactions on Systems, Man and Cybernetics, Part C, Applications and Reviews*, Vol. 30, No. 4, pp. 397–407.

Kaber, D. B., Onal, E., and Endsley, M. R. (2000), "Design of Automation for Telerobots and the Effect on Performance, Operator Situation Awareness, and Subjective Workload," *Human Factors and Ergonomics in Manufacturing*, Vol. 10, No. 4, pp. 409–430.

Kaber, D. B., Riley, J. M., Tan, K. W., and Endsley, M. R. (2001), "On the Design of Adaptive Automation for Complex Systems," *International Journal of Cognitive Ergonomics*, Vol. 5, No. 1, pp. 37–57.

Kantowitz, B. H., and Sorkin, R. D. (1987), "Allocation of Functions," in *Handbook of Human Factors*, G. Salvendy, Ed., Wiley, New York, pp. 355–369.

Kirlik, A. (1993), "Modeling Strategic Behavior in Human—Automation Interaction: Why an 'Aid' Can (and Should) Go Unused," *Human Factors*, Vol. 35, No. 2, pp. 221–242.

Klein, G., Ross, K. G., Moon, B. M., Klein, D. E., Hoffman, R. R., and Hollnagel, E. (2003) "Macrocognition," *IEEE Intelligent Systems*, Vol. 18, No. 3, pp. 81–85.

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., and Feltovich, P. J. (2004), "Ten Challenges for Making Automation A 'Team Player' In Joint Human-Agent Activity," *IEEE Intelligent Systems*, Vol. 19, No. 6, pp. 91–95.

Landauer, T. K. (1995), *The Trouble with Computers: Usefulness, Usability, and Productivity*, MIT Press, Cambridge, MA.

Lee, J. D. (1997), "A Functional Description of ATIS/CVO Systems to Accommodate Driver Needs and Limits," in *Ergonomics and Safety of Intelligent Driver Interfaces*, Y. I. Noy, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 63–84.

Lee, J. D. (2001), "Emerging Challenges in Cognitive Ergonomics: Managing Swarms of Self-Organizing Agent-Based Automation," *Theoretical Issues in Ergonomics Science*, Vol. 2, No. 3, pp. 238–250.

Lee, J. D. (2006). "Affect, Attention, and Automation," in *Attention: From Theory to Practice*, A. Kramer, D. Wiegmann, and A. Kirlik, Eds., Oxford University Press, New York.

Lee, J. D., and Gao, J. (2006), "Trust, Automation, and Cooperation in Supply Chains," *Supply Chain Forum: An International Journal*, Vol. 6, No. 2, pp. 82–89.

Lee, J. D., and Kantowitz, B. K. (2005) "Network Analysis of Information Flows to Integrate In-Vehicle Information Systems," *International Journal of Vehicle Information and Communication Systems*, Vol. 1, No. 1/2, pp. 24–43.

Lee, J. D., and Moray, N. (1992), "Trust, Control Strategies and Allocation of Function in Human-Machine Systems," *Ergonomics*, Vol. 35, No. 10, pp. 1243–1270.

Lee, J. D., and Moray, N. (1994), "Trust, Self-Confidence, and Operators' Adaptation to Automation," *International Journal of Human-Computer Studies*, Vol. 40, pp. 153–184.

Lee, J. D., and Morgan, J. (1994), "Identifying Clumsy Automation at the Macro Level: Development of a Tool to Estimate Ship Staffing Requirements," in *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*, Vol. 2, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 878–882.

Lee, J. D., and Sanquist, T. F. (1996), "Maritime Automation," in *Automation and Human Performance*, R. Parasuraman and M. Mouloua, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 365–384.

Lee, J. D., and Sanquist, T. F. (2000), "Augmenting the Operator Function Model with Cognitive Operations: Assessing the Cognitive Demands of Technological Innovation in Ship Navigation," *IEEE Transactions on Systems, Man and Cybernetics, Part A, Systems and Humans*, Vol. 30, No. 3, pp. 273–285.

Lee, J. D., and See, K. A. (2004), "Trust in Technology: Designing for Appropriate Reliance," *Human Factors*, Vol. 46, No. 1, pp. 50–80.

Lee, J. D., and Seppelt, B. D. (2009), "Human Factors of Automation Design," in *Handbook of Automation Design*, S. Nof, Ed., Springer, New York, pp. 417–436.

Lee, J. D., and Strayer, D. L. (2004), "Preface to a Special Section on Driver Distraction," *Human Factors*, Vol. 46, pp. 583–586.

Lee, H. L., and Whang, S. J. (2000), "Information Sharing in a Supply Chain," *International Journal of Technology Management*, Vol. 20, Nos. 3-4, pp. 373–387.

Lee, J. D., Gore, B. F., and Campbell, J. L. (1999) "Display Alternatives for In-Vehicle Warning and Sign Information: Message Style, Location, and Modality," *Transportation Human Factors Journal*, Vol. 1, No. 4, pp. 347–377.

Lee, J. D., Hoffman, J. D., and Hayes, E. (2004), "Collision Warning Design to Mitigate Driver Distraction," in *Proceedings of CHI*, Association for Computing Machinery, Vienna, Austria, ACM, New York, pp. 65–72.

Lees, M. N., and Lee, J. D. (2007), "The Influence of Distraction and Driving Context ON Driver Response to Imperfect Collision Warning Systems," *Ergonomics*, Vol. 50, No. 8, pp. 1264–1286.

Leveson, N. G. (1995), *Safeware: System Safety and Computers*, Addison-Wesley, Reading, MA.

Lewandowsky, S., Mundy, M., and Tan, G. (2000), "The Dynamics of Trust: Comparing Humans to Automation," *Journal of Experimental Psychology: Applied*, Vol. 6, No. 2, pp. 104–123.

Lewis, M. (1998), "Designing for Human-Agent Interaction," *AI Magazine*, Vol. 19, No. 2, pp. 67–78.

Lorch, M., and Kafura, D. (2002), "Supporting Secure Ad-Hoc User Collaboration in Grid Environments," *Grid Computing*, Vol. 2536, pp. 181–193.

Lutzhoft, M. H., and Dekker, S. W. A. (2002), "On Your Watch: Automation on the Bridge," *Journal of Navigation*, Vol. 55, No. 1, pp. 83–96.

MacMillan, J., Entin, E. B., and Serfaty, D. (1994), "Operator Reliance on Automated Support for Target Acquisition," in *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*, Vol. 2, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1285–1289.

Madhavan, P., and Wiegman, D. A. (2007), "Similarities and Differences between Human-Human and Human-Automation Trust: An Integrative Review," *Theoretical Issues in Ergonomics Science*, Vol. 8, No. 4, pp. 277–301.

Madhavan, P., Wiegmann, D. A., and Lacson, F. C. (2006), "Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids," *Human Factors*, Vol. 48, No. 2, pp. 241–256.

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995), "An Integrative Model of Organizational Trust," *Academy of Management Review*, Vol. 20, No. 3, pp. 709–734.

McFadden, S., Vimalachandran, A., and Blackmore, E. (2003) "Factors Affecting Performance on a Target Monitoring Task Employing an Automatic Tracker," *Ergonomics*, Vol. 47, No. 3, pp. 257–280.

Metzger, U., and Parasuraman, R. (2001), "The Role of the Air Traffic Controller in Future Air Traffic Management: An Empirical Study of Active Control Versus Passive Monitoring," *Human Factors*, Vol. 43, No. 4, pp. 519–528.

Meyer, J. (2001), "Effects of Warning Validity and Proximity on Responses to Warnings," *Human Factors*, Vol. 43, No. 4, pp. 563–572.

Meyer, J. (2004), "Conceptual Issues in the Study of Dynamic Hazard Warnings," *Human Factors*, Vol. 46, No. 2, pp. 196–204.

Meyer, J., and Bitan, Y. (2002), "Why Better Operators Receive Worse Warnings," *Human Factors*, Vol. 44, No. 3, pp. 343–353.

Michon, J. A. (1985), "A Critical View of Driver Behavior Models: What Do We Know, What Should We Do?" in *Human Behavior and Traffic Safety*, L. Evans and R. C. Schwing, Eds., Plenum, New York, pp. 485–520.

Miller, C. A. (2002), "Definitions and Dimensions of Etiquette," in *Etiquette for Human-Computer Work*, C. Miller, Ed., Technical Report FS-02-02, American Association for Artificial Intelligence, Menlo Park, CA, pp. 1–7.

Miller, C. A., and Funk, H. B. (2001), "Associates with Etiquette: Communication to Make Human-Automation Interaction More Natural, Productive and Polite," in *Proceedings of the 8th European Conference on Cognitive Science Approaches to Process Control*, Munich, September 24-26, pp. 1–8.

Min, T. W., and Yin, H. K. (1998), "A Decentralized Approach for Cooperative Sweeping by Multiple Mobile Robots," in *Proceedings of the 1998 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Victoria, B.C., Canada, pp. 380–385.

Minsky, M. (1986), *The Society of Mind*, Simon h& Schuster, New York.

Mollenhauer, M. A., Dingus, T. A., Carney, C., Hankey, J. M., and Jahns, S. (1997), "Antilock Brake Systems: An Assessment of Training on Driver Effectiveness," *Accident Analysis and Prevention*, Vol. 29, No. 1, pp. 97–108.

Montgomery, D. A., and Sorkin, R. D. (1996), "Observer Sensitivity to Element Reliability in a Multielement Visual Display," *Human Factors*, Vol. 38, No. 3, pp. 484–494.

Moray, N. (2003) "Monitoring, Complacency, Scepticism and Eutactic Behaviour," *International Journal of Industrial Ergonomics*, Vol. 31, No. 3, pp. 175–178.

Moray, N., Inagaki, T., and Itoh, M. (2000), "Adaptive Automation, Trust, and Self-Confidence in Fault Management of Time-Critical Tasks," *Journal of Experimental Psychology: Applied*, Vol. 6, No. 1, pp. 44–58.

Mosier, K. L., Skitka, L. J., Heers, S., and Burdick, M. (1998), "Automation Bias: Decision Making and Performance in High-Tech Cockpits," *International Journal of Aviation Psychology*, Vol. 8, No. 1, pp. 47–63.

Muir, B. M., and Moray, N. (1996), "Trust in Automation, 2: Experimental Studies of Trust and Human Intervention in a Process Control Simulation," *Ergonomics*, Vol. 39, No. 3, pp. 429–460.

Mulder, M., Van Paassen, M. M., and Abbink, D. A. (2008), "Haptic Gas Pedal Feedback," *Ergonomics*, Vol. 51, No. 11, pp. 1710–1720.

Mulkerin, T. (2003) "Free Flight Is in the Future: Large-Scale Controller Pilot Data Link Communications Emulation Testbed," *IEEE Aerospace and Electronic Systems Magazine*, Vol. 18, No. 9, pp. 23–27.

Murray, J., and Liu, Y. (1997), "Hortatory Operations in Highway Traffic Management," *IEEE Transactions on Systems, Man and Cybernetics, Part A, Systems and Humans*, Vol. 27, No. 3, pp. 340–350.

Nagel, D. C. (1988), "Human Error in Aviation Operations," in *Human Factors in Aviation*, E. Weiner and D. Nagel, Eds., Academic, New York, pp. 263–303.

Nass, C., and Lee, K. N. (2001), "Does Computer-Synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity-Attraction, and Consistency-Attraction," *Journal of Experimental Psychology: Applied*, Vol. 7, No. 3, pp. 171–181.

National Transportation Safety Board (NTSB) (1990), "Marine Accident Report: Grounding of the U.S. Tankship Exxon Valdez on Bligh Reef, Prince William Sound, Valdez, Alaska, March 24, 1989," Report NTSB/MAR90/04, NTSB, Washington, DC.

National Transportation Safety Board (NTSB) (1997), "Marine Accident Report: Grounding of the Panamanian Passenger Ship Royal Majesty on Rose and Crown Shoal near Nantucket, Massachusetts, June 10, 1995," Report NTSB/MAR97/01, NTSB, Washington, DC.

Nickerson, R. S. (1999)in *Automation, Technology and Human Performance*, M. W. Scerbo and M. Mouloua, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 11–19.

Nikolic, M. I., and Sarter, N. B. (2001), "Peripheral Visual Feedback: A Powerful Means of Supporting Effective Attention Allocation in Event-Driven, Data-Rich Environments," *Human Factors*, Vol. 43, No. 1, pp. 30–38.

Norman, D. A. (1990), "The 'Problem' with Automation: Inappropriate Feedback and Interaction, Not 'Overautomation,'" *Philosophical Transactions of the Royal Society, London, Series B, Biological Sciences*, Vol. 327, No. 1241, pp. 585–593.

Norman, D. A. (1993), *Things That Make Us Smart*, Addison-Wesley, Reading, MA.

Norman, D. A. (2003) "Interaction Design for Automobile Interiors," available: http://www.jnd.org/dn.mss/InteractDsgnAutos.html, accessed April 8, 2004.

Ockerman, J. J. (1999) "Over-Reliance Issues with Task-Guidance Systems," in *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting*, Vol. 2, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1192–1196.

Olson, W. A., and Sarter, N. B. (2000), "Automation Management Strategies: Pilot Preferences and Operational Experiences," *International Journal of Aviation Psychology*, Vol. 10, No. 4, pp. 327–341.

Olson, W. A., and Sarter, N. B. (2001), "Management by Consent in Human-Machine Systems: When and Why It Breaks Down," *Human Factors*, Vol. 43, No. 2, pp. 255–266.

Parasuraman, R., and Riley, V. (1997), "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors*, Vol. 39, No. 2, pp. 230–253.

Parasuraman, R., Molloy, R., and Singh, I. (1993), "Performance Consequences of Automation-Induced 'Complacency'" *international Journal of Aviation Psychology*, Vol. 3, No. 1, pp. 1–23.

Parasuraman, R., Mouloua, M., and Molloy, R. (1994), "Monitoring Automation Failures in Human-Machine Systems," in *Human Performance in Automated Systems: Current Research and Trends*, M. Mouloua and R. Parasuraman, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 45–49.

Parasuraman, R., Mouloua, M., and Molloy, R. (1996), "Effects of Adaptive Task Allocation on Monitoring of Automated Systems," *Human Factors*, Vol. 38, No. 4, pp. 665–679.

Parasuraman, R., Hancock, P. A., and Olofinboba, O. (1997), "Alarm Effectiveness in Driver-Centred Collision-Warning Systems," *Ergonomics*, Vol. 40, No. 3, pp. 390–399.

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000), "A Model for Types and Levels of Human Interaction with Automation," *IEEE Transactions on Systems, Man and Cybernetics, Part A, Systems and Humans*, Vol. 30, No. 3, pp. 286–297.

Patterson, E. S. (1999) "A Simulation Study of Computer-Supported Inferential Analysis Under Data Overload," in *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting*, Vol. 1, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 363–368.

Payne, J. W., Bettman, J. R., Coupey, E., and Johnson, E. J. (1992), "A Constructive Process View of Decision-Making: Multiple Strategies in Judgment and Choice," *Acta Psychologica*, Vol. 80, Nos. 1-3, pp. 107–141.

Perrow, C. (1984), *Normal Accidents*, Basic Books, New York.

Pickrell, D., and Schimek, P. (1999) "Growth in Motor Vehicle Ownership and Use: Evidence from the Nationwide Personal Transportation Survey," *Journal of Transportation and Statistics*, Vol. 2, No. 1, pp. 1–18.

Prinzel, L. J., Freeman, F. C., Scerbo, M. W., Mikulka, P. J., and Pope, A. T. (2000), "A Closed-Loop System for Examining Psychophysiological Measures for Adaptive Task Allocation," *International Journal of Aviation Psychology*, Vol. 10, No. 4, pp. 393–410.

Ranney, T. A. (1994), "Models of Driving Behavior: A Review of Their Evolution," *Accident Analysis and Prevention*, Vol. 26, No. 6, pp. 733–750.

Rasmussen, J. (1986), *Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering*, North-Holland, New York.

Reeves, B., and Nass, C. (1996), *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, Cambridge University Press, New York.

Rempel, J. K., Holmes, J. G., and Zanna, M. P. (1985), "Trust in Close Relationships," *Journal of Personality and Social Psychology*, Vol. 49, No. 1, pp. 95–112.

Resnick, M. (1991), *Turtles, Termites, and Traffic Jams: Explorations in Massively Parallel Microworlds*, MIT Press, Cambridge, MA.

Riley, V. (1989), "A General Model of Mixed-Initiative Human-Machine Systems," in *Proceedings of the Human Factors and Ergonomics Society 33rd Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 124–128.

Riley, V. (2001), "A New Language for Pilot Interfaces," *Ergonomics in Design*, Vol. 9, No. 2, pp. 21–27.

Riley, V. A. (1994), "Human Use of Automation," Ph.D. dissertation, University of Minnesota, Minneapolis, MN.

Ross, W., and LaCroix, J. (1996), "Multiple Meanings of Trust in Negotiation Theory and Research: A Literature Review and Integrative Model," *International Journal of Conflict Management*, Vol. 7, No. 4, pp. 314–360.

Roth, E. M., and Woods, D. D. (1988), "Aiding Human Performance, I: From Cognitive Analysis to Support Systems," *Human Factors*, Vol. 51, No. 1, pp. 39–64.

Roth, E. M., Bennett, K. B., and Woods, D. D. (1987), "Human Interaction with an 'Intelligent' Machine," *International Journal of Man-Machine Studies*, Vol. 27, pp. 479–526.

Roth, E. M., Bennett, K. B., and Woods, D. D. (1988), "Human Interaction with an 'Intelligent' Machine," in *Cognitive Engineering in Complex, Dynamic Worlds*, G. Mancini, D. D. Woods, and E. Hollnagel, Eds., Academic Press, London, pp. 23–69.

Rotter, J. B. (1967), "A New Scale for the Measurement of Interpersonal Trust," *Journal of Personality*, Vol. 35, No. 4, pp. 651–665.

Rouse, W. B. (1988), "Adaptive Aiding for Human/Computer Control," *Human Factors*, Vol. 30, pp. 431–443.

Sagberg, F., Fosser, S., and Saetermo, I. A. F. (1997), "An Investigation of Behavioural Adaptation to Airbags and Antilock Brakes among Taxi Drivers," *Accident Analysis and Prevention*, Vol. 29, No. 3, pp. 293–302.

Sarter, N. B. (2000), "The Need for Multisensory Interfaces in Support of Effective Attention Allocation in Highly Dynamic Event-Driven Domains: The Case of Cockpit Automation," *International Journal of Aviation Psychology*, Vol. 10, No. 3, pp. 231–245.

Sarter, N. B., and Schroeder, B. (2001), "Supporting Decision Making and Action Selection under Time Pressure and Uncertainty: The Case of In-Flight Icing," *Human Factors*, Vol. 43, No. 4, pp. 573–583.

Sarter, N. B., and Woods, D. D. (1994), "Decomposing Automation: Autonomy, Authority, Observability and Perceived Animacy," in *Human Performance in Automated Systems: Current Research and Trends*, M. Mouloua and R. Parasuraman, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 22–27.

Sarter, N. B., and Woods, D. D. (1995), "How in the World Did We Ever Get in That Mode? Mode Error and Awareness in Supervisory Control," *Human Factors*, Vol. 37, No. 1, pp. 5–19.

Sarter, N. B., and Woods, D. D. (1997), "Team Play with a Powerful and Independent Agent: Operational Experiences and Automation Surprises on the Airbus A-320," *Human Factors*, Vol. 39, No. 4, pp. 553–569.

Sarter, N. B., and Woods, D. D. (2000), "Team Play with a Powerful and Independent Agent: A Full-Mission Simulation Study," *Human Factors*, Vol. 42, No. 3, pp. 390–402.

Sarter, N. B., Woods, D. D., and Billings, C. E. (1997), "Automation Surprises," in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., Wiley, New York, pp. 1926–1943.

Sarter, N. B., Mumaw, R. J., and Wickens, C. D. (2007). "Pilots' Monitoring Strategies and Performance on Automated Flight Decks: An Empirical Study Combining Behavioral and Eye-Tracking Data," *Human Factors*, Vol. 49, No. 3, pp. 347–357.

Scallen, S. F., and Hancock, P. A. (2001), "Implementing Adaptive Function Allocation," *International Journal of Aviation Psychology*, Vol. 11, No. 2, pp. 197–221.

Scerbo, M. W. (1996), "Theoretical Perspectives on Adaptive Automation," in *Automation and Human Performance*, R. Parasuraman and M. Mouloua, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 37–63.

Schuffel, J. J., Boer, P. A., and van Breda, L. (1988), "The Ship's Wheelhouse of the Nineties: The Navigation Performance and Mental Workload of the Officer of the Watch," *Journal of Navigation*, Vol. 42, No. 1, pp. 60–72.

Seeley, T. D. (1997), "Honey Bee Colonies Are Group-Level Adaptive Units," *American Naturalist*, Vol. 150, pp. S22–S41.

Seifert, C. M., and Hutchins, E. L. (1992), "Error as Opportunity: Learning in a Cooperative Task," *Human-Computer Interaction*, Vol. 7, pp. 409–435.

Sharit, J. (2003) "Perspectives on Computer Aiding in Cognitive Work Domains: Toward Predictions of Effectiveness and Use," *Ergonomics*, Vol. 46, Nos. 1-3, pp. 126–140.

Sheridan, T. B. (1992), *Telerobotics, Automation, and Human Supervisory Control*, MIT Press, Cambridge, MA.

Sheridan, T. B. (2000), "Function Allocation: Algorithm, Alchemy or Apostasy?" *International Journal of Human-Computer Studies*, Vol. 52, No. 2, pp. 203–216.

Sheridan, T. B. (2002), *Humans and Automation*, Wiley, New York.

Sheridan, T. B., and Ferrell, W. R. (1974), *Man-Machine Systems: Information, Control, and Decision Models of Human Performance*, MIT Press, Cambridge, MA.

Sheridan, T. B., and Hennessy, R. T. (1984), *Research and Modeling of Supervisory Control Behavior*, National Academy Press, Washington, DC.

Sheridan, T. B., and Parasuraman, R. (2000), "Human Versus Automation in Responding to Failures: An Expected-Value Analysis," *Human Factors*, Vol. 42, No. 3, pp. 403–407.

Sheridan, T. B., and Verplank, W. L. (1978), "Human and Computer Control of Undersea Teleoperators," Technical Report, Man-Machine Systems Laboratory, Department of Mechanical Engineering, MIT, Cambridge, MA.

Singh, D. T., and Singh, P. P. (1997), "Aiding DSS Users in the Use of Complex OR Models," *Annals of Operations Research*, Vol. 72, pp. 5–27.

Sitkin, S. B., and Roth, N. L. (1993), "Explaining the Limited Effectiveness of Legalistic 'Remedies' for Trust/Distrust," *Organization Science*, Vol. 4, No. 3, pp. 367–392.

Skitka, L. J., Mosier, K. L., and Burdick, M. (1999) "Does Automation Bias Decision-Making?" *International Journal of Human-Computer Studies*, Vol. 51, No. 5, pp. 991–1006.

Skitka, L. J., Mosier, K., and Burdick, M. D. (2000a), "Accountability and Automation Bias," *International Journal of Human-Computer Studies*, Vol. 52, No. 4, pp. 701–717.

Skitka, L. J., Mosier, K. L., Burdick, M., and Rosenblatt, B. (2000b), "Automation Bias and Errors: Are Crews Better Than Individuals?" *International Journal of Aviation Psychology*, Vol. 10, No. 1, pp. 85–97.

Sklar, A. E., and Sarter, N. B. (1999) "Good Vibrations: Tactile Feedback in Support of Attention Allocation and Human-Automation Coordination in Event-Driven Domains," *Human Factors*, Vol. 41, No. 4, pp. 543–552.

Smilowitz, K. (2006), "Multi-Resource Routing with Flexible Tasks: An Application in Drayage Operations," IEMS Working Paper.

Smith, P. J., McCoy, E., and Layton, C. (1997), "Brittleness in the Design of Cooperative Problem-Solving Systems: The Effects on User Performance," *IEEE Transactions on Systems, Man and Cybernetics, Part A, Systems and Humans*, Vol. 27, No. 3, pp. 360–371.

Sniezek, J. A., Wilkins, D. C., Wadlington, P. L., and Baumann, M. R. (2002), "Training for Crisis Decision-Making: Psychological Issues and Computer-Based Solutions," *Journal of Management Information Systems*, Vol. 18, No. 4, pp. 147–168.

Stanton, N. A., and Pinto, M. (2000), "Behavioural Compensation by Drivers of a Simulator When Using a Vision Enhancement System," *Ergonomics*, Vol. 43, No. 9, pp. 1359–1370.

Stanton, N. A., and Young, M. S. (1998), "Vehicle Automation and Driving Performance," *Ergonomics*, Vol. 41, No. 7, pp. 1014–1028.

Stanton, N. A., Young, M., and McCaulder, B. (1997), "Drive-by-Wire: The Case of Driver Workload and Reclaiming Control with Adaptive Cruise Control," *Safety Science*, Vol. 27, No. 2-3, pp. 149–159.

Sterman, J. D. (1989), "Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision-Making Experiment," *Management Science*, Vol. 35, No. 3, pp. 321–339.

Stickland, T. R., Britton, N. F., and Franks, N. R. (1995), "Complex Trails and Simple Algorithms in Ant Foraging," *Proceedings of the Royal Society of London, Series B, Biological Sciences*, Vol. 260, No. 1357, pp. 53–58.

Strain, J., and Eason, K. (2000), "Exploring the Implications of Allocation of Function for Human Resource Management in the Royal Navy," *International Journal of Human-Computer Studies*, Vol. 52, No. 2, pp. 319–334.

Sugihara, K., and Suzuki, I. (1990), "Distributed Motion Coordination of Multiple Mobile Robots," in *Proceedings of the 5th IEEE International Symposium on Intelligent Control*, Philadelphia, Pennsylvania, pp. 138–143.

Tenner, E. (1996), *Why Things Bite Back: Technology and the Revenge of Unanticipated Consequences*, Alfred A. Knopf, New York.

Tenney, Y. J., Rogers, W. H., and Pew, R. W. (1998), "Pilot Opinions on Cockpit Automation Issues," *International Journal of Aviation Psychology*, Vol. 8, No. 2, pp. 103–120.

Todd, P., and Benbasat, I. (1999) "Evaluating the Impact of DSS, Cognitive Effort, and Incentives on Strategy Selection," *Information Systems Research*, Vol. 10, No. 4, pp. 356–374.

Todd, P., and Benbasat, I. (2000), "Inducing Compensatory Information Processing through Decision Aids That Facilitate Effort Reduction: An Experimental Assessment," *Journal of Behavioral Decision Making*, Vol. 13, No. 1, pp. 91–106.

Tseng, S., and Fogg, B. J. (1999) "Credibility and Computing Technology," *Communications of the ACM*, Vol. 42, No. 5, pp. 39–44.

Vicente, K. J. (1990), "Coherence- and Correspondence-Driven Work Domains: Implications for Systems Design," *Behaviour and Information Technology*, Vol. 9, pp. 493–502.

Vicente, K. J. (1999) *Cognitive Work Analysis: Towards Safe, Productive, and Healthy Computer-Based Work*, Lawrence Erlbaum Associates, Mahwah, NJ.

Wang, J. A., Knipling, R. R., and Blincoe, L. J. (1999) "The Dimensions of Motor Vehicle Crash Risk," *Journal of Transportation and Statistics*, Vol. 2, No. 1, pp. 19–44.

Ward, N. J. (2000), "Automation of Task Processes: An Example of Intelligent Transportation Systems," *Human Factors and Ergonomics in Manufacturing*, Vol. 10, No. 4, pp. 395–408.

Wickens, C. D., and Kessel, C. (1981), "Failure Detection in Dynamic Systems," in *Human Detection and Diagnosis of System Failures*, J. Rasmussen and W. B. Rouse, Eds., Plenum, New York, pp. 155–169.

Wickens, C.D., Rice, S., Keller, D., Hutchins, S., Hughes, J. and Clayton, K. (2009). "False Alerts in the Air Traffic Control Traffic Conflict Alerting System: Is There a Cry Wolf Effect?," *Human Factors*, Vol. 51, No. 4, pp. 446–62.

Wiener, E. L. (1989), "Human Factors of Advanced Technology ("Glass Cockpit") Transport Aircraft," NASA Contractor Report 177528, NASA Ames Research Center, Moffett Field, CA.

Wilde, G. J. S. (1988), "Risk Homeostasis Theory and Traffic Accidents: Propositions, Deductions and Discussion of Dissension in Recent Reactions," *Ergonomics*, Vol. 31, No. 4, pp. 441–468.

Wilde, G. J. S. (1989), "Accident Countermeasures and Behavioral Compensation: The Position of Risk Homeostasis Theory," *Journal of Occupational Accidents*, Vol. 10, No. 4, pp. 267–292.

Woods, D. D. (1987), "Commentary: Cognitive Engineering in Complex and Dynamic Worlds," *International Journal of Man-Machine Studies*, Vol. 27, pp. 571–585.

Woods, D. D. (1991), "Nosocomial Automation: Technology-Induced Complexity and Human Performance," in *Proceedings of the 1991 IEEE International Conference on Systems, Man, and Cybernetics*, Charlottesville, VA, pp. 1279–1282.

Woods, D. D. (1994), "Automation: Apparent Simplicity, Real Complexity," in *Human Performance in Automated Systems: Current Research and Trends*, M. Mouloua and R. Parasuraman, Eds., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 3–17.

Woods, D. D., Potter, S. S., Johannesen, L., and Holloway, M. (1991), "Human Interaction with Intelligent Systems: Trends, Problems, New Directions," CSEL Report 1991-001, Ohio State University, Columbus, OH.

Woods, D. D., Patterson, E. S., Roth, E. M., and Christoffersen, K. (1999) "Can We Ever Escape from Data Overload? A Cognitive Systems Diagnosis," in *Proceedings of the Human Factors and Ergonomics 43rd Annual Meeting*, Vol. 1, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 174–179.

Yeh, M., and Wickens, C. D. (2001), "Display Signaling in Augmented Reality: Effects of Cue Reliability and Image Realism on Attention Allocation and Trust Calibration," *Human Factors*, Vol. 43, pp. 355–365.

Young, M. S., and Stanton, N. A. (2007), "Driving Automation: Learning from Aviation about Design Philosophies," *International Journal of Vehicle Design*, Vol. 45, No. 3, pp. 323–338.

Zhao, X. D., and Xie, J. X. (2002), "Forecasting Errors and the Value of Information Sharing in a Supply Chain," *International Journal of Production Research*, Vol. 40, No. 2, pp. 311–335.

Zhou, T. S., Lu, J. H., Chen, L. N., Jing, Z. J., and Tang, Y. (2003) "On the Optimal Solutions for Power Flow Equations," *International Journal of Electrical Power and Energy Systems*, Vol. 25, No. 7, pp. 533–541.

Zuboff, S. (1988), *In the Age of Smart Machines: The Future of Work Technology and Power*, Basic Books, New York.

# CHAPTER 60

# HUMAN FACTORS IN MANUFACTURING

**Dieter Spath, Martin Braun, and Katrin Meinken**
**Fraunhofer Institute for Industrial Engineering**
**Stuttgart, Germany**

## 1 MANUFACTURING

### 1.1 Basic Definitions

*Manufacturing* is a human-driven transformation process. Using energy and manpower, this process creates consumer goods of economic value from naturally or rawly produced materials (Westkämper and Warnecke, 2002). The manufacturing area is more and more moving away from force-focused physical activity in favor of cognitive control activity. The main challenges coming with this change will be addressed in this contribution regarding the ergonomic work design.

Manufacturing is a part of production which has a major function in enterprises. It includes the tasks of production, assembly, logistics, planning, control, maintainance, and quality management. Additionally human resource management needs to be considered. Comprehensive production contemplation refers not only to technological and organizational aspects but also to the work's social and cultural values. The production process can only run optimally if integration of these factors has been achieved (Spath, 2003).

### 1.2 Historical Overview of Production Engineering

Three "industrial revolutions" characterize the historical development of production engineering (see Figure 1):

- During the *first industrial revolution*, power machines were developed. Characteristic for this period is the aspect of energy conversion.

- During the *second industrial revolution*, the organization of production was at the forefront. As far as organizational work measures, time became a most important production factor.

- During the *third industrial revolution*, information technology and automation were developed. Characteristic for this era is the application of information.

### 1.2.1 Early Industrialization

For decades production systems have been influenced by the "principles of scientific management" of Taylor (1911). Taylor examined the effects of monetary incentive systems and work division on performance. The basic assumption was that the average worker is motivated to work efficient mainly by financial aspects. This assumption leads to a consequent division of manual and mental work. In this way, the production system became independent from the know-how of the skilled worker. Complex tasks divided into sufficiently small subtasks could be done by almost everyone. This division of work called for responsible management, which today is still an influential characteristic of many businesses. The most important characteristics of the Tayloristic work structure are:

- Division of planning and implementing tasks
- Individual incentives (e.g., piece wage)
- Hierarchical system
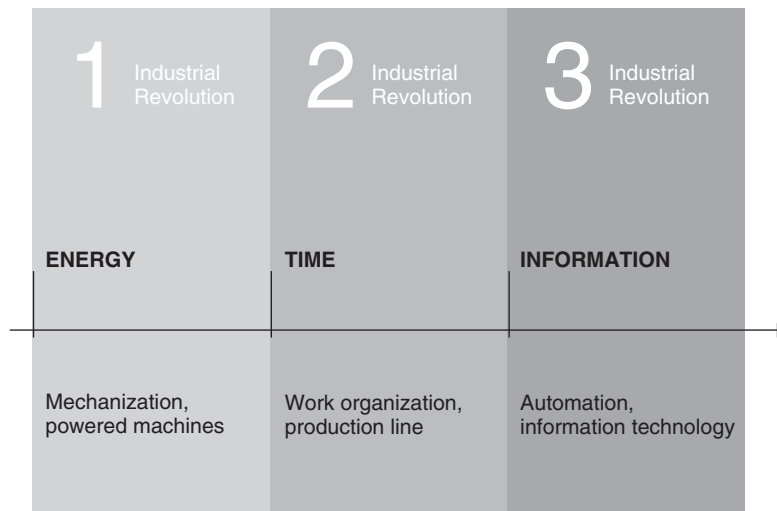
**Figure 1**  History of industrial development.

- Low work content
- Low skill levels
- "One best way" for every work sequence

It was not until the middle of the twentieth century that there was a gradual turn from the Taylorism division-of-work structure. In the 1960s, team work was used in place of these concepts. Team work calls for the individual to strive toward autonomy and self-realization.

The complex nature of these challenges required appropriate work design strategies. Although numerous promising design concepts were developed, it is to be noted that the obtained findings were implemented only insufficiently.

### 1.2.2  Present Positions

Production engineering is significantly influenced by progressive technical development, dynamic market conditions, and individual values regarding his or her work. At the same time, the technical organisational and human factors influencing the production process are not seen as being isolated from each other but rather influence each other. The most important factors influencing production design will be discussed below.

***The Human***  Over the past few years, the application of flexible production systems has become more widespread as a result of small lot sizes and an increasing variety of products. A high level of mechanization as well as intense informational relationships is characteristic of flexible production systems. Modified working situations for the human arise: While physical load was at the foreground earlier in time, the human today has to face additional psychological stress (Braun, 2008). Examples are as follows:

- Maintenance and surveillance of large-scale plants with increasing complexity
- Attention when working with dangerous materials or during dangerous processes

- Exposure to industrial robots
- Working with computer networking systems
- Innovative information systems
- Management and leadership paradigms (e.g., management by objectives)

Despite the change in working conditions, physical strain still plays an important role. Physical strain results, for example, from the manual handling of loads, by unfavorable or forced body movements, and postures which can lead to health hazards.

Additionally, companies are confronted with the effects of demographic development, that is, a partially overaged population of workers. In the future production tasks will be predominantly carried out by older employees. Therefore, maintaining health and qualification is more important than ever (Kern and Braun, 2006).

***Market***  The buyer's market has confronted manufacturing businesses with new challenges. Product life cycles are shortened and the time to market is constantly decreasing. These factors cause increases in employees' complaints about time pressure and stress.

***Technology***  In manufacturing, increasing automation and use of efficient information technology (IT) are particularly important. The problem is how individual work can be designed through existing and developing technology operations. It is assumed that the planning and operation of production systems need a paradigm shift toward a balance between individual and technical factors (Bullinger, 2003).

### 1.2.3  Perspectives in Manufacturing

The market-driven development and introduction of new products are crucial in order to obtain competitive economic advantages. At the same time, increased efficiency must be promoted. These can be achieved by the strategies discussed below (Westkämper, 2004).

***Time to Market*** In order to attain and use innovations with the goal of having a timely advantage over competitors, the manufacturing processes must be effective. Methods and tools have to be supplied in order to reduce the time as well as the cost of development and planning (Spath et al., 2003). As seminal are regarded:

- The synchronization of product and production development
- The market orientation and objective orientation considering the entire process chain
- An optimization of product development at an early stage
- Integrated management systems for engineering and production, including generative processes for prototyping

***Flexible Enterprise*** Customer orientation is a strategic success factor. The fulfillment of customer demands and short supply of products are preconditions for a substantial economic benefit. In dynamic markets only a flexible and adaptive organization delivers justifiable economic results. Above all, the resources (i.e., manufacturing resources, employees, material) must constantly be adjusted to demand. Currently, reaction times are often in the middle range. Bureaucracy and long logistical routes hinder flexibility.

Short-term production structure flexibility can be achieved by production networks, self-organization, self-optimization, and the use of intelligent production methods. Buzzwords such as "agile manufacturing" characterize the discussion of product paradigms which cause a higher dynamic and flexibility.

### Performance and Precision in Manufacturing

Over the past few years, performance has improved in many areas of manufacturing technology. Starting points are the usage of information systems for process management, self-learning, quality control, just-in-time systems, and machine and system diagnostics. The development of materials, sensor systems, and actuating elements as well as the awareness of interactions between process parameters and achievable performance and precision is of significant importance. At a high technological and performance level significant time and cost savings can only be achieved by the cooperation of skilled individuals.

***Automation and Humanization*** The idea of full automated enterprises existed for years, inspired by the development of automation technology. The supply and disposal of workplaces with material and information processing from design to implementation were completely automated. These assumptions failed due to the high complexity of automated systems.

As a result, numerous enterprises changed their strategy by emphasizing the qualifications of employees. In this way, effective and increasing short-term performances could be achieved. At the same time, enterprises learned how to optimize operations and control processes by applying human resources better.

***Adaptive Production*** Not only is adaptive production affected by turbulences in the markets but also these turbulences should be used to develop a competitive advantage. Adaptive production systems have the following characteristics:

- Information, as a production factor, is becoming increasingly important.
- Economic growth is achieved more by system considerations than by individual optimization.
- Employee qualification development is becoming more important.
- Team work is being increased.
- The implementation and development of human resources become central issues.

## 1.3 Production Systems

*Production systems* were developed in order to deal with expected operational challenges. Their appearance is a result of an integrated industrial application of various methodical approaches.

### 1.3.1 Terms and Definitions

A *system* is an interrelated combination of elements or procedures which either exist in nature or are artificial. A system and its subsystems are confined by the system boundaries. Combinations and interactions within a system are usually described in a process-oriented manner.

Systemic design requires that elements and methods are connected to each other. Therefore, an optimal material and information flow is an important consideration. The output of one method serves as an input for the next method. Integrating material and information flow results in optimizing transfer processes concerning time and quality.

When dealing with production systems the individual, the organization, and the technology should be taken into consideration (Spath, 2003). Moreover various management concepts have been integrated (see Figure 2).

A production system covers at least the production process, which includes functions of planning, manufacturing and assembly, control, logistics, and quality control. These business processes are often expanded to integrate, for example, the acquisition process, the research and development process, and the sales process.

### 1.3.2 Design Principles

Production systems are adapted specifically to markets, products, technologies, and corporate cultures. The systematic application of design principles will ensure that all system elements fit together, resulting in a coherent system. Next we discuss the principles for designing production systems (Scholtz et al., 2003).

***Autonomous Teamwork*** Comprehensive and challenging tasks are often fulfilled best if the team is working autonomously. The team takes on planning tasks such as work distribution, the order in which assignments are processed, appointment compliance,
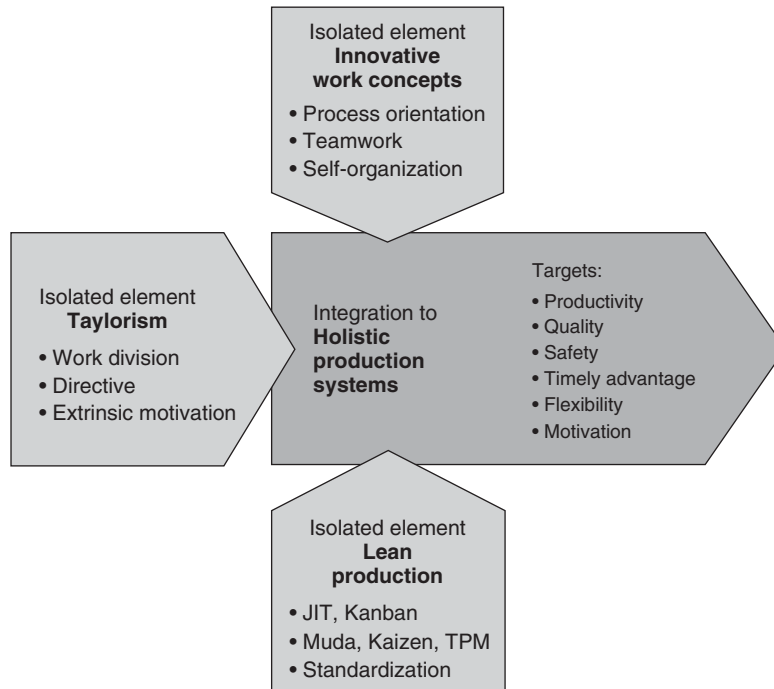
**Figure 2**  Development of integrated production systems (Korge and Scholtz, 2004).

acknowledgement of lot size per time unit, material disposition, and so on. In addition, they adopt tasks such as machine maintenance and repair, cleaning and transportation tasks, and quality control. Qualification activities ensure that group members are able to fulfill these tasks. Members more inefficient than others are not pushed to the edge in this system. The concept corresponds to a long-time existing demand to create working conditions which enhance an employee's development potential. Figure 3 illustrates a group working at an ergonomically and functional optimized workstation.

***Process Integration***  Process integration is defined as enabling the cooperation of production functions, which replaces the consideration of isolated functions. Process integration implies that production reacts to supply and demand. Information systems support process orientation (e.g., logistical tasks).

***Just in Time***  Just in time (JIT), that is, the concept of "timely delivery as required" is an essential part of process-oriented work. This concept is related to having limited storage capacity. Minimal capital is tied up for the storage of parts, and only the amount of product needed in one process step is provided in a specific time frame. The JIT concept includes the suppliers and only the required parts are produced at the right time in the desired quality and free of waste. The most important methods and concepts of JIT are:

- One-piece flow (piece production)
- Principle of "first in, first out (FIFO)"
- Small-sized transport equipment



**Figure 3**  Group work assembly operation at an ergonomically optimized workstation. (Courtesy of Fraunhofer IAO.)

- Rapid changeover
- Kanban

***Continuous Improvement Process***  The continuous improvement process (CIP) is a fundamental element of work organization. The attitude is to be always

in search of better problem solving. All workers are asked to supply ideas for the process improvement.

The concept of CIP originated in Japan where it is called "Kaizen," which means "getting good through a thousand little improvements." In contrast to the established employee suggestion system, CIP encourages all workers to make even minor recommendations. Basically, it is an attitude that directs employees to critically consider existing processes, analyze them, and find solutions. Improvements are communicated to all concerned.

CIP can be a fixed component in teamwork, for example, a quality circle. Mostly the employees are motivated by incentives to provide ideas. It is important to note that the improvement process might lead to disadvantages for the employees (e.g., job elimination). To motivate employees to participate in the CIP, they should be made aware that no disadvantages can arise from their participation. Therefore, the CIP should be aimed not at cost savings, personnel reduction, and work compression but at improved work organization, employee qualification, ergonomic work design, stress reduction, and so on.

***Professional Work Routines*** To maintain results from the CIP and to prevent, for example, the so-called brain drain, professional work routines are used. They define the manner and protocol by a work process is carried out. In this way professional work routines ensure a sustainable implementation of improvements.

Important concepts and methods regarding professional work routines are standardized working papers, standardized shift change, standardized equipment, standardized quality problem operations, and defined rhythms for preventive maintenance and documentation.

The standardization aims to improve the process management. It is wise to reduce the number of solutions to as few as possible. The necessary degree of flexibility with respect to the customer needs may not, however, be inadequately confined by standardized solutions.

***Target Management*** Target management means making agreements with individual employees or teams about the expected performance during a given period. After the given period, it is decided whether or not these goals have been reached. Many factors can be stipulated ahead of time, for example, the amount of revenue expected during a given period, the date until which a product is to be manufactured, and when an assignment has to be completed. Similarly, there are stipulations regarding employee performance, for example, in order to increase operational readiness. The extent and degree of the achieved objectives form the foundation for further personnel decisions, such as promotion prospects, transfers, and cancellations. In combination with performance management the target management forms an instrument for performance control. A main characteristic of target management is that employees decide nearly autonomously how to reach the goals.

Target agreements should be commonly developed to create mutual trust of the employee. The employee knows why reaching these goals is of importance for the business. Goals should not be inconsistent or too detailed; rather they should be important, plausible, and easy to evaluate.

***Robust Processes*** Robust processes aim to deliver products and benefits reliably and without errors (e.g., "zero-error goal") to the customer. This task can only be reached by a preventive quality policy. Preventive methods should help to identify and avoid errors before the start of production. By analyzing the source of the error and eliminating it, errors are prevented from reoccurring. Well-known robust process concepts and methods are the quality circle, account force diagram, marginal samples, quality alarm, machine breakpoint, total productive maintenance (TPM), poka yoke, failure mode and effects analysis (FMEA), and quality agreements.

## 1.4 Human Factors in Production

The significance of human factors can be appraised by the workers' contributions to the economic success of a businesss enterprise.

### 1.4.1 Economic Requirements

Businesses have an economic interest in keeping healthy, motivated, and qualified workers because they will have a positive effect on production efficiency. Frustrated, passive, or aggressive workers are less motivated and at the same time are susceptible to diseases. Diseases and accidents weaken an enterprise's productivity. Business surveys show a strong correlation between loss time and personnel costs. An appropriate compensation of absent personnel is in many sectors not possible without conflicts, which can lead to loss of flexibility, lack of quality, and production disruptions and in the end to assignment loss. In the course of flexible strategies and a decomposition of decreasing personnel compensation buffers, the economical consequences of absenteeism and insufficient employee engagement are becoming increasingly important (Braun, 2003).

### 1.4.2 Motives for Performance and Cooperation

Employees feel motivated to do good work if their personal interests agree with those of the company and their performance is acknowledged. If there is disagreement during the course of changing processes, resistance may develop. Resistance to change leads to considerable loss of performance and cooperation. As a result. managers of major enterprises expend between 50 and 80% of their working time to override internal constraints (Spath et al., 2003). A survey of approximately 2000 employees taken by Gallup (2010) in Germany shows that about 80% of those interviewed were insufficiently involved in their company. Approximately 70% of the interviewed employees lacked engagement and were passive when dealing with their managers. Fifteen percent of those interviewed were often aggressive toward their superiors, displaying this by poor productivity. Such negative attitudes resulted in:

- A high number of absences
- Readiness to leave the company as soon as an opportunity arises
- Nonexistent career intentions with current employer

- A bad attitude toward others
- A small chance of recommending the workplace to friends or relatives
- A small chance of recommending the product
- No enjoyment at work

As a result of lower productivity and high number of absences, the macroeconomic loss resulting from unengaged employees was over 10% of the yearly gross national product. Reasons for the lack of dedication are obscure expectations of managers as well as an insufficient acceptance. Also most interviewed employees had the feeling that their opinions and ideas were not acknowledged by their manager.

### 1.4.3 Focusing on the Healthy Individual

When discussing cooperation, flexible reaction, and innovation, the worker is irreplaceable due to his or her creativity and communicative capabilities. The employees' interest and will to change are essential factors for every increase in productivity. Many enterprises have recognized that healthy, qualified, and efficiency-motivated employees are one of their main assets.

Empirical comparative studies support that technology does not play such a meaningful role during the development of an enterprise. Actually, technology does not trigger innovations but merely expedites them (Collins and Porras, 1994). A successful enterprise has to ensure reliable orientation and demanding motivation for its employees during a change process. Experience has shown that problems such as motivation, engagement, and diversification are often solved when the work fits the preconditions of the human.

From the above-mentioned production concepts, it can be concluded that the process of making businesses automated and flexible requires consideration of human-oriented work design.

## 2 HUMAN-ORIENTED WORK DESIGN

### 2.1 Objective Target

The goal of human-oriented work design is to balance the strain of the worker. By doing this, human performance potential for the production of goods and services is utilized and it is balanced against the premature wear of performance capabilities. This involves the use of technical, medical, psychological, as well as social and ecological knowledge (Bullinger, 1994). It is to be noted that the scientific concepts of work design have been researched since the 1970s (Helander, 1995; Karwowski and Salvendy, 1998; Karwowski, 2001; Salvendy, 2006). Therefore it is supposed that the reader has sufficient knowledge of these human-oriented design concepts. Since their implementation is still of basic relevance, these concepts will be presented shortly within this chapter and referenced accordingly.

Human-oriented work design deals with the worker, the detection of his or her skills and abilities, and the analysis of variables that may influence performance. Further tasks concern the design of technical equipment and organizational structures used for work. The goal is optimal customization of equipment and structures to the workers' abilities and skills. This results from three operational and design approaches:

- Fitting the work to the human by designing working conditions
- Fitting the human to the work through qualification and job assignment
- Fitting the workers with each another; this can only indirectly result from organizational and technical work design

A five-stage target system is used for the design and evaluation of human-oriented work. The criteria are interdependent since criteria of lower levels must be achieved before criteria of higher levels can be applied. The criteria are (Luczak et al., 1987):

- *Without Damage*. Work has to be tolerable and free of damage. Also long-term effects have to be considered. The possible damage to workers during their lifetime must be assessed over a period of time. Work time, work intensity, and environmental conditions are particularly significant.
- *Feasibility*. Work tasks, particularly tool handling, must be feasible. Individual abilities and skills may lead to different kinds of strain. For this purpose, limits are generally established by human biomechanics or available mental capacity.
- *Reasonability*. Reasonableness is an individual factor which can only be answered by each employee. The individual experience, however, is related to the cultural environment and previous know-how. According to this criterion, workers should have some latitude over their work task design and work environment.
- *Satisfaction*. Work should be satisfying and supportive. Through work design this can be reached by considering the psychological and cultural environment. Here acknowledgment, motivation, reward, and superior leadership behavior are to be considered.
- *Social Compatibility*. Social compatibility means that the employees are involved in work design relating to their cooperative organization. Task-oriented work structures implement this criterion, particularly since the employees are involved in the work design process.

### 2.2 Humanization and Rationalization

Human-oriented work design similarly follows humanization and rationalization objectives. *Rationalization* describes the substitution of inherited procedures by using more practical and better thought-out procedures.

*Humanization* refers to the civilization and design of the workplace regarding the well-being of the worker. Work design follows both humanization and rationalization objectives of work systems. For this, it is necessary to find compatible conditions for both objectives (Kern et al., 2005).

In past years, under the premise of humanization, enterprises increased their efforts to accommodate the work organization to the growing demands of the individuals for larger task variety and self-organization. Thus the worker should be motivated by his or her work. As a result, enterprises hope for better product quality through a decline in absenteeism and fluctuation. Enterprises ultimately expect a better level of quality of their business processes.

## 2.3 Strategies of Human-Oriented Work Design

Focusing on the worker the following criteria should be implemented:

- Feasibility: anthropometrical, psychophysical, and biomechanical limits for short loading times
- Tolerability: physiological and medical limits for long loading times
- Reasonability: sociological, group-specified, and individual limits for long loading times
- Satisfaction: individual social–psychological limits with long and short valid times

These explanations make clear that the two first-mentioned criteria, feasibility and tolerability, are achieved by measures of engineering, while the other two criteria, reasonability and satisfaction, are achieved by psychological approaches. However, it has to be noted that both sets of approaches cannot be separated from one another.

When considering a manual assembly operation, feasibility limits arise for the worker which result from required movement speed and movement accuracy. In Figure 4 this situation is illustrated. At the point where the combination of the resulting work task's stress parameters does not lead to a feasible work situation, automation has to be implemented. Where work is achievable but not tolerable, the work content must be designed, that is, the work must be restructured. Ergonomically optimal conditions are aimed for where tolerable combinations exist.

The three design approaches differ when talking about humanization and rationalization. Thereby, the limit in which work is achievable but is not tolerable comprises most of the problems. The scheme also can be adapted to informational work.

### 2.3.1 Ergonomics

The aim of ergonomics is to protect the worker from impairments, especially by eliminating influences which constrain performance or cause physical damage. So far, good results regarding this task have been achieved. Particularly, anthropometric *workplace design* methods are very well engineered. Therefore, it is surprising to still find workplaces where important measurement questions remain unanswered. Computer-aided human models among others are installed for workplace design. These human models also include anthropometric and biomechanical modules and databases for posture evaluation (see Figure 5).
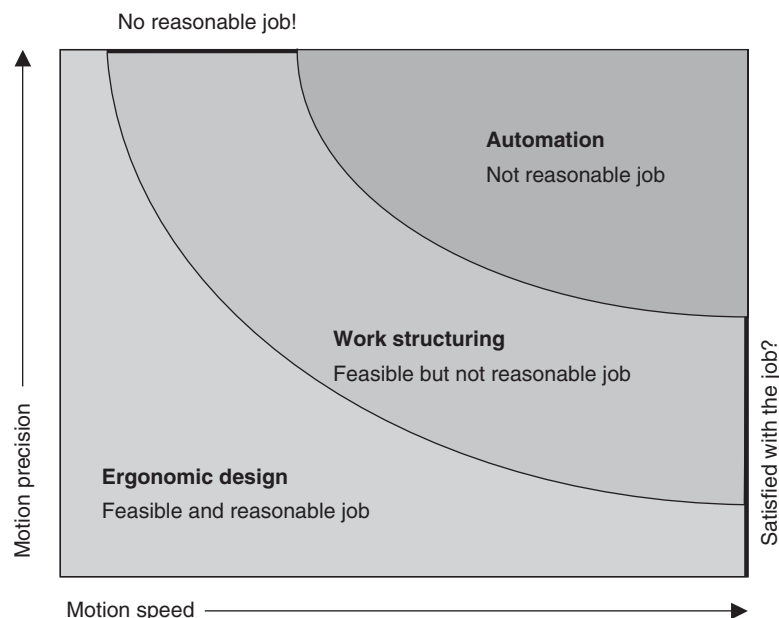


**Figure 4** Dimensions of work design strategies.

**Figure 5** Anthropometric workplace design using the virtual human model ANTHROPOS on a Picasso-3D-Projection System. (Courtesy of Fraunhofer IAO.)

*Product ergonomics* deals with ergonomic product design with respect to the task to be performed. This is particularly important for consumer products and when using hand-held tools in order to have optimal performance. In this case the ergonomic design complements the functional design (Bullinger, 1997).

Due to multiple requirements, ergonomic work design has become increasingly complex. This complexity can only be accomplished using an integrated method (see Figure 6).

Integration takes place in three dimensions: (1) Integration of requirements is comprised of all relevant design influence factors and requirements. (2) In methodological integration, it is essential to define and combine established procedures and methods during the design phases. (3) The process and communication for a design project will ultimately be adjusted in line with organizational integration.

### 2.3.2 Work Structuring

Work structuring involves the organization of work as well as their requirements so that the work contents comply with the capabilities and skills of the employees (Bullinger and Braun, 2001). The definition of work structuring allows different interpretations that are dependent on the objectives concerning humanization and rationalization, including:

- Solving economic problems (such as insufficient flexibility, poor production activity, lacking quality)
- Solving personnel problems (such as high numbers of fluctuation, lacking work morale, signs of dissatisfaction) Redesigning technical systems and the operational organization

At first glance, it appears as if economic and human problems are handled the same way. But an exact analysis of human-oriented problems shows that their removal also enhances performance. Therefore, human-oriented design measures always concern economic efficiency motives.

### 2.3.3 Automation

The goal of automation is to transfer functions done by humans to machines. The level of automation is measured by the number of functions previously performed by humans being done by machines. The prosperity in industrialized countries is due to the productivity increase through automation. This fact cannot be denied in the discussion about the consequences of automation. However, it accounts for the increasing need to work on solving automation-related problems.

Finding an automation strategy that is acceptable under human and economic aspects is difficult. The problem becomes clear once it is considered that, because of economic and technical reasons, automation
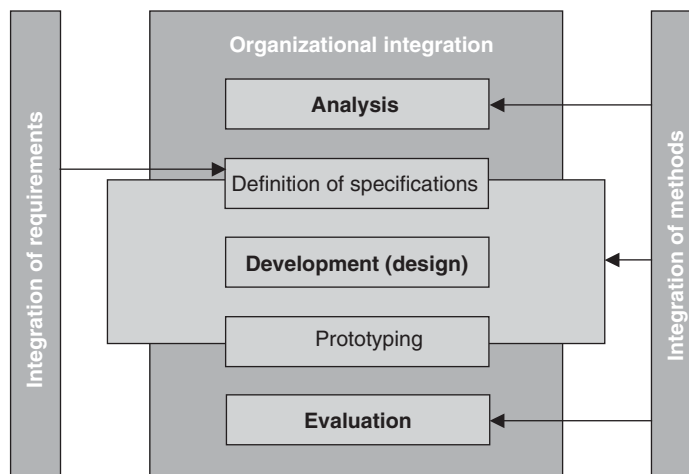


**Figure 6** Aspects of integrated ergonomic work design.

can only occur in a way that would make humans unnecessary. In many cases monotonous tasks remain for humans since machines can carry out the more complicated work with higher speed and precision. Also a workplace with ecological damage will expose humans to possible hazards.

It is obvious that an automation strategy has to consider the work structure. Thereby, the planning should emphasize the work structure while automation constitutes an alternative solution to the problem. This is away to avoid keeping undemanding tasks in highly mechanized structures.

Another basic requirement is to improve the automation technology in a way that workers are relieved from specific tasks, especially working in a straining environment. Using automated handling machines is an example. The effects of the humanization of automation are as follows (Spath et al., 2009):

- Disappearance of monotonous tasks
- Disappearance of heavy physical strain, resulting from unfavorable body position and exertion, for example, when lifting heavy loads
- Disappearance of unfavorable environmental influences, for example, through heat, filth, and noise
- Decline of accident risk

## 2.4 Work System Design

The systemic design approach was introduced in Section 1.1. To comprehend and design complex systems, the use of systemic models has proven valuable (Spath and Dill, 2002). Basic principles of systemic work design relating to complex production systems will be subsequently discussed.

### 2.4.1 Work System and Its Elements

The human, the workplace, the work environment, and the work organization all shape the production process. They are linked to each other in order to build the work system. This systemic model is able to provide systematization for analysis, planning, designing, and evaluation measures. It can be applied to an individual workplace (first-order work system) but also to a network enterprise (nth-order work system). The work system and its single elements as well as its related design approaches will be subsequently presented (see Figure 7).

Basic work system elements are:

- A *work task* is an assignment for humans to exercise a job which serves to achieve objectives. It illustrates the purpose of the work system. The work task is often referred as the target work result.
- *Tools* are equipment, machines, organizational tools, and so on, which are in any way involved in a work system in order to accomplish the work task. From a systemic point of view, tools are inactive elements.
- The *workplace* is a spatial area where one or more people are designated to accomplish a task.
- Physical, chemical, biological, social, and cultural conditions are identified by the *work environment*, which surrounds the workplace. These conditions influence the system behavior and the features of the elements.
- The *human* is the active element of the work system. It is only the human who is able to bring other system elements into action.

The traditional work system design methods are primarily oriented on isolated system elements. The
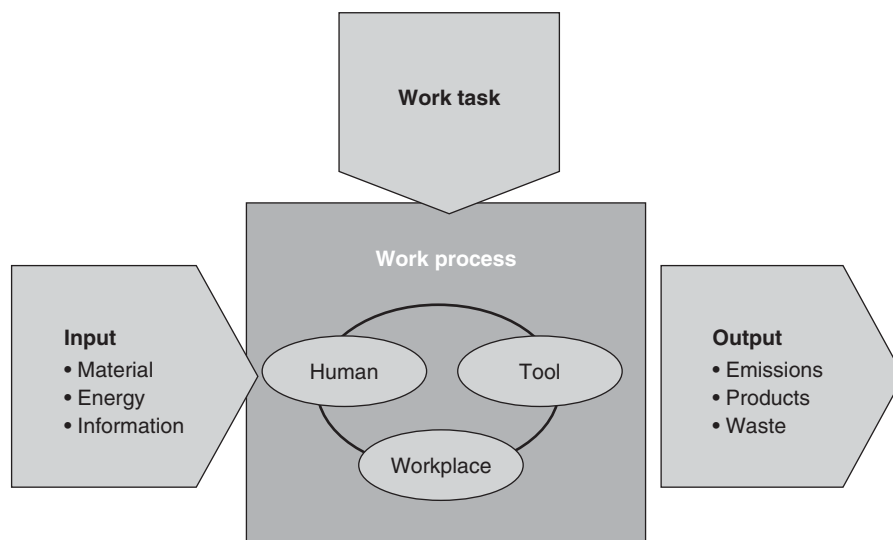


**Figure 7** Work system and its elements.

previous remarks make clear, however, that only integrative work design methods in a systemic context can lead to human-oriented and economic effective work systems.

### 2.4.2 Design Methods

Work system design aims for an optimal interaction of human, tools, and tasks. This goal is to be accomplished by consideration of the worker's capability as well as his or her individual and social needs. By minimizing impairments, human-oriented work design helps to continuously promote the worker's health and performance. Furthermore, to improve work system effectiveness, work design has the goal to increase efficiency and reliability by optimizing human–machine interfaces and by influencing the behavior at work.

In the following, starting with the system model, basic methods for work design will be presented. First, single work system elements will be considered. Subsequently, information on the design of the complete work system will be given. Thereby, the interaction of system elements is to be taken into consideration: The change of one element has an influence on all the other elements. Closer considerations of the presented methods are to be found in the literature (Helander, 1995; Luczak, 1998).

### 2.4.3 The Worker

The human, being the active element of the work system, holds particular importance. An effective work system is always a human-oriented system. In order to design a human-oriented system, basic function characteristics of the human must be identified. Here physical and mental dimensions are to be differentiated, even though these two dimensions are closely related to one another.

***Stress and Strain***   The simplified stress–strain model, presented in Figure 8, is an appropriate model to describe stress (i.e., workload), individual capabilities, and strain of the worker.

The term *stress* defines all external demands on humans resulting from the workplace, the work process, and all physical environmental influences. The term *strain* is defined as the reaction of the organism on stress which is affected by individual human characteristics.
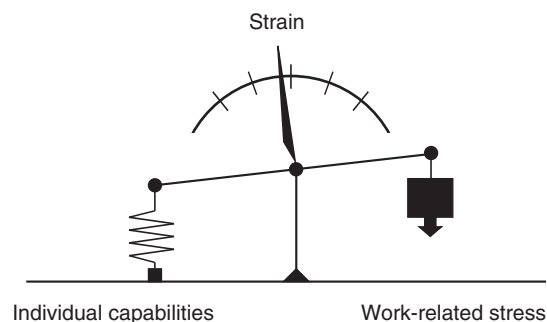


**Figure 8**   Relationship between stress and strain.

The individual capability is the factor that combines stress with strain. Human physical and mental capabilities are, however, not constant variables but may change.

The total amount of human stress at work results from the level of stress as well as the duration of exposure. Stress is divided into stress that is measurable quantitatively and nonquantitavely. Quantitative stress can be measured by using physical methods. Nonquantitave stress can often only be described.

A direct strain measurement is not possible since every stress can result in a different strain with respect to different people. However, the analysis and quantification of strain are important so that:

- Work tolerability can be evaluated.
- (Durable) capacity limit values can be determined.
- Critical strain reactions can be avoided during work design.
- Physiologically adequate design of rest time is possible.

Strains can indirectly be measured by physiological parameters (e.g., frequency of heart beat, hormone concentration in blood), performance analyses, and subjective techniques (e.g., standardized questionnaires). The goal of human-oriented work design is to create an adequate balance between strain overload and underload.

***Capability and Motivation***   Physiological and psychological capacities define human capability (Luczak, 1998). Human capability fluctuates for each person and between individuals, increasing with training and declining with fatigue. An appropriate break design maintains the capability during work (Spath et al., 2003). Human motivation is no less relevant for efficient production. Even well-educated employees bring little added value when they are not willing to perform, that is, they are not motivated (see Section 1.4).

Numerous existing models describe capability and motivation. At this time, the psychological *theory of mental regulation of action* should be mentioned. This regulation theory predominantly relates to work tasks. Therefore, it refers to observable and conscious execution processes. Unconscious mental processes, for example, formation of opinion, remain unconsidered. The theory of mental regulation of action assumes that human action is target oriented, that it corresponds to external matters, involves social connections, and shows process characteristics. Consequently, the theory describes running processes of action, from setting the goal, and presumes an interaction of mental processes and observable activity. Hereby, two regulation procedures are differentiated, as shown in Figure 9 (Hacker, 1998):

- Stimulus regulation, that is, it is determined when an action is conducted. Individual motives, attitudes, and preferences are the basis for these procedures.
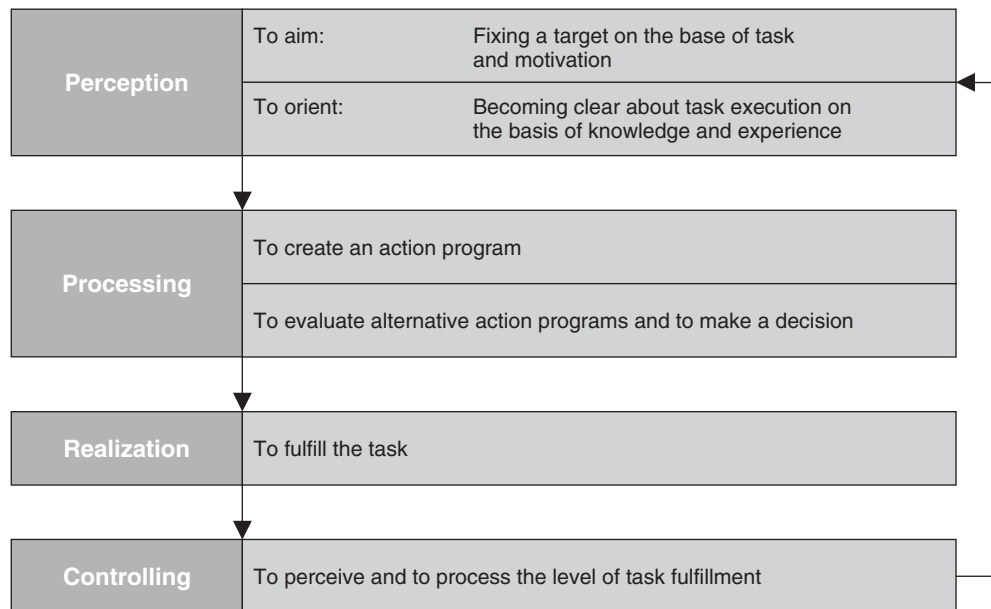
**Figure 9** Structure of the mental process.

- Execution regulation, that is, it is determined how an action is done. Individual subgoals, mean-methods-choice, and execution control are the basis for these procedures.

According to this model, human action is a control circuit which is significantly influenced by motivation, knowledge, information, and experience.

The theory of mental regulation of action assumes that task requirements and stress are independent of one another. Psychological requirements should give the worker the possibility to independently decide goals and methods. The most important influencing factors are task variety and communication. Mental stresses at interfaces between the individual and the organization are defined as restraints. Regulation restraints describe discrepancies between working conditions and goal attainment as well as excessive demands, including time pressure and monotonous jobs.

This follows the simple realization that people who perform their work as a result of their inner motivation and conviction are considerably more productive than those who are only subjected to work. It shows that *motivation* is a deciding factor in the relationship between the individual and the work. Motivation is the sum of action, behavior, and behavioral tendencies. Contrary to the biological stimulus of humans, motivation and individual motives can be learned and respectively arranged in the socialization process. Therefore, for greater efficiency it is important to encourage employees.

Motivational theories only explain part of human behavior within the stimulus–reaction scheme. Because of their diffusion, the theories of Maslow (1987) and Herzberg et al. (1967) are mentioned.

Maslow (1987) established a hierarchy of individual requirements. A requirement level must be fulfilled for the next level of requirement to become relevant in terms of motivation. Thus, according to this theory, it makes little sense, for instance, to motivate the employee to self-realization using a general offer of further training when the security of their position (e.g., resulting from economic reasons) is not given.

Herzberg et al. (1967) differentiate between hygiene factors, whose nonobservance leads to work dissatisfaction, and motivators, whose compliance leads to work satisfaction, respectively motivation. Thus, poor internal enterprise politics lead to intense work dissatisfaction, while good internal politics prevent this from happening. This, however, does not conversely lead to work satisfaction. Acknowledgment from colleagues has strong motivating effects, but lack of acknowledgment does not necessarily lead to dissatisfaction.

For business practices, the following design concepts are derived from psychological theories:

- *Reduction of Time Constraints*. Using buffer banks between individual workstations, employees have the opportunity to work detached from the fixed-time work cycles.
- *Job Rotation*. The work content of individual jobs does not change, but employees exchange their workplaces systematically.
- *Job Enlargement.* The job content is increased. Employees are given more similar assignments of the same qualification level, leading to longer time work cycles.
- *Job Enrichment.* The job content is changed in a way that individual employees have a greater

task variety, resulting in higher qualification requirements.

- *Autonomous Team Work*. A job assignment is given to the work team. This assignment is separated into subtasks within the team. The team can organize the work itself within certain limits (i.e., time targets, technically marginal conditions). This type of work structuring offers a good chance for individual work design. However, it holds the risk of social conflicts arising within the group.

### *Consequences of Unbalanced Workload*

Fatigue, monotony, mental saturation, reduced vigilance, and stress are closely related to the problem of strain. These categories describe both a procedural happening and an internal condition of the human.

*Fatigue* describes stoppage of motivation, resulting from a continuous task performed over the course of hours up to a day. Considering the nature of the strain, it has to be differentiated between physical and psychological fatigue. Physical fatigue is attributed to displacement of the physiological–chemical balance of an organism. Psychological fatigue is an impairment of regulation. It is an aftereffect of psychologically demanding tasks which are characterized by information reception and processing. Monitoring, inspection, and control tasks, which are particularly found in process control, are mainly considered to be psychologically demanding tasks. For such tasks alertness adaptations and focus on defined task contents are typical. As a general rule, cognitively overstraining conditions bring sporadic functional ability and decreases in cognition, memory, and thinking. Fatigue-caused capability impairments can be temporarilyy eliminated by job rotation, environmental influences, or stimulants. They are completely eliminated only by sleep (Richter and Hacker, 1998).

*Monotony* is similar to fatigue but is generated by lack of stimulation or by conditions with minimal changes in stimulus structure. Symptoms of monotony are fatigue, sleepiness, aversion, and attention decline. The decrease in motivation and reaction ability associated with monotony is reflected by unsteady and decreasing performance. Monotony arises when the fulfillment of a job does not allow a complete solution and does not offer enough possibilities for a mental debate of the task [International Organization for Standardization (ISO) 10075; ISO, 1991].

According to ISO 10075, *mental saturation* is a condition of nervousness, restlessness. and intensely affective decline in an undemanding, repetitive task or situation. The concerned person perceives his or her job as being senseless; aversion and irritation are a result. The continuation of the task is carried out reluctantly. In the long run, psychological saturation leads to the employee's "internal termination" by refusing his or her own initiative and operational readiness. These sentiments are triggered by monotonous tasks which are under the employee's qualification level. Permanent troubles, unfulfilled needs, and unreached personal goals, however, also can lead to mental saturation.

*Diminishing vigilance* (also called hypovigilance) is a condition of reduced mental activation which results from little variation monitoring tasks. It is the consequence of qualitatively undemanding tasks with a highly passive work content, low amount of environmental stimulus, and lack of environmental diversification which result when concentrating on few inputs. The mental tension required for the deliberate balance of functional impairment with reduced vigilance presents an additional source of mental exhaustion. Hypovigilance systems, which recognize such conditions and support people accordingly, for example, by targeted activation, are being investigated and developed (Hagenmeyer, 2007).

Monotony, mental saturation, and diminishing vigilance resemble fatigue in that they result from an underdemand of human capability; they can be removed by expanding the task variety (Spath et al., 2003).

*Stress* is a condition of the organism which develops when the person has recognized that his or her well-being or integrity is in danger and he or she must use all available energy for self-defense (Cofer and Appley, 1964). All situations experienced as being unpleasant or threatening can trigger stress. At the same time, the amount of stress depends on the intensity and amount of time the person has been exposed to similar situations, the disposition, and the situational factors. The human is not able to support long-term stress. The physiological mechanism of stress reaction breaks down at the point when the stress factor cannot be removed by means of coping or avoidance for a short term. Long and continuous exposure to work-induced stress leads to excessively increased levels of alertness which do not sufficiently subside after work finishes. The consequences result in sleep disorders, impulse liability, and internal unrest. Long-term memory is weakened, and muscle activation including speech becomes aggravated. The responsiveness of perception is reduced. Consequences are erroneous actions, erroneous estimation, and anomiae.

### 2.4.4 Working Environment

Relevant factors of ergonomic working environment design are lighting, noise, mechanical vibrations, climate, harmful substances, and radiation. The sphere of influence for these factors is primarily in the direct environment of the workplace and in the used work tools. For example, the decrease in machine noise is primarily a constructive problem. Since the machine is, however, located at the workplace, its emission of sound also influences the quality of the workplace.

Some environmental factors are purposefully used for work design (see Figure 10). Others have undesirable effects. The goal is to reduce intensity, exposure time, and impact frequency in a way that avoids excessive strain. At the same time, eliminating all environmental influences can have disadvantageous consequences. As an example, psychic problems arise for inhabitants of apartments as a result of extreme isolation from outside noise. Complete elimination of noise and vibration emission while using an electrical razor irritates the user and might lead him to reject the product. In general,

| Environmental Factors | Potential desired effects | | Potential undesired effects | | | |
|---|---|---|---|---|---|---|
| **Legend:** ● Correlation ○ No correlation | Support of task operation | Well-being | Disturbance | Malfunction | Health damage | Bodily injury |
| Lighting | ● | ● | ● | ● | ○ | |
| Coloring | ● | ● | ○ | ● | | |
| Climate | ○ | ● | ● | ○ | ● | |
| Contact temperatures | | ● | ● | | | ● |
| Air/air quality | ○ | ● | ● | ○ | ● | |
| Noise | | ○ | ● | ● | ● | ● |
| Vibrations | | ○ | ● | ● | ● | ● |
| Acceleration forces | | | ● | ● | | ● |
| Moistness | | | ● | ● | ● | |
| Dirtiness | | | ● | ○ | ○ | |

**Figure 10** Environmental factors at the workplace and their effects (Bullinger, 1994).

environmental influences should not to be eliminated but be optimized.

The three environmental factors climate, noise, and lighting are most frequently identified as being problematic in manufacturing. These factors are discussed in the following.

**Lighting** Most human sensory perception happens over the visual canal. Appropriate lighting is required for visual information intake. By using appropriate lighting, performance and work safety can be enhanced and visual strain can be reduced.

Visual perception is directly dependent on light intensity. Therefore, light intensity must be adjusted to the visual task. Depending on the aging process, elderly people require more light in order to be able to carry out visual tasks precisely and rapidly. The work design must ensure that light intensity is sufficiently adapted to the conditions.

A visual object is only recognizable if it has a minimum contrast, that is, a luminous density difference within its environment. The capability to perceive contrasts depends on the object's size, luminous density, perception time, and level of adaptation. The higher the level of lighting, the greater the contrast has to be. If the contrast is too strong, then glare arises.

Direct glare results from directly glancing at a luminous source. Here the absolute light density value is often too high (e.g., when looking at the sun). Reflex glare is a result of luminous source reflections on reflective surfaces. Light density differences, which are too large in the visual field, that is, intense contrasts, lead to relative glare. Furthermore, relative glare causes eye strain which results from adaptation.

The contrast depends on the surface condition of the observed object, angle of light, and distribution of light density. Good contrast reproduction is achieved by dimmed material surfaces and by laterally arranged lighting above the workplace. The direction of light, especially at workplaces with visual display units, is important so that reflections do not occur on the visual display.

In practice, for lighting design the following are helpful:

- Adequate level of light intensity
- Harmonious distribution of light intensity
- Glare limitation
- High contrast reproduction
- Proper direction of light
- Accurate amount of shadiness
- proper lighting color and appropriate color reproduction
- High degree of energy efficiency

**Noise** Within the variety of physical environmental factors appearing as strain variables for the worker, noise is the most significant. In the last decades, noise

has become a public problem due to the high numbers of compensable occupational illnesses as a consequence of long-lasting noise exposure and emerging noise-induced hearing loss.

Continual loss of hearing results from recurring noise exposure or in rare cases from short exposure of high noise levels. Mental reactions such as disturbance and annoyance can already be observed at low noise pressure levels. These reactions mainly depend on the position of the concerned person to the origin of noise and his or her momentary disposition (e.g., mood and tension). At approximately 65 dB, there are reactions of the vegetative system, for example, a change in breathing rate. An irreversible effect on hearing loss is not excluded when noise exceeds 85 dB. Hearing loss makes it difficult to sense acoustic signals and speech. This can lead to higher accident risk and changes in behavior toward fellow workers.

Sound emissions and their effect on people are directly related. The effect of noise on people is assessed by the noise rating level of an 8-h work shift.

The goal of ergonomic work design is to prevent the development of noise. Basic *primary* measures are preferred in order to reduce noise. These measures prevent noise from developing, for example, by using electrical motors instead of pneumatics. Since these measures often have constructive machine and equipment demands, they are particularly considered during production planning and related machine construction; in ongoing production, primary measures are usually very expensive. In this case *secondary* measures, which prevent sound spread, are aimed for. For instance, noisy machines can be grouped in a separate space or machines can be encapsulated. *Tertiary* measures such as ear plugs should be used when all technically possible and economically justifiable efforts concerning noise reduction did not lead to a noise level below 85 dB.

***Climate***   Climate is an important environmental factor at the workplace. Its importance is a result of multiple interactions with the human. In fact, the number of workplaces under extreme climatic conditions has declined (e.g., foundry). But in the field of workplace design calls for a comfortable climate are increasing and ergonomic recommendations and rules are necessary.

Climate is not a consistent dimension; rather it is a generic term influenced by air temperature, humidity, air movement, and thermal radiation.

The various climate factors are integrated into one variable by a cumulative climate indicator. The principle of the cumulative climate indicator is based upon different combinations of three variables: air temperature, humidity, and air movement.

In principle, the human body tries to establish equilibrium between body heat and external climatic influences. The goal of this regulation is to maintain the normal temperature of approximately $37°C$ related to a certain level of comfort. Within low environmental temperatures, the lack of warmth is balanced by an increase of body temperature. Within a higher environmental temperature, the body tries to eliminate excess warmth in the environment by transpiration.

In the production environment *temperatures* between $-50°C$ and $+50°C$ can be found, whereas workplaces with more extreme thermal radiation (e.g., a foundry) are not considered. This climate variety underlines the need for adequate climate design.

For the climate assessment, subjective influence variables beyond those relating to environmental climate are also considered. Among these variables are clothing and work intensity as well as the condition and constitution of each person. In the assessment of climate comfort it should be considered that individual differences in climate perception can arise: For example, in a survey of normally clothed office workers ($n = 1296$), a majority of those surveyed felt a temperature of approximately $21°C$ to be neither too cold nor too warm. This is also considered to be a state of neutral temperature. The interesting result of the survey was that only 20% of those surveyed found this temperature to be too warm and approximately 20% too cold. Consequently, a significant portion of those surveyed were dissatisfied with the climate (Fanger, 1972). Thus, climate sensation is highly individual, whereas on average summer high temperatures are felt to be more pleasant than those in winter.

Ultimately, climate design should strive to create damage-free, better executable, and in general achievable conditions. During ideal conditions, however, a comfortable climate is established when body heat balance turns out to be neutral.

### 2.4.5   Tools

In principle, a tool is differentiated in the hand side and the work side, whereas ergonomics has its focus on the hand side.

A deductive approach (i.e., from the general to the specific) works best when designing an ergonomic hand side. Inductive approaches that start with decisions about shape are usually doomed to fail or require extensive reworking, which is time consuming and expensive. Ergonomic products are optimized for the human user, taking individual abilities and skills into account, helping to prevent one-sided stress during work and increase efficiency.

The relevant measured variables on the hand and work sides of a tool affect:

- Body position, posture, and range of motion
- Hand position, grip type, and connection type
- Handle shape, dimensions, material, and surface
- Function direction and force direction
- Accuracy, movement speed, and resistance

A systematic approach is essential to ensure that these variables are implemented into the design process in the right amount and order. The design process starts with the consideration of the task, that is, the examination of working conditions. Only after further detail analysis do design parameters (i.e., shape, dimensions, material, and surface) actually become part of the process (see Figure 11).
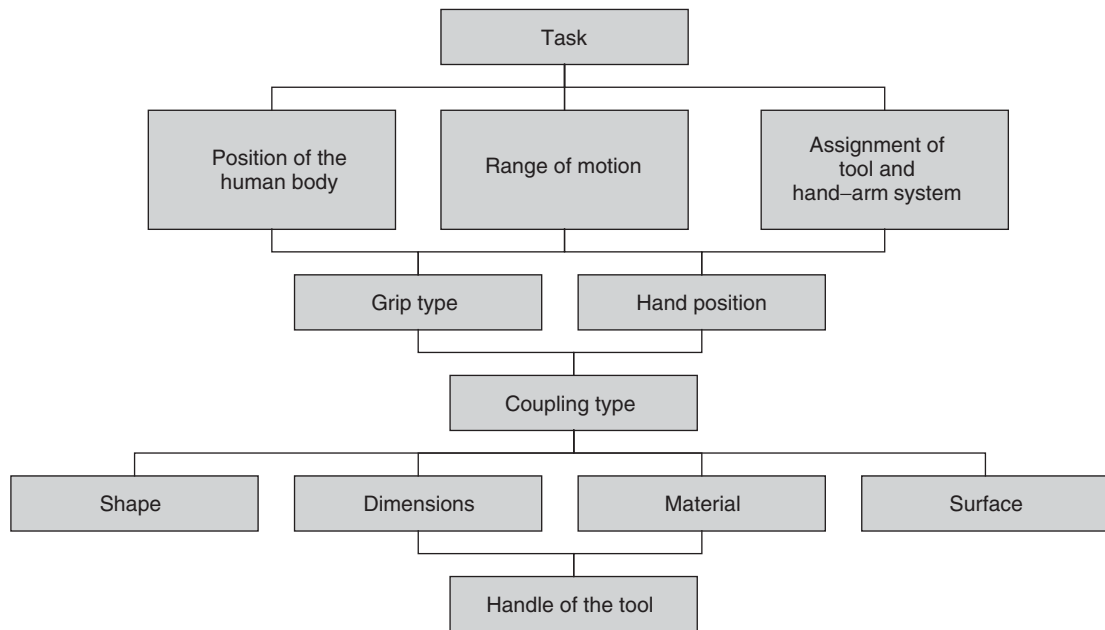
**Figure 11** Algorithm for the design of work tools (Bullinger, 1994).

In this way, tool handle design becomes a creative and analytical process instead of a purely technical and aesthetic task. The procedure conforms to classical scientific engineering methods, which may also be complemented with usability engineering practices (Meinken et al., 2008).

### 2.4.6 Workplace Design

A basic requirement for the use of manpower is an ergonomic workplace. Workplaces which are arranged ergonomically inadequate not only affect occupational safety and health but also limit effective assignment.

Because musculoskeletal diseases cause the most illness-related absences from work, it is important that special attention be paid to workplace design with respect to dimensions and forces.

The advantages of workplaces designed ergonomically can be assessed directly and indirectly. Processing time is decreased and number of absence is reduced. Both human body dimensions and physical forces are main aspects of ergonomic work design.

Human body dimensions are especially important as here the minimal values are not always as fundamentally decisive as physical strength.

***Anthropometry*** Anthropometry is the scientific approach of the different human body dimensions and their exact determination. Since everybody has different anthropometric values, the only way to create an ergonomically correct and efficient working environment is to adjust workplaces individually. The most important criteria for determining workplace dimensions are shown in Figure 12.

Human physical measurements vary. Therefore, defining an "average human" is not applicable as designs often have to accommodate both the smallest and largest users. If designs were based on average dimensions, half of the population would be worried about hitting their heads on doorframes, while the other half would fear not being able to reach an emergency power switch in the event of an accident. The definition of upper and lower limits for broad population groups has proven to be a good feasible method. This consideration leads to the term "percentile."

The customary limits for the adjustment area of an object, adapted to the human body, are the 5th and 95th percentile. Since the variance of the residual extreme groups is overproportionally large, 95% of the users can be addressed by only approximately one fourth of the entire variation range.

However, when using anthropometrical data, it should be kept in mind that it is gradually changing. The current acceleration phenomenon results in the fact that physical dimensions are slowly and continually increasing.

The action space for body parts is confined by the anatomically maximal rotary range, respectively displacement range and bent angles. Since details about optimal ranges and angles are required in the workplace design, the maximal action space plays a minor role in ergonomics. There are neutral positions between the extreme values, in which muscle activity as well as tendon and ligament strain is minimal. Muscle exhaustion is at its lowest level in these positions, which are generally considered to be subjectively comfortable but do not have to inevitably be in the respective centralized position.
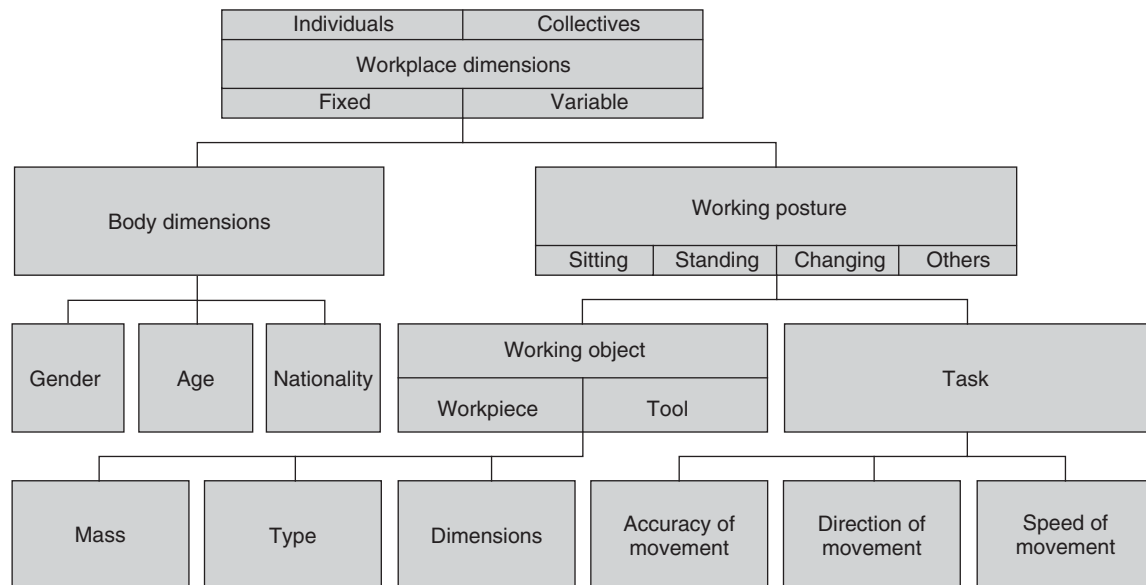
**Figure 12** Factors of workplace measure specification (Bullinger, 1994).

This also applies to the visual space. The work process must be visually controlled and checked for nearly all tasks. Therefore, besides body position and body posture, the visual space parameters are important for workplace design.

***Dimensioning of Workplaces*** The nature of the work task primarily influences the choice of working height and the decision of whether to set up a sitting, standing, or sitting–standing workplace. In principle, minor motional limitations exist at a standing workplace. For this reason, great force development is possible. The sitting workplace is favorable for precisional work and reduces posture work. From a physiological viewpoint, the sitting workplace should be preferred to the standing workplace since work chairs corresponding to ergonomic requirements can reduce the number of continuous muscle contraction. Standing steady does not cause much muscle activity but does strain the ankles and affected tendons and ligaments, which can lead to increased blood pressure in the legs.

It should be noted that no standing or sitting posture can be comfortably maintained for a long period. Therefore, a workplace should be set up such that it is possible for the worker to alternate between sitting and standing (sitting–standing workplace). Thereby it is possible for the worker to have a balanced amount of movement which is more favorable for the human body than motionless forced posture. This fact is expressed by the ergonomic proverb "the most ergonomic (sitting) position is the next one."

Forced posture, however, results from:

- Not enough free space at the workplace (i.e., foot space, leg space)
- Unfavorable working height (i.e., height of chair, height of table)

- Unfavorable position of work objects (i.e., too large of a frontal or side displacement, too high or too low)
- Arrangement of work tools and manual controls far from the body
- Unfavorable position of displays
- Limited space of movement (e.g., projecting components of machinery)

A workplace adjusted to human dimensions is necessary in order to provide natural body posture and motion sequences during work.

Because of the multiple influencing factors to the dimensioning of workplaces, each respective method has a different relevance. Some methods are appropriate for a quick qualitative examination of workplaces, while others are complex and produce detailed results. An overview of methods and tools for workplace design is shown in Figure 13.

At this point two established methods are presented:

- *Calculation Using Body Measurement Charts.* Using body measurement charts, specific workstation measurements are defined whereby certain measures pertain to the 5th percentile (e.g., hand and foot operating space), whereas others pertain to the 95th percentile (open space for feet and knees). Summing up, it is necessary that a small person be able to reach all relevant elements, while a large person should have ample room and not hit anything at the same workplace.
- *Computer-Based Methods.* A variety of computer-aided design (CAD) system applications opens up a basis for efficient workplace design. In particular, variant designs and quick changes of particular components of a work system improve

| Influencing factors | Method/instrument | Workplace dimensions |
|---|---|---|
| • Anthropometry<br><br>• Task<br><br>• Body posture<br><br>• Movements<br><br>• Viewing conditions | • Calculation using anthropometric charts<br><br>• Recommendations for the dimensioning of workplaces<br><br>• Somatography<br><br>• IT-based methods | • Workspace<br><br>• Desk/seat height<br><br>• Motional range<br><br>• Movement space<br><br>• Safe distances |
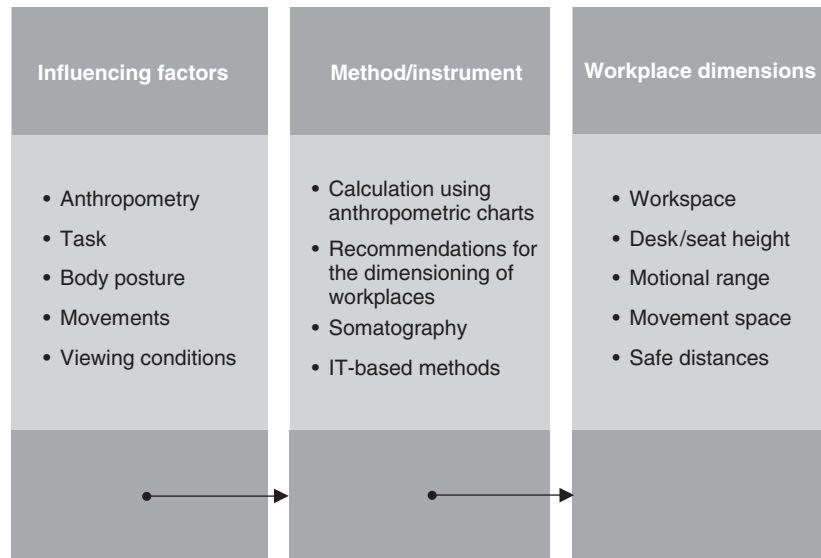
**Figure 13** Methods for workplace design.

the ergonomic design of the entire work system. Depending upon type and performance of available hardware and software, human models with different degrees of detail are available. Two-dimensional (2D) models, corresponding in principle to digital templates, are standard. On the other hand 3D human models such as JACK (Allbeck/Badler 2002) or RAMSIS [Forschungsvereinigung Automobiltechnik (FAT), 1995] allow a detailed ergonomic analysis of the entire workplace system (see Figure 14).

### 2.4.7 Manual Load Handling

Despite the handling technology application, heavy loads have to be moved often during manufacturing.



**Figure 14** IT-based workplace design using VirtualANTHROPOS. (Courtesy of Fraunhofer IAO.)

Here, back stress is so intense that serious health effects cannot be excluded: Acute limited functional impairments can appear (e.g., strained muscles and blockage of vertebrae joints while lifting loads). Furthermore, chronic impairments along with steadily increasing continuous medical conditions can appear (e.g., deteriorating intevertebral disks, expansion of ligaments, tendosynovitis, and muscle tension). These cause pain and often limit human flexibility. They can lead to longer periods of disablement [National Institute for Occcupational Safety and Health (NIOSH), 1994].

Strains are effectively reduced when constant or heavy lifting is avoided. This applies especially for young people (because of reduced strain capacity of the spine) and women (because of lower average strain capacity of the spine in comparison to men).

Where this is not possible, appropriate measures have to be taken on the basis of a work analysis in order to keep health hazards as low as possible. A variety of tested and proven methods are available in order to conduct an ergonomic assessment. These methods include operational terms and conditions and corresponding legal guidelines. Resulting measures to be taken can be carried out with the aid of technical, organizational, or individual means. Recommendations for the manual handling of loads are:

- Supply of appropriate work tools and utilities (lifting belts, lifting platforms, etc.)
- Opportunity that the load can be carried and transported close to the body
- Favorable load lifting, respectively depositing heights between 70 and 110 cm over the ground
- Adequate motional range for load handling
- Alternating between straining and relieving tasks
- Adequate relaxation time

Informing employees about the right way of lifting and carrying loads is necessary so that they can avoid health impairments. Here, information about straight posture of the spine is important leading to an equal distribution of the stress. Heightened strain on the spine results from the weight of the load; load distribution is uneven when the torso is bent. For this reason, the danger of intervertebral disk or vertebral body damage increases considerably.

### 2.4.8 Human–Machine Interface

All components of a work system for functional interaction between the human and technical system are defined by the term *human–machine interface*. Processes to be supervised and controlled by humans generate a multitude of information such as operating status or task conditions. Human receptors take in this information both directly and indirectly. This information is processed in the brain and is usually referred back to the process by an operation.

Information is transferred to the technical system in an operation by using either so-called actuator components or control elements (i.e., switches, levers, buttons) or by using complex informational input systems (such as speech input systems, graphic charts, and keyboards). The system can in turn communicate with the user by using visual, acoustic, or haptic modes of interaction. Today more and more automatic announcements and multi-/hypermedia systems are in use due to highly sophisticated language and image processing systems as well as highly effective data storage.

When designing the modes of interaction, the different individual attributes are to be considered: Human expectations are attributed to inherited or, in the technical and social environment, acquired behavioral stereotypes which are dependent upon certain population affiliations (e.g., left-handed people). These behavioral stereotypes are introduced into the work system as part of the individual's proficiency requirements. If one wants to use them during the task or at least respond to them, they must be considered during display and manual control design, that is, the display and manual controls must be designed compatibly.

Compatibility of informational input and output of work tools is accomplished when certain human expectations regarding spatial as well as mapping of dynamic procedures correspond. In the dynamic case, movement compatibility is of major importance. A well-known example is steering a vehicle. When turning the steering wheel right, the vehicle is expected to turn right; the same holds true for the converse.

In order to obtain an option and design of informational input and output elements customized to the work task, it is necessary to analyze the task. The arrangement and design of input and output media result from the work procedure.

In past years, during the course of increasing automation, a trend from manual tasks to control and monitoring tasks has appeared. Meanwhile, microcontrollers are being built into machines. For this reason, however, the mode of the human–machine interaction (HCI) is also



**Figure 15** Depiction of a typical process control workstation. (Courtesy of BKB Göppingen.)

changing. Classical machine interfaces, which mainly use discrete operating elements, are nowadays only found rarely. They are replaced by display screens on the machine as well as in control centers (see Figure 15). As a result of this, ergonomic software considerations (often referred to as HCI) are becoming increasingly important for the division of production. Thus, the design of information flow under cognitive considerations is still a challenge for HCI.

If in the past the human and machine were considered to be independent partners, now the human is seen as the central component of the human–machine system to which the machine is to be adjusted (Oborne and Arnold, 2001). Since the human, however, learns through interaction with the machine, the best possible mode of interaction is also changing constantly. It is here that the human–machine system is considered to be dynamic. "So the goal is to create supportive dynamic environments that enable individuals to work at their safest and most effective levels; not just to design the environment to fit the individual in some static sense" (Oborne and Arnold, 2001).

The central tool used to accomplish this goal is *usability engineering*. Lin et al. (1997) define usability as "the ease with which a . . . product can be used to perform its designated task by its users at a specific criterion." According to ISO (2006) 9241-110, four steps in the design process will ensure usability:

- understand and specify the context of use, which leads to
- specify the user and the organizational requirements, which leads to
- define product design solutions, which leads to
- evaluate the designs against the requirements.

For monitoring tasks, constantly appearing in process control, a good interaction system design implies a significant safety aspect. Information must be taken and processed reliably according to its importance. For this purpose, the human must not be overstrained or unchallenged. In the worst case, a decrease in vigilance

can result from monotony when the user is not able to act quickly enough in an acute critical situation. As a result, more damages can threaten the human and environment. In addition to ergonomic design of modes of interaction, such as the geometric arrangement of display elements, vigilance management systems can bring an important safety contribution by detecting decreasing vigilance and warning the respective user, correspondingly by maintaining vigilance through appropriate activation.

## 3 DESIGN PRINCIPLES FOR MANUFACTURING SYSTEMS

So far, the manufacturing system, its elements, and the approved methods for ergonomic work design were introduced. These methods, however, correspond primarily to isolated elements of a work system. For the design of manufacturing systems, limitation to individual elements is not sufficient. In fact, systemic consideration is required, as it happens within an integrative design process. Design principles and measures of ergonomic and efficient manufacturing system development, whose conceptual basics were already named, will be subsequently presented.

### 3.1 From Execution- to Object-Oriented Work Content

Organizational production according to the principles of Taylorism leads to high division of work, which results in less work. As a result, employees only identify with the work and its results to a minor degree. Besides these human problems, the high division of work also leads to organizational problems resulting from a multitude of interfaces. An example of this is the sometimes troublesome adjustment of the cycle times.

*Team work*, embedded in decentralized enterprise structures, can serve as a solution to this problem. These groups are established according to the object principle. An example is the complete assembly of all attached parts and aggregates for a dashboard (object) of a car (product). In the further production process, the completely assembled dashboard is installed as a whole in the body work of the car.

This procedure generates small units. In addition, the scope of action and decision making for each worker expands and groups are able to manage and coordinate operating processes independently.

The object-oriented concept has most notably become accepted in the form of production and assembly insulars. Thereby, the team has the task of the entire production process, assembling a spectrum of particles. They also take on tasks from which the individual employee is excluded by the performance principle of Taylorism. The additional work content occurs as a result of:

- Job enlargement, by overtaking preliminary and downstream functions (e.g., material provision, examination of the particles)
- Job enrichment, by overtaking the production and preparing functions (e.g., setup and maintenance of installation, material disposition)

### 3.2 From One- to Multidimensional Work Qualifications

In the work system, complex work structures such as production and assembly insulars also require employee organizational skills aside from technical skills and abilities. These are needed since employees handle greater work content in terms of the scope of services and efficiency diversity within these production structures.

Cooperation within the group, however, also employs new demands on the cooperation and coordination capabilities of the individual. It is clear that the introduction and implementation of more innovative production structures with conventional, strongly subject-oriented training concepts on a wider basis are problematic since a sufficient number of qualified employees are not available right away. Here, a training concept is required for enterprises in order to provide technically well-trained employees with strong organizational and social skills.

### 3.3 From Fixed to Flexible Working Times

The reduction of working hours, mainly in industrial countries, calls for the development of more flexible work models so that the work time factor does not become a disadvantage for the enterprise. This results in pressure to use intensive production systems to remain competitive in the international market.

In the past few years, the following working hour designs in production are recognized:

- Flexible work time models
- Seasonalization of working time
- Working time differentiation (maintenance of preferably high working time volume for highly qualified employees, of whom not very many can be found in the job market)
- Part-time work
- Individualized working time models instead of collective working time models

Note that such flexible working time models must be attractive so that employees accept them, that is, the employees have to understand the advantages that arise from these models.

### 3.4 From Reactive to Preventive Work Design

Work design should protect employees from work-related risks. Therefore occupational health and safety (OHS) criteria have been established in enterprises. In the past, OHS predominantly focused on human protection from hazards. A reactive procedure, however, does not meet the requirements of a human-oriented work design. It has been recognized that the preventive dimension of OHS has to be strengthened, that is, the protection of humans' health is achieved through the avoidance and elimination of hazards before damage occurs.

Practice shows that the implementations of these conclusions are not yet adequate. This is mainly because hazard risks need to be identified in an early phase

of production system planning so that the human-oriented realization of the working system can be based on it. Otherwise, for economic reasons, the practical adjustment of ergonomic design measures might become impossible during the course of the project.

## 4 HUMAN RESOURCES MANAGEMENT

In this section, essential aspects of human-oriented work design are discussed in detail in human resource management perspectives. One focus of human resource management rests upon the accomplishment of demographic change challenges, which are related to measures of qualification and occupational health.

### 4.1 Accomplishing Sociodemographic Change

Sociodemographical development in nearly all industrial nations leads to a fractional increase in the elderly (worker) population. More dramatic than the decline in the absolute number of working people is the change in age composition as the number of younger workers slowly but continually decreases and the number of elderly people able to work will continue to grow in the coming years. With this in mind, there will be an increase in the average age of staff and shortages when recruiting a younger work force. Youth-centered enterprise principles, short-term success-oriented personnel policies, and general conditions that favor the retirement of older workers do not measure up to this development (Pack et al., 2000).

The age in professional life mostly becomes a problem when employees remain in stressful tasks for a long time and the required specific capability is used up to the point that individual resources do not satisfy the work requirements any more.

On-going physical and psychological stress in badly designed work systems is partly responsible for physical decline and decreasing cognitive flexibility. Thus, for example, a lack of physical demand resulting from unbalanced body posture such as continuously sitting while working leads to the reduction of physical ability efficiency and ultimately to the same result as capacity overload, namely musculoskeletal illnesses.

The condition of an employee's health is not primarily appropriated by their age but rather by the result of past work conditions. A reduction in older employees' work efficiency ability does not generally apply; it always refers to specific tasks and work requirements. For example, a machine operator with damaged intervertebral disks might no longer be able to operate a machine but might be able to work in administration.

There is no standard solution for designing age-based jobs, personnel placement, and work hours. The acceptable method for an enterprise depends upon specific initial conditions of the enterprise. In general, however, a basic sensibility toward the topic of age is important.

Age-based work design remains an illusion as long as production strategies exclusively aim for short-term economies of scale and profit increases. Indeed, such strategies cause long-term damage to the enterprises themselves, as misunderstood examples of *lean management* show. Here, the principally correct idea of reducing the number of hierarchal levels caused a number of personnel reductions that resulted in a massive lack of qualified employees (Kern and Braun, 2006).

In order to face the risks of sociodemographic change for performance and innovation ability, work design and human resource management have to be able to support employee psychological and physical capacity during their entire professional lifetime and open up a larger degree of specific capabilities for elder employees. During work design, it has to be considered that with any work, middle to long term, depending on requirements, the physical and psychological capacity derived through learning, training, and degradation processes can change. Therefore, work should be designed in a way that an assorted amount of body movement as well as varied mental specifications are required to accomplish assignments.

Effective age-based work design should be set up not only for elderly people, who are already concerned with performance limitations, but also for younger people so as to counteract health damages as early as possible. For this purpose, a rethinking of both employees and managers is necessary. Criteria for appropriate personal development processes should confirm that through different work specifications:

- New knowledge is gained.
- Developing fixations concerning health-damaging strain and stress situations are halted.
- New social configuration (work groups and the like) are experienced and thereby new social key qualifications are learned.
- Individual willingness and ability to deal with new working situations and requirements are supported.

### 4.2 Qualification

In manufacturing systems, knowledge represents an important resource for the production of goods and services. Knowledge and capability significantly influence the innovation power and competitiveness of an enterprise.

A superior target of operational qualification measures is the development of comprehensive decision-making and responsibility as a sum of technical, method, social, and self-competence. Standard knowledge is losing importance as a basis for operational decision making. In the course of further manufacturing procedures, operational decision making will be based upon prognosis and diagnosis, which enables development of professional decision-making strategies.

The required availability of task-specific decision making and responsibility within complex production systems places high demands on the qualification. According to this, qualification processes are geared to the premises of decentralized organizational concepts:

- Qualification increasingly takes place in open knowledge networks, which are marked by exchange of knowledge and experience.

- Qualification strategies are more and more based on short-term learning content and on the integration of learning and implementation phases.
- Qualification methods incorporate learners' self-organization, for example, by means of computer-aided self-learning phases.

While in traditional qualification strategies knowledge is taught "on stock" and is timely implemented and spatially separated, an up-to-date gain in competence is more and more task oriented.

Poor predictability of future qualification places increasing demands on the flexibility and rapidity of learning. Up-to-date qualification concepts abandon teaching the concepts completely but use only elements of knowledge to develop of problem solution strategies, enabling the learner to reflect on new problems. A dynamic knowledge transfer is requested on the one hand by the increasing amount of required knowledge during the course of one's professional life and on the other hand by the short-lived validity of knowledge. It appears that specific professional knowledge becomes old relatively quickly and life-long learning is consequently becoming significantly more important. On-the-job training as a supplement to vocational training has become the norm.

### 4.3 Occupational Health Prevention

#### 4.3.1 Stress Situation

Considering all chances of innovative work structures, their negative effects on the employees' health situation cannot be overlooked. Increasing requirements on time and local flexibility as well as increasing performance demands are only a few effects of the structural change. Hereby, mental stress and chronic illnesses step into the forefront. These factors can lead to loss in human performance.

In past years mental health disorders are becoming common among workers. More than half of employees in the European Union complain about physical health damages caused by work (see Figure 16). In Germany, losses resulting from work-related health disorders are estimated to be over three billion Euros a year; stress resulting from work is the cause of 7% of early retirements. The most common causes of underperformance and absences are related to mental disorders (WHO, 2000).

In the workplace, the mental health and well-being of employees gain importance as resources worthy of being sustained and protected.

#### 4.3.2 Definition of Health

Traditional diction defines health as the absence of illness. According to timely understanding, health includes the goals of competence development as well as broad physical, mental, and social well-being (WHO, 1986). Health includes the ability to define and pursue individual long-term targets, the ability to adapt to changing environmental conditions, and the ability to take part in such changes. Therefore, a healthy person is target oriented and someone who actively acts within his or her world and develops further. Health relies on personnel and organizational resources. Occupational health is a precondition of coping with work challenges as well as a result of adequate working conditions.

#### 4.3.3 Prevention Strategies

For a long time occupational health prevention was primarily determined by the defense of acute risk hazards, that is, accidents at work, as well as by the impact of individual stress and illness factors with a definite effect on health. In addition to these risks, it is important to focus on the complex correlations of work factors and health effects. These complex correlations do not comply with the prototype of the specific cause–effect relationship. They can only be controlled by detailed regulations in a limited way. Furthermore, the health resources concept sets priorities by abandoning a risk-oriented approach (Braun, 2003).

Requests for occupational health prevention are the stabilization, that is, the enhancement of physical and mental health resources, as well as the mobilization of individual capabilities. Physical and mental resources should be developed in order to be able to cope with health disorders (see Figure 17). In addition to avoiding illness caused by work, the stabilization of health with respect to potentially illness-causing influences and
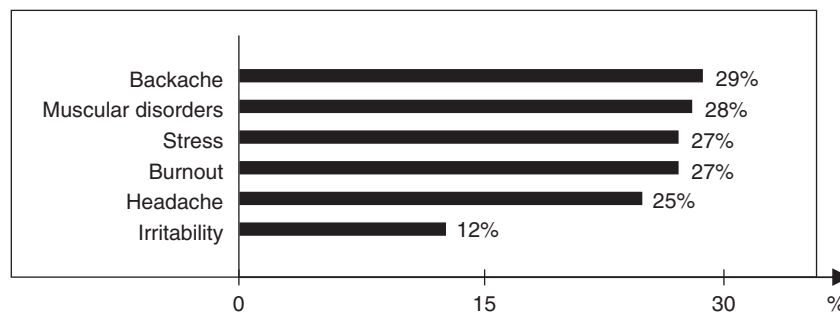


**Figure 16** Health problems caused by work. Fraction of complaints per 100 employees surveyed in countries in the European Union. (Data from European Foundation, 2005.)
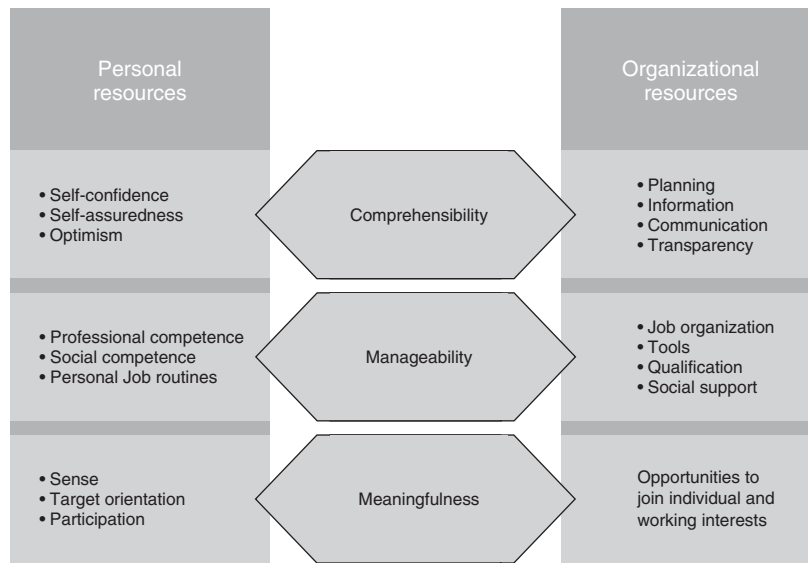
**Figure 17**   Organizational and personal health resources according to the coherence model by Antonovsky (Spath et al., 2003).

premature weakening processes caused by work are considered.

The resource-oriented definition of the health concept focuses on the human ability for competence development as a basis for health. According to this, health prevention is aimed at enabling humans to have greater self-responsibility for strengthening their health and, moreover, to qualify them for this. Advancement of work satisfaction and social well-being, resulting from cooperation possibilities and a positive setup of the operational environment, come into central focus.

## 5   CONCLUSION

In past years, production enterprises made numerous efforts to increase their performance, efficiency, and flexibility and thereby satisfy market demands. Fundamental characteristics of manufacturing concepts are business process design and team work. Despite great endeavors and obvious success, a number of deficiencies still emerge as a closer look at manufacturing processes and results shows, for example, long cycle time and time of delivery.

In view of the obvious limits of technology-oriented strategies, it has been recently recognized that the targeted success in manufacturing can only be reached through the integrated actions of the human, technology, and organization. Production systems integrate elements of the organization of assembly, process, work, quality control, and continuous improvement, which for a long time have been only considered isolated, as one system. Design principles, methods, and tools have to be integrated into the production system. Nevertheless the established methods of human-oriented work design

have to be regarded as a necessary requirement for economic success.

One aspect that is critical for success is the development of production systems "from the operational practice." An efficient production system can only be realized when considering the specifications of the enterprise. Understanding the business culture and human values make up the specific requirements for successful implementation. It is necessary to involve all participants, to identify good practices, and to make this the business standard. In this way, the system is comprehensible and practicable for all participants.

The comprehension of human success factors such as qualification, information, and participation is of particular importance. On the one hand, this calls for meaning, commitment, and readiness to take on responsibility, and, on the other hand, it results in win–win situations as well as in leadership through personal commitment.

In the past, manufacturing enterprises have primarily relied upon quantitative or "hard" success factors. They are comprehensible by means of strategy papers, plans, organization charts, job descriptions, operating instructions, and target systems. Currently, a tendency toward a stronger emphasis on qualitative or "soft" success factors can be seen which are geared to the conditions of the worker. These factors include abilities, values, cultures, and participation. The soft success factors are often impossible to be described definitively. Although these soft factors rather remain latent, they crucially affect the performance of businesses, as numerous examples document: These companies are very successful in relation to economic efficiency and competitive ability, customer satisfaction, and employee health.

In view of the multiple business influences it is obvious that restructuring measures and technological advantages are no longer sufficient for maintaining the enterprise's competitiveness. Ultimately, the healthy, motivated, and qualified human generates distinctive and, hence, decisive competitive products and services. Consequently, human factors remain a focus of business enterprise strategies.

## REFERENCES

Allbeck, J., and Badler, N. (2002), "Embodied Autonomous Agents," in *Handbook of Virtual Environments*, K. Stanney, Ed., Lawrence Erlbaum Associates, Mahwah, NJ, pp. 313–332.

Antonovsky, A. (1979), *Health, Stress and Coping*, Wiley, New York.

Bevan, N. (1999), "Quality in Use: Meeting User Needs for Quality," *Journal of System Software*, Vol. 49, pp. 89–96

Braun, M. (2003), "Gesundheitspräventive Arbeitsgestaltung und Unternehmensentwicklung," *Das Gesundheitswesen*, Vol. 65, No. 12, pp. 698–703.

Braun, M. (2008), "Gesundheit aus arbeitswissenschaftlicher Perspektive," in *Gesundheit—Gesundheiten? Eine Orientierungshilfe*, I. Biendarra and M. Weeren, Eds., Königshausen und Neumann, Würzburg, pp. 125–165.

Bullinger, H.-J. (1994), *Ergonomie—Produkt- und Arbeitsplatzgestaltung*, Teubner, Stuttgart.

Bullinger, H.-J. (1997), "Mechanische Werkzeuge und Maschinen," in *Handbuch der Arbeitswissenschaft*, H. Luczak and W. Volpert, Eds., Schäffer Poeschel, Stuttgart, pp. 598–601.

Bullinger, H.-J. (2003), *Früherkennung von Qualifikationserfordernissen in Europa*, Bertelsmann, Bielefeld.

Bullinger, H.-J., and Braun, M. (2001), "Arbeitswissenschaft in der sich wandelnden Arbeitswelt," in *Erträge der interdisziplinären Technikforschung*, G. Ropohl, Ed., Schmidt, Berlin, pp. 109–124.

Cofer, C., and Appley, M. (1964), *Motivation—Theory and Research*, Wiley, New York.

Collins, J., and Porras, J. (1994), *Built to Last: Successful Habits of Visionary Companies*, Harper Collins, New York.

European Foundation for the Improvement of Living and Working Conditions (EFILWG) (2005), "Third European Working Conditions Survey," EFILWG, Dublin.

Fanger, P. O. (1972), *Thermal Comfort: Analysis and Applications in Environmental Engineering*, McGraw-Hill, New York.

Forschungsvereinigung Automobiltechnik (FAT) (1995), *RAMSIS—ein System zur Erhebung und Vermessung dreidimensionaler Körperhaltungen von Menschen zur ergonomischen Auslegung von Bedien- und Sitzplätzen im Auto*, FAT-Bericht, Frankfurt.

Gallup (2010), "Gallup Engagement Index 2009," Gallup, Potsdam:.

Hacker, W. (1998), *Allgemeine Arbeitspsychologie: Psychische Regulation von Arbeitstätigkeiten*, Huber, Bern.

Hagenmeyer, L. (2007), "Development of a Multimodal, Universal Human Machine Interface for Hypovigilance Management Systems," Dissertation, Jost-Jetter, Heimsheim.

Helander, M. (1995), *A Guide to the Ergonomics of Manufacturing*, Taylor and Francis, London.

Herzberg, F., Mausner, B., and Snyderman, B. (1967), *The Motivation to Work*, Wiley, New York.

International Organization for Standardization (ISO) (1991), "Ergonomic Principles Related to Mental Work-Load," ISO 10075, ISO, Geneva.

International Organization for Standardization (ISO) (2006), "Ergonomics of Human-System Interaction—Part 110: Dialogue Principles," ISO 9241-110, ISO, Geneva.

Karwowski, W., Ed. (2001), *International Encyclopedia of Ergonomics and Human Factors*, Taylor and Francis, London.

Karwowski, W., and Salvendy, G., Eds. (1998), *Ergonomics in Manufacturing—Raising Productivity through Workplace Improvement*, Society of Manufacturing Engineers, Dearborn, WI.

Kern, P., and Braun, M. (2006), "Arbeiten bis 67—Herausforderungen für die betriebliche Gesundheitsförderung," *Die BKK*, Vol. 94, No. 5, pp. 240–245.

Kern, P.; Schmauder, M., and Braun, M. (2005), *Einführung in den Arbeitsschutz*, Hanser, München.

Korge, A., and Scholtz, O. (2004), "Ganzheitliche Produktionssysteme," *Werkstattstechnik Online*, Vol. 94, Nos. 1/2, pp. 2–6.

Lin, H. X., Chong, Y.-Y., and Salvendy, G. (1997), "A Proposed Index of Usability: A Method for Comparing the Relative Usability of Different Software Systems," *Behaviour and Information Technology*, Vol. 16, pp. 267–278

Luczak, H. (1998), *Arbeitswissenschaft*, Springer, Berlin.

Luczak, H., Volpert, W., Raeithel, A., and Schwier, W. (1987), *Arbeitswissenschaft, Kerndefinition—Gegenstandsbereich—Forschungsgebiete*, RKW, Eschborn.

Maslow, A. (1987), *Motivation and Personality*, 3rd ed., Harper Collins, New York.

Meinken, K., Rix, A., Widlroither, H., Plihal, U., and Muellerleile, A. (2008), "Ergonomic Design of a Multilevel Writing System for School Children," in *Conference Proceedings of International Conference on Applied Human Factors and Ergonomics*, W. Karwowski, Ed., AHFE International, Las Vegas, NV, July 14-17, 2008, USA Publishing, Kansas City.

National Institute for Occupational Safety and Health (NIOSH) (1994), *Applications Manual for the Revised NIOSH Lifting Equation*, Publication No. 94-110, NIOSH, Washington, DC.

Oborne, D. J., and Arnold, K. M. (2001), "Human-Machine Interaction: Usability and User Needs of the System," in *Handbook of Industrial, Work and Organizational Psychology*, N. Anderson, D. S. Ones, H. K. Sinangil, and C. Viswesvaran, Eds., Sage Publications, London, pp. 336–347.

Pack, J., Buck, H., Kistler, E., Mendius, H., Morschhäuser, M., and Wolff, H. (2000), *Zukunftsreport demographischer Wandel—Innovationsfähigkeit in einer alternden Gesellschaft*, Federal Ministry for Research, Bonn.

Richter, P., and Hacker, W. (1998), *Belastung und Beanspruchung—Stress, Ermüdung und Burnout im Arbeitsleben*, Asanger, Heidelberg.

Salvendy, G., Ed. (2006), *Handbook of Human Factors and Ergonomics*, 3rd ed., Wiley, New York.

Scholz, O., Korge, A., and Schlauss, S. (2003), "Was ein Produktionssystem ausmacht," in *Ganzheitlich produzieren*, D. Spath, Ed., Logis, Stuttgart, pp. 53–84.

Spath, D. (2003), "Revolution durch Evolution," in *Ganzheitlich produzieren*, D. Spath, Ed., Logis, Stuttgart, pp. 15–44.

Spath, D., and Dill, C. (2002), "Ist Flexibilität genug? Turbulenzen sind nur mit systemischem Denken zu bewältigen," in *Erfolg in Netzwerken*, J. Milberg and G. Schuh, Eds., Springer, Berlin, pp. 161–175.

Spath, D., Braun, M., and Grunewald, P. (2003), *Gesundheits- und leistungsförderliche Gestaltung geistiger Arbeit*, Schmidt, Berlin.

Spath, D., Braun, M., and Bauer, W. (2009), "Integrated Human and Automation Systems, Incl. Automation Usability; Human Interaction and Work Design in (Semi-) Automated Systems," in *Handbook for Automation*, S. Nof, Ed., Springer, New York, pp. 571–598.

Taylor, F. W. (1911), *The Principles of Scientific Management*, Harper, New York.

Westkämper, E. (2004), "Structural Change in Manufacturing—Caused By Turbulent Influencing Factors," in *Proceedings for the International Conference on Competitive Manufacturing (COMA)*, Stellenbosch, February, 4-6, 2004, D. Dimitrov, Ed., University of Stellenbosch, pp. 21–27.

Westkämper, E., and Warnecke, H.-J. (2002), *Einführung in die Fertigungstechnik*, 5th ed., Teubner, Stuttgart.

World Health Organization (WHO) (1986), "Ottawa Charter for Health Promotion," WHO, Ottawa.

World Health Organization (WHO) (2000), "International Labour Organization: Mental Health and Work: Impact, Issues and Good Practices," WHO, Geneva.

# CHAPTER 61

# HUMAN FACTORS AND ERGONOMICS IN AVIATION

**Steven J. Landry**
**School of Industrial Engineering**
**Purdue University**
**West Lafayette, Indiana**

## 1   INTRODUCTION: BRIEF HISTORY OF HUMAN FACTORS PROBLEMS IN AVIATION

Aviation is inherently a complex sociotechnical system. It involves the operation of numerous interacting vehicles by human operators within a complex system managed by other human operators. The performance of those human operators is tightly coupled to the safety and performance of the individual vehicles and the overall system. As such, human factors problems and the solutions to those problems have been of paramount importance to the air transportation system.

Overall, this chapter focuses on the principles that have been developed with respect to aviation, how those principles have been applied, and what challenges lie ahead. These principles and their application are described starting in Section 2, which describes principles related to the flight deck. Section 3 describes principles associated with air traffic control, and Section 4 describes principles associated with maintenance and safety management. Section 5 describes emerging issues, and Section 6 summarizes what we have learned and what future issues are appearing on the horizon.

Understanding these principles, however, requires an understanding of their genesis. In this section, a brief historical review of the problems that precipitated the need for these principles is provided. That review is, as is the case in aviation, notably driven by fatal accidents. Due to, in part, the complexity and safety criticality of aviation, change does not occur easily—it often takes catastrophe to push system managers to make large changes.

In addition, aircraft themselves are quite robust to failure. Serious aviation accidents are astonishingly rare; of those, accidents in which the pilots had no ability to save the aircraft are a small minority. In many cases, a properly designed system with a properly trained pilot would have been sufficient to save the aircraft.

In covering such principles, however, several aspects of human factors in aviation are not covered here. In particular, except for the principles related to circadian rhythms, the human factors of passenger travel in aviation, including the design of aircraft to improve passenger travel, are not discussed. (It is, however, a notable development that the Boeing 787 aircraft has incorporated features, such as daylight-simulating lighting and a lower cabin altitude, designed to improve passenger comfort.) Also, the human factors of the job of the flight attendant and that of the "ground side," such as passenger loading, baggage handling, aircraft marshalling, refueling, and deicing, are not discussed.

For a discussion of the human factors of passengers and principles for the cabin crew, see Hawkins (1987). Numerous texts that discuss ergonomics, including this handbook, exist and the principles within those texts can be applied to the ground side and passenger cabin. Notably, the Federal Aviation Administration (FAA, 2007) has published a manual for airport operations human factors, and some specific aspects have been investigated in more detail (e.g., Korkmaz et al., 2006). A description of the principles for the layout of passenger terminals, along with a good description of airport ground-side operations, is found in Kazda and Caves (2007).

### 1.1  First Flight–1940

In 1903 the Wright brothers took what is generally considered to be the first piloted, powered, and sustained flight. Aviation advanced rapidly, with Orville Wright living to see the dawn of supersonic jet propulsion aircraft.

Numerous problems, most related to the design of the vehicle itself, had to be overcome to allow for this rapid advancement. Safety became a paramount concern, with pilot errors a notable cause of crashes [National Advisory Committee for Aeronautics (NACA), 1921]. With the use of aircraft in two world wars, the performance of the vehicle and the pilot were also of great interest to researchers and engineers. One example of this is the identification of the problem of "situation awareness" by Oswald Boelcke, a World War I ace, who indicated that "the pilot must acquire the habit of 'taking in' unconsciously the general progress of the whole multi-aircraft dogfight going on around the individual combat in which the pilot will become involved ... [so that] no time [is] wasted in assessment of the general situation after the end of an individual combat (Hacker, 1984, p. 1)."

In recognition of the great promise of aviation, the United States founded NACA in 1915. NACA became a repository for worldwide aviation research for several decades.

### 1.1.1  NACA: Early Work on Human Factors

At NACA, several problems were identified early on in the development of aviation. The physiology of flight, including altitude sickness, became an issue as aircraft began to be operated at altitudes above 10,000 ft. A NACA report relayed the results of tests in pneumatic chambers by German researchers who found that severe symptoms, including unconsciousness, manifest themselves above 6500 m (NACA, 1921). (It is now known that altitude sickness can manifest itself as low as 2500 m.)

Hersey described challenges with instrumentation. Specifically, Hersey noted that pilots' use of instrumentation can be inappropriate, where important but poorly designed instruments may be ignored, and unimportant but salient instruments capturing the attention of the pilot. More generally, Hersey wrote that, "the reaction of the aviator to his [sic] instruments has to be considered ... [the instruments] must be ... readily intelligible to the pilot (and) must not make an appreciable demand on his [sic] time or attention" (Hersey, 1923, p. 482).

Researchers also identified that the challenge of interpreting instruments is compounded by the workload and time pressure experienced by the pilot. It was clear that instruments must be made for "the easiest possible reading and manipulation ... (due to) the extremely short time at the disposal of the pilot for adjustment (NACA, 1921, p. 5)."

Pilot error was very clearly a large source of serious accidents. In a 1924 report, errors in piloting were the largest single cause of serious accidents, with nearly two thirds of such accidents associated with pilot error during training and one third of such accidents in civil flying (Devaluez, 1924). A similarly large percentage (54%) of accidents were attributed to pilot errors in a slightly later study of accidents in French aviation (Brunat, 1927). Generally, these accidents were not viewed as a problem of the interaction of pilot and vehicle but, as is common today as well, are attributed to poor piloting skills and judgment. The suggested remedy for such problems was viewed to be training and selection, rather than better design of the vehicle and its instrumentation.

### 1.1.2  Equipment: Instrumentation and Automation

At *Concours de la Sécurité en Aéroplane* in France in 1914, Lawrence Sperry and his assistant, Emil Cachin, flew by the air show crowd utilizing a new device, a gyroscopic stabilizer, with their hands in the air to demonstrate that they were not touching the controls. (The gyroscopic stabilizer had previously been invented by Herman Anschütz-Kaempfe and manufactured by the Sperry Gyroscope Company for use on large naval vessels; see Hughes, 1971.) On a second pass, Cachin stood on the wing, with Sperry again holding his hands in the air; on the third pass both men stood on the wings, with the pilots' seats empty.

This automation utilized a simple feedback control system, a concept which had existed for centuries but whose utility did not extend to aircraft until Sperry developed a relatively light gyroscopic device. The autopilot was operated by a simple switch, and its only function was to stabilize the three axes of the aircraft (pitch, yaw, and roll). Since then, autopilots have increased in complexity but have remained standard equipment on commercial and military aircraft. Not only has the design of autopilots and their operation been of interest to human factors engineers and researchers, but the methods developed to perfect control automation have been applied to try to model the human pilot.

The autopilot is a somewhat rare example of a development that predated the problems that necessitated its use. Aircraft were considered generally stable and easy to fly for properly trained pilots (Warner, 1922a), and pilots and researchers did not see a substantial need or desire for an automatic pilot even almost a decade later (Warner, 1922b). However, much like driving a car, flying an aircraft requires constant vigilance, except that in an aircraft tracking must be done on the vertical axis as well as the horizontal axis and with respect to speed. Aircraft are also generally subjected to more substantial disruptions since their motion is easily disturbed by

changes in the wind. Because of this, it was clear from the very early days of flight that fatigue was a significant problem in controlling an aircraft on flights of even modest (1-h) duration (Devaluez, 1924). The autopilot would fill this need as flights grew in duration and as aircraft grew in complexity.

Along with the development of the aerodynamic and propulsion aspects of the aircraft, the development of instrumentation was considered crucial. A number of early studies were commissioned for the naming and taxonomy of instrumentation. From these papers, a number of human factors problems addressed or considered by early instrumentation were discussed, including:

- Ease of reading (Hersey, 1923)
- Errors in reading multiple-pointer round dials for altitude (Hunt, 1926)
- Trade-offs between space and usability (Subcommittee on Instruments, 1932)
- Spatial disorientation (Eaton, 1921)
- Usable and precise navigation and orientation instruments (Eaton, 1921)
- Use of vertical-scale instruments (Subcommittee on Instruments, 1932)
- Limits on the workload needed to interpret instruments (Subcommittee on Instruments, 1932)
- Discrete (vs. continuous) motion of instrument indicators (Subcommittee on Instruments, 1932)
- Arrangement of instruments, including Gestalt effects (Subcommittee on Instruments, 1932)
- Proper graduation and orientation of airspeed indicators (Beij, 1933)
- Visual field measurement and importance (Gough, 1936)
- Physical forces required to operate controls (Gough and Beard, 1936)
- Need for stall warning devices (Thompson, 1938)

Near the start of World War II, an aviation psychology research unit was established in England led by Sir Frederick Bartlett, and in the U.S., the National Research Council started a Committee on Aviation Psychology, led by Jack Jenkins (Roscoe, 1997a). These units focused much of their effort on pilot selection and training; they also were the first to deal with the problem of pilot error as a design problem instead of a personnel or training problem.

## 1.2 1950–1970

At the end of World War II, Lieutenant Colonel Paul Fitts edited a report commissioned by the Committee on Aviation Psychology that outlined the primary human factors challenges to aviation in 1951 (Fitts, 1951). Fitts was at the time leading a new research unit at Wright Field in Ohio on aviation psychology, which would produce a number of important results in aviation human factors. The report provides great insight into aviation problems at the beginning of the 1950s. As can be seen

from a comparison of the list generated by the Fitts panel, many of the problems identified prior to 1940 had not been satisfactorily addressed.

### 1.2.1 Fitts Panel

The Fitts (1951) report opens with a quote from a nineteenth-century textbook on applied mechanics which argues that the engineering of systems has largely focused on the machine and tools and ignored the operator, despite the operator being the most important part of the human–machine system. Morris Viteles, the Chairman of the Committee on Aviation Psychology, then states (p. iv):

> During the past decade, primarily in response to military requirements, there has been a growing withdrawal from this classical psychological viewpoint, and ever-increasing acceptance of the principle that "machines should be made for men; not men forcibly adapted to machines." This basic principle of "human engineering" is not new, particularly to industrial psychologists, although it has been consistently neglected, largely because engineers concerned with the development of the machine and scientists concerned with the individual who is to operate the machine have worked in almost complete insulation from one another. The resulting disregard of physiological and sensory handicaps; of fundamental principles of perception; of basic patterns of motor coordination; of human limitations in the integration of complex responses, etc. has at times led to the production of mechanical monstrosities which tax the capabilities of human operators and hinder the integration of man and machine into a system designed for most effective accomplishment of designated tasks.

The report then goes on to summarize the findings of the last several decades and to identify a long-range research program to address problems related to the navigation of aircraft in the air traffic system. The report may be unique in its scope and authority.

The report identified a number of human factors problems that needed to be solved, which will be numbered here so as to reference them in the remainder of the chapter. The problems identified include:

1. What is the proper role of humans in the air traffic system?
2. What tasks appear to exceed human capacities, either trained or untrained?
3. How do we measure human performance and behavior in the air traffic system?
4. How do we measure the information-handling capacity of pilots and controller, and how reliable are these measures?
5. What is the information-handling capacity of pilots and controllers?
6. What redundancy is needed in information presentation?

7. What information is needed by pilots and controllers, what is the required accuracy of that information, how much time do pilots and controllers have to use that information, and what actions or decisions do pilots and controllers make based on that information?

8. What are the relative advantages of single- versus multifunction instruments?

9. What is the best way to encode and present information on those instruments, including quantitative, spatial, status, and control information?

10. In what ways can communication be improved, including intelligibility, encoding, and the use of multiple modalities?

11. What variability, delay, and error can be attributed to the human operators in the air traffic system?

Over the next two decades, Fitts and a member of the review team named Stan Roscoe, among others, were instrumental in developing methods for evaluating flight deck and air traffic controller workload and performance. These efforts had a direct and substantial impact on the development of aircraft during the 1960s, including the Boeing 737, 727, and others.

### 1.2.2 Radar in ATC

The Fitts report also queried controllers about what equipment they would like to see adopted in the air traffic control (ATC) system. On the subject of radar, which had been developed in England during World War II and was being considered for use in the air traffic system, only 34% of the en route controllers desired such equipment. In 1951, aircraft were monitored by using flight status boards, on which the current and future (estimated) positions of aircraft, as reported by radio, were updated. Aircraft were separated based on the projected positions, but, with increasing air traffic, this method was becoming more and more difficult to apply effectively.

In 1956, two commercial aircraft, one a TWA Super Constellation and the other a United Airlines DC-7, collided over the Grand Canyon. Both aircraft were initially under the control of air traffic authorities, but, due to the need to fly around thunderstorms, the TWA flight requested, and was granted, visual flight rules, which meant the pilot was responsible for avoiding other aircraft. The TWA flight was struck by the DC-7 at 21,000 ft, disabling both aircraft and resulting in the death of all passengers and crew. This accident, in combination with several other collisions or near collisions, helped catalyze a radical change in the air traffic control system. Among the innovations mandated was the introduction of radar to monitor aircraft positions.

The introduction of radar is probably the single most radical change to have occurred in the air traffic control system to date. It fundamentally altered a controller's tasks and the nature of the controller's job and allowed for the introduction of a number of new technologies over the succeeding decades.

### 1.2.3 Altimeter Reading Problems/Design

Problem 9 on the Fitts report list was written broadly, but one of the specific problems in mind was almost certainly the difficulty in reading altimeters. Altimeters typically had multiple pointers on a single dial to indicate the hundreds, thousands, and (sometimes) ten thousands of feet. For example, in Figure 1, the altimeter shows an altitude of 10,180 ft.

The difficulty in reading such altimeters was the cause of two accidents in 1958 and was most likely a principal cause of two accidents between 1965 and 1967. The two crashes in 1958 involved aircraft carrying only crew. The 1965 accident involved a flight from LaGuardia Airport in New York to O'Hare Airport in Chicago. In that accident, a Boeing 727 crashed into Lake Michigan while on descent, killing 30. In 1967, a Caravelle was en route from Malaga, Spain, to London Heathrow, but descended slowly into trees in West Sussex, England, killing all 37 people on board. The accident report indicated the ease with which a pilot could mistake an indication of 6000 ft for one of 16,000 ft due to the lack of salience of the 10,000-ft pointer.

The solution to the problem, ultimately involving, in part, the vertical-scale instruments that had been under consideration since 1932, would also involve several of the other problems on the list. In particular, the salience of indications of problems would lead to the introduction of various types of alerting systems, which then spawned their own set of problems.

### 1.3 1970s — 1980s

From 1970 to 1990, the number of U.S. scheduled departures rose by 40%, from just over 5 million in 1970 to just under 7 million in 1990. During that same period, U.S. revenue passenger enplanements rose by 270%, from about 170,000,000 to 465,000,000. (European air



**Figure 1**  Three-pointer altimeter.

traffic rose in similar fashion.) This suggests that, in addition to an increase of 2 million flights, aircraft were larger and were carrying more passengers than ever before. The airspace was becoming more dense, and the consequences of accidents were growing.

Two other factors drove change in the air transportation system during the 1970s and 1980s. First, due to the Clean Air Act of 1970 in the U.S., which had followed earlier efforts in Europe such as the 1952 Clean Air Act in Britain, and the Arab Oil Embargo of 1973, airlines came to understand that new, more efficient aircraft were needed. Second, computing technology was being introduced in a form that would revolutionize instrumentation and which could be adopted for use in commercial aviation.

### 1.3.1 Elimination of Flight Engineer

As of 1970, all large commercial aircraft had three-person crews, which included a flight engineer, who was responsible for monitoring and controlling the fuel, pressurization, electrical, and hydraulic systems. The flight engineer was also expected to help troubleshoot any in-flight emergencies with the pilot. (The copilot would concentrate on flying the aircraft.) Due in part to advances in understanding and analyzing workload, and due in part to a desire to reduce salary costs for their client airlines, aircraft manufacturers began considering the possibility of reducing the crew size to two and eliminating the flight engineer.

This led to a great deal of human factors analysis and numerous hearings to resolve labor concerns. The general trade-off focused on two problems:

12. What is the relationship between flight deck workload and performance?
13. Does the ability to spread workload among three crew members overcome the additional interpersonal complexity of three-member teams?

The results of this analysis convinced regulators that two-person flight decks were no less safe than three-person flight decks. Boeing and Airbus, the two primary manufacturers of commercial aircraft, adapted the designs of their new aircraft and modified many older aircraft to eliminate the flight engineer position.

The pilots now became responsible for monitoring the automated systems using panels such as that shown in Figure 2. Although this potentially adds to the workload of pilots under some conditions, there have been no significant indications of safety problems related to the elimination of the flight engineer.

### 1.3.2 CRM

United Airlines Flight 173, A DC-8 en route from Stapleton Airport in Denver to Portland Airport, encountered unusual indications upon lowering the landing gear [National Transportation Safety Board (NTSB), 1979]. While the crew was troubleshooting the problem, the aircraft entered holding. Despite the relative simplicity of the problem—the gear problem was merely an indication problem—it would be nearly another
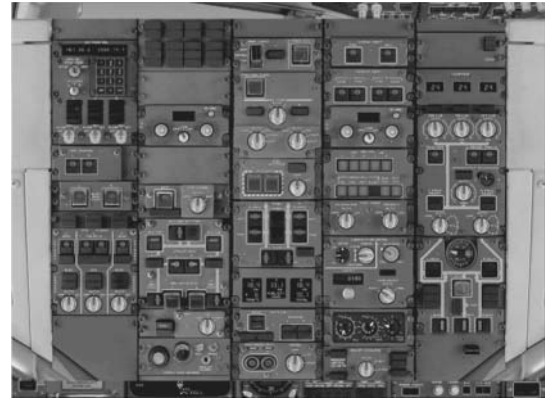


**Figure 2** Flight deck overhead panel showing systems controls. (This is a screenshot of copyrighted computer software; the copyright for the contents of the software is held by Level-D Simulations and Flight One Software, Inc.)

hour before the aircraft crashed due to fuel starvation. Although several queries and comments were made by the first officer and flight engineer about the fuel status, the captain did not appear to understand the actual fuel status. This accident highlighted a particular problem that had appeared to plague other incidents:

14. How do we ensure proper communication and coordination between crew members?

As a result of this incident, the NTSB recommended that all airlines ensure that crew members are trained in flight deck resource management. United Airlines complied, instituting a program in crew resource management (CRM) that was replicated by all other airlines and within the military. Good crew resource management is widely considered to be a significant factor in the relatively successful crash of a severely crippled DC-10 in Sioux City in 1989.

### 1.3.3 Alerting Systems

In an incident that presaged the crash of United 173, a Lockheed L-1011 aircraft, flying as Eastern Airlines Flight 401, encountered a similar problem with a landing gear indicator. The aircraft was placed on autopilot while troubleshooting occurred. However, during the troubleshooting the autopilot was inadvertently disconnected and the aircraft entered a very slow descent. Although an altitude warning was issued, it apparently was not heard due to fixation on the landing gear problem. In addition to the CRM-related problem, this accident highlighted a number of human factors issues related to automation and warning systems, including:

15. How do we ensure that pilots are aware of the status of the automation?
16. How do we ensure that alerts are sufficiently salient that pilots are aware of the situation?

This incident was one in a number of incidents that resulted in the introduction of new alerting systems onboard aircraft. In 1974, TWA flight 514 crashed on approach to Washington Dulles International Airport. This flight was one of the last in a series of "controlled flight into terrain" (CFIT) accidents that led to the requirement that commercial aircraft carry an operating ground proximity warning system. (That series of incidents included the 1970 crash of Southern Airways Flight 923, in which the Marshall University football team was killed.)

An encounter of a Lockheed L-1011 airliner with 152 passengers on board with "wind shear," a sudden shift in wind direction, speed, or both resulted in a crash that killed 135 people, including 1 person on the ground. This crash of Delta 191 in 1985 helped lead to the requirement for airliners to utilize wind shear detection systems.

Lastly, Aeroméxico Flight 498 collided with a general aviation aircraft approaching Los Angeles International Airport in August 1986. The crash claimed the lives of 82 persons and followed a number of in-flight collisions, including a very similar incident involving a Pacific Southwest Boeing 727 in which 144 persons died in San Diego, California, in 1978. These accidents led to the requirement that commercial aircraft be equipped with collision warning systems.

The introduction of multiple alerting systems, in addition to the numerous other alerting and status information displays in the aircraft, added complexity to the flight deck. With more alerting systems on board the flight deck, alerts were more likely to sound, even when unnecessary (false alarms), resulting in distraction to the pilots. In addition, aircraft malfunctions would sometimes result in a number of simultaneous alerts. While there have not been fatal accidents associated with these problems, they were listed as two of the primary factors in a number of incidents which are considered precursors to accidents (Rehmann, 1995). These two problems are:

17.  How do we design alerts so that pilots can properly prioritize actions when multiple alerts occur simultaneously?

18.  How do we design alerts so that they do not distract the pilot from higher priority duties?

### 1.3.4    Multifunction Displays

Cathode ray and computing technology enabled the replacement of single-sensor, single-instrument (SSSI) displays with multifunctional displays (MFDs). The adoption of MFDs was consistent with a number of human factors principles that arose from studying problems 8 and 9 but also added a new task to the flight deck—that of navigating through displays that now had *depth*. Moreover, controls for these displays had to be added. These two problems have close ties to usability research:

19.  What is the best design for controlling MFDs on a flight deck?

20.  What is the best interface design for MFDs on a flight deck?

These problems were somewhat unique to the flight deck due to the conditions on the flight deck, which include things such as that vision should be directed outside, there are a number of periods of very high workload, and both high- and low-frequency movement are prevalent.

### 1.3.5    Heads-Up Displays

The requirement for a good visual field outside the flight deck, the topic of earlier research, was present in commercial aircraft but was especially important in military aircraft. Early examples of information placed in the line of sight of the pilot such that the pilot could look through the information to the outside came in World War II when gun aiming was displayed on a transparent plate. (There were versions developed that projected information on the windscreen, but these were not adopted for use until after the war.)

In the 1950s, improvements were made to the projection of gun-sighting information on military aircraft. It was noticed that pilots could use the information to better fly the aircraft as well, so in the 1960s the information contained in the display was increased to include all the basic information. The symbology for this "heads-up display" (HUD), a name that distinguished the behavior of pilots from one who must look down to find the same information, was formalized by a French test pilot, Gilbert Klopfstein, in the 1960s and adapted during the 1970s for use as a primary flight instrument display (Deaton et al., 1989). However, research was needed to understand the HUD and how it differed from traditional instrumentation. In particular, the following problems were of interest:

21.  What effect does the use of a HUD have on flight performance?

22.  To what extent does the HUD improve detection of important environmental information external to the aircraft?

### 1.3.6    Communication

The deadliest accident in aviation history occurred in 1977 when two Boeing 747s, one with 396 persons on board and one with 248 persons on board, collided at a foggy Tenerife airport due to a fairly simple communication failure. The first aircraft, a KLM flight, mistook a route clearance as permission to take off while the second aircraft, a Pan Am flight, was still on the runway. The two aircraft collided, resulting in only 61 survivors, all of which were on the Pan Am flight. This accident highlights problem 10 on the Fitts panel list. As will be discussed, runway incursion incidents, primarily due to communications problems, continue to be one of the most vexing problems in air transportation.

23.  What principles should be used for communication between the different operators in the air traffic system?

## 1.4   1990s–Present

The last two decades have seen substantial changes to the system. The air traffic system has been undergoing modernization, as have the aircraft, with the introduction of new systems and with increasing reliance on automation and computing technology. These advances, however, have not come without difficulties.

### 1.4.1   Automated Collision Avoidance Systems

The danger of midair collisions reduced greatly after the introduction of radar at the end of the 1950s. Since that time, in the United States no two aircraft under radar control have collided. However, there have been collisions outside of the United States and near misses are not infrequent even within the United States.

To combat this danger, automated collision avoidance systems (ACASs) have been mandated on most commercial aircraft by national and international regulatory agencies. In actuality, the concept of an onboard collision avoidance system had been around since the 1950s as one response to the collision over the Grand Canyon in 1956. The systems were, for one reason or another, unreliable but continued to be developed and tested throughout the next decade (White, 1968).

It was not until the FAA focused researchers on systems that utilized the aircraft's transponder information, followed by the series of midair collisions mentioned above, that a system was actually developed and fielded (Kuchar and Drumm, 2007). Aircraft began flying with the systems in 1990.

However, aircraft have still collided, despite the presence of the system. These cases included problems with opaque failures of the system and with procedures for executing the resolutions of the system. In addition, it has been shown that pilots do not often comply with the resolution of the alerting system, which could be a precursor indication (Kuchar and Drumm, 2007). These problems can be summarized as follows:

24. What can be expected in terms of pilot response to a collision avoidance warning?

### 1.4.2   Mode Confusion and Envelope Protection

With new aircraft and with a two-person flight deck, autopilots were increasing in both capability and complexity. Autopilots were now capable of operating in numerous "modes" which do not have intuitive or salient differences but which can have profound consequences for the behavior of the aircraft.

In Toulouse, France, in 1994, a new Airbus A330 was being flight tested for certification of flight under engine failure conditions. The autopilot unexpectedly entered a mode where it attempted to capture the altitude input to it, which was 2000 ft, and pitched up to do so. The pilots were not aware of the mode and did not increase power or disconnect the autopilot until the aircraft had stalled. It crashed with no survivors.

A China Air A300 approaching Nagoya Airport in Japan, also in 1994, crashed after it inadvertently entered into a "go-around" mode. The pilots, not aware of the mode, attempted to continue the approach. With the autopilot and pilots fighting for control, the aircraft stalled and crashed, killing 264 persons.

More recently, two Bombardier Dash 8 aircraft have nearly crashed after their autopilots entered into modes other than those expected by the pilots. These incidents highlight a significant problem:

25. How do we design autopilots to prevent or minimize mode confusion?

### 1.4.3   AAS Failure in ATC

In the 1990s, the FAA contracted IBM to build an "advanced automation system" (AAS) to replace the aging automation infrastructure of the air traffic system. In addition to modernization, the system must have improved functionality and extensibility (Debelack et al., 1995). The procurement was terminated in 1994, and approximately $1.5 billion of the $2.6 billion invested was written off as a loss.

The program was criticized as overly ambitious; it tried to radically alter all aspects of the air traffic system nearly simultaneously. In addition, the program was caught in a nonterminating loop of specification and refinement. Since then, the FAA's approach has been to build and deploy smaller systems and capabilities in a more evolutionary approach.

### 1.4.4   Next-Generation Air Traffic

Recently, in response to projections of substantial delays in the air traffic system due to increasing demand in the face of static capacity, Europe and the United States have embarked on parallel plans to modernize the air transportation system. Implementation of changes suggested as part of these systems, while posing some technical challenges, will almost certainly hinge primarily on resolving a number of human factors issues. The systems proposed contain several significant changes, including:

- Trajectory-based operations—flying aircraft with respect to four-dimensional, three spatial dimensions plus time, trajectories
- Automated separation assurance—the allocation of the air traffic controller's primary function, separation assurance, to automation either on the ground or on the flight deck
- Increased coordinated information to pilots, controllers, and other stakeholders regarding weather and air traffic
- Robustness to disruptions such as weather and traffic congestion

These changes are expected to increase the capacity of the system to twice its current level or more. At the same time, the system must remain as safe or safer than the current system. This, however, poses a significant problem:

26. How do we predict the safety of a complex human-integrated system?

## 2  FLIGHT DECK HUMAN FACTORS

Much of the research and development in aviation human factors has been applied to the flight deck. Such work has been driven by the rapid expansion and advancement of flying in general, including commercially, where hundreds of millions of passengers are moved every year, and in the military, where the application of good human factors can make the difference between success and failure.

Substantial work has been done on the design and use of automation and the performance of the pilot reflecting human factors work that cuts across perception, information processing, decision making, workload, human error, usability, and physical ergonomics. The methods applied to study automation have included task analysis, modeling, information theory, control theory, and anthropometry.

### 2.1  Automation

Various types of automation have been introduced onto the flight deck. Perhaps the earliest forms of automation were autopilots, for control of the aircraft. These have proven highly successful. Next, various forms of information automation, mostly in the form of alerting systems, were deployed in response to a number of accidents. Lastly, decision support systems have been added to the flight deck.

#### 2.1.1  Control Automation

Autopilots have become an indispensable part of commercial flying. They allow for great reliability in maintaining the desired course, speed, and altitude and greatly assist in landing aircraft in very poor weather conditions. Overall, they relieve pilots of a great deal of workload. Autopilots, however, are not without problems, as indicated earlier. A number of principles have been developed based on research into these problems.

***Operators May Overrely on Reliable Automation***
Autopilots fit very clearly into the conception of a proper role for automation. Stabilizing an aircraft in the horizontal, vertical, and speed axes required extensive monitoring of these axes, a task for which humans are not well adapted, a finding which is well documented in the human factors research literature (Sheridan, 2002). Allocating this task to automation, then, should improve overall performance. For the tasks of stabilizing the axes of the aircraft, this is clearly true.

However, the humans would then have to monitor the automation, which was very reliable. Again, human vigilance is generally poor for such tasks. When such automation fails, the humans must detect and resolve the issues, problems that contributed to the L-1011 crash mentioned earlier. Research has led to the following principle:

***If Automation Has a Multiplicity of Modes, the Current Mode and Its Implications Should Be Transparent to the Operator***  The problem of mode confusion, as mentioned, is relatively recent. Generally, mode confusion occurs when the operator's

mental model is not coincident with the behavior of the system (Bredereke and Lankenau, 2005). In advanced, multimode automation, the operator must be provided with sufficient information to easily determine the mode in which the system is operating and must understand the consequences of being in that mode. Some of the mode problems that resulted in accidents were a consequence of subtle, easily confused cues as to the mode of the aircraft or even of opaque implications of being in a particular mode.

Better feedback alone, however, is generally not sufficient. In addition to having sufficient feedback, overcoming mode errors appears to involve training, workload management, and perhaps the use of multiple modalities (Sarter and Woods, 1995).

In general, although the problem is understood, the solution to the problem is not yet entirely clear. Complications in solving this problem include workload, vigilance, and the dynamic nature of mental models.

***Periods of Automatic Control Reduce the Operator's Situation Awareness***  Situation awareness (SA) refers to the pilot's knowledge of that information that is necessary for good performance. SA, whose measurement is discussed in a subsequent section, includes the knowledge of the status of important elements, knowledge of the meaning of that status, and knowledge of the future implications of that status (Endsley, 1995). SA has been extensively applied to aviation, where the concept has been discussed since the World War I (Hacker, 1984), although research on the topic did not occur until the 1980s.

The emphasis on situation awareness has grown in concert with the increase in information available to pilots and controllers and as automation further displaces the human operators from the work. As operators are removed from the task, they have been shown to lose situation awareness and performance declines (Endsley and Kiris, 1995). This tendency can be mitigated by either keeping the operator involved manually in the task or periodically allocating manual control to the operator.

#### 2.1.2  Information Automation and Systems

In addition to control automation, systems to support the information needs of pilots and controllers have been introduced. These include the provision of status information through single- and multiple-function displays and alerting systems.

One of the earliest problems to be considered by researchers involved information displayed on gauges. These issues included how to display the information, the arrangement of the displays, and whether to integrate information.

***Digital Displays Are Preferred When the Specific Reading Needs To Be Identified; When Precision Is Not Required, Fixed-Scale Displays with Moving Pointers Are Better***  Researchers have found that, when the precise numerical value is required from an instrument, digital displays are better than dial-type displays (Simmonds et al., 1981). In such cases, it is necessary to ensure that the display of the value

is of sufficient duration to allow the operator to read the value.

In other cases, dial-type displays are generally better. A number of subprinciples for these displays have been found (Sanders and McCormick, 1993), including that having the pointer move over a fixed scale is generally better than having the scale move against a fixed pointer unless the entire scale cannot be displayed; the movement should be consistent with the physical interpretation of the scale, such as vertical motion being presented on a vertical scale (Roscoe, 1968); multiple pointers on the same scale are generally not desirable and can result in reversal errors, as was the case with the altimeter-misreading accidents mentioned (Heglin, 1973); and for qualitative or check-reading instruments, where the precise value is not necessary, analog displays with the grouping and orientation designed to be consistent with a strong Gestalt, such as at the 9 or 12 o'clock position, produced the best performance (Dashevsky, 1964). Vertical scales are particularly effective for qualitative reading (Elkin, 1959), and performance can be enhanced by using color or coded markings that have intuitive meaning to the operator (Sabeh et al., 1958).

### Primary Flight Instruments Should Be Arranged in a "T" in the Center of the Pilot's Visual Field

A number of studies, primarily by Fitts, utilized eye tracking to determine where to position particular instruments. It was found that certain instruments, specifically the artificial horizon, the navigation display, the altimeter, and the airspeed indicator, were accessed the most, and best performance was obtained by positioning them directly in front of the pilot almost at eye level (Cole et al., 1954; Fitts et al., 1950). These studies led the Civil Aeronautics Board, whose name would later change to the FAA, to require a "basic T" arrangement for these instruments on all flight decks [Civil Aeronautics Board (CAB), 1953]. This original list included the vertical speed indicator and flight path deviation, which were later dropped from the list, although they still generally appear in the basic T arrangement.

### Place Related Information in Close Proximity to Allow for Easier Integration

Additional eye movement studies were conducted to examine whether certain pieces of information were "linked" or somehow related to one another. Information on separate instruments but which were closely related should be located next to one another whenever possible (Chapanis, 1959). Automation has since been developed roughly around this principle to help lay out instrument panels (Mendel and Sheridan, 1986; Pulat and Ayoub, 1979).

In general, locating information together when such information must be integrated is considered consistent with the "proximity compatibility principle" (Wickens and Carswell, 1995). In addition to locating related information proximally, the principle also covers cases in which the display has multiple, related pieces of information on it, such as the common integration of altitude and vertical speed on the same instrument. Multifunction displays are typically designed to be compatible with the proximity compatibility principle,

as one display usually only contains, on one screen, related information.

### MFDs Should Have Proper Organization and Minimal Depth

MFDs, although satisfying the proximity compatibility principle, could potentially suffer from poor organization (Seidler and Wickens, 1992). In addition, because the content of MFDs can be controlled manually or by computer, the display may have multiple pages (i.e., depth). If such a display is poorly organized, does not comply with usability guidelines, or has too much depth, the pilot could become lost while navigating the display pages (Allen, 1983; Francis, 2000; Schneiderman, 1998). Moreover, the operation of the display may require the pilot to have their head down, that is, not looking outside, which is generally considered poor practice.

### Operators May Not Be Able to Extract More Than Four to Six Distinct Pieces of Information from a Display

Miller (1956) roughly identified the number of objects one can keep in short-term memory as being approximately 7, although how to interpret this number has been the subject of some debate (Cowan, 2000), and clustering of related information can result in an apparently much larger set of information. When relating this to a pilot or controller's attention and processing of a visual scene, there is substantial debate, but it seems as if a similar number can be extracted (VanRullen and Koch, 2003).

The literature to support this principle is not settled. In general, operators likely quickly obtain a general but not detailed representation of a scene, then focus on specific items of interest (VanRullen and Thorpe, 2001). As each item of interest is focused on, its representation is suppressed in subsequent processing for a short period of time (Tipper and Driver, 1988).

### The Time Sequence of Operator Attention to Display Information Can Be Modeled as a Function of Expected Value, Including Salience, Effort, Expectancy, and Value

The factors that influence where attention is applied include size, color, and salience. A recent model applied to flight decks has found very good correspondence between predictions of areas of eye fixations and assessments of the salience and value of the information, of the effort needed to move to that fixation location, and the expectancy of what the pilot perceives is necessary (Horrey et al., 2006; Wickens et al., 2003).

### Minimize Clutter or Provide Decluttering Capabilities

While this may seem intuitive, some complexity has been identified with respect to the concept of clutter. In general, the response time of operators when given a stimulus that is "cluttered" by proximate non-stimulus items is longer than when the clutter does not exist (Eriksen and Eriksen, 1974). On displays with a great deal of information, pilots have been found to take longer to locate an item and may be unable to find items at all (Wickens, 2003). In addition, subjective ratings of complex instruments very often include criticisms of excessive clutter (Abbott et al., 1980).

However, clutter can come in more subtle forms. On an instrument, when given a particular task goal, one can identify "stimulus" versus nonstimulus items. However, it is, in general, difficult to know a priori how to categorize information in this way. In addition, operators have been found to be capable of extracting information effectively from extremely complex displays under some conditions (Landry et al., 2007).

HUDs represent an interesting case of clutter. On traditional instruments, additional symbology placed on the display can create clutter. On a HUD, either the symbology or the environmental information being viewed by the pilot through the HUD could be considered clutter. Research has suggested that this type of clutter does not affect pilot performance for nominal events (Levy et al., 1998; Ververs and Wickens, 1998) but may affect them for unexpected events (Wickens and Long, 1995).

***When Displaying Spatial Relationships, the Display Should Be Consistent with the Perceived Information in the Environment and Should Have Movement Consistent with a Person's Mental Model of the Motion of Elements in the System*** One of the concerns of instrument designers in the 1950s was related to the artificial horizon or "attitude indicator" (ADI). This instrument shows the orientation of the aircraft with respect to horizon, including the pitch and bank of the aircraft. (An example is shown in Figure 3.) Specifically, the instrument has a depiction of the horizon, with an overlay of aircraft wings.

Some designs for this instrument had a fixed earth depiction and the wings moved relative to that fixed earth, while others had fixed wings and the earth moved relative to the wings. These two depictions conflicted variously with the idea of either compatible motion or



**Figure 3** Attitude indicator.

pictorial realism, principles which had been developed for the general design of spatial instruments (Roscoe et al., 1981). While researchers found a compromise design and eventually demonstrated an improvement in performance using it (Beringer et al., 1975; Fogel, 1959), it was never implemented due to the strength of the legacy of the existing instruments. A number of accidents have been tied to this problem (Bryan et al., 1954; Roscoe, 1997b).

Also consistent with these principles is the idea that navigation problems are largely representational and they are more effectively solved by "distributing" some cognition to the environment. That is, some of the cognition needed to accomplish the task should be integrated into the display and not be required of the operator (Norman, 1993; Zhang and Norman, 1994). As such, good displays should eliminate the need for operators to perform complex calculations in their head. From this, one can understand why map-type displays are much more useful for navigation than instruments that only show relative position from a known location.

## 2.2 Performance

Operator performance is typically one of the primary measures for human factors researchers. As such, methods for evaluating, modeling, and predicting performance are extremely useful for system design and analysis.

### 2.2.1 Control-Theoretic Approaches

***Human Operators May Be Best Modeled as an Integrated Part of the System*** The mathematics of control systems and their reliability have advanced substantially over the decades since the first autopilot. Control theory allows for keen insight into the expected operating behavior of the system in which a feedback control system is embedded. The transfer functions of control systems provided an understanding of whether the system is *stable*, that is, that its outputs are always bounded for a given input.

In the 1950s and 1960s substantial effort was put into trying to determine the closed-loop transfer function of the pilot, such as is shown in Figure 4*a*. However, McRuer (McRuer and Graham, 1965) discovered that, while a general transfer function for the pilot could not be found, the transfer function for the pilot–aircraft system *combined*, as shown in Figure 4*b*, was fairly simple—the system, with the pilot controlling it, responded to an error by integrating the error with a certain amount of delay and gain.

While this result may not seem overly profound, the general concept is extremely important. As stated by Sheridan (Gerovitch, 2003):

> In 1957–60, [a] man named Duane McRuer, got a bright idea that the human really was so adaptable that he or she could adapt to different types of dynamic systems, and the right way to model the human was to model the whole closed loop: the human and the airplane, or whatever system he was controlling, together as a single entity. This was a great insight.
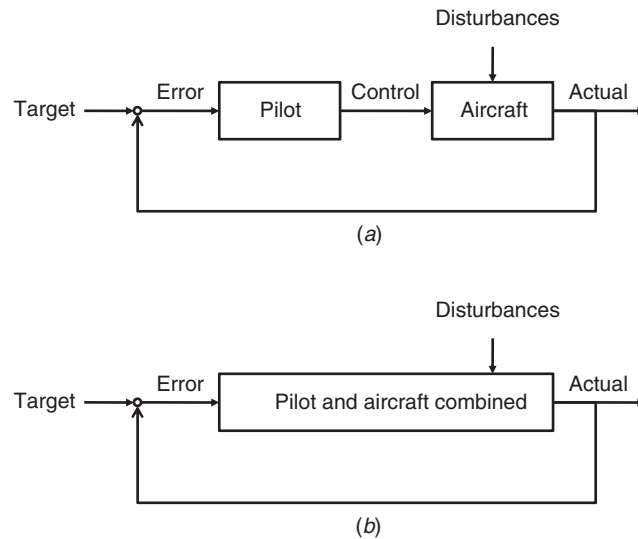
**Figure 4** (*a*) Closed-loop system with pilot as controller; (*b*) closed-loop system with pilot embedded in system.

Subsequently, control theory was applied to more complex tasks than simply tracking one axis, as was done in the McRuer work. This led to "optimal control models" of human responses (e.g., Kleinman et al., 1970). These, however, were difficult to apply to realistic problems of any substantial complexity, and work along this track in aviation has not progressed significantly since the 1970s.

### 2.2.2 Alerting System Response

The number of alerting systems on board aircraft has steadily increased over the decades since the first retractable landing gear, which most likely precipitated the first alerting system, that for warning of the landing gear not being safely deployed or stowed. A number of principles have been developed over the years with respect to alerting systems.

***Alerts Can Be Set Based to Warn of Noncompliance, to Warn of Exceeding Some Threshold of Risk, or to Control Some Performance Metric*** A number of aviation alerting systems base an alert on deviation of the aircraft from its expected behavior, usually subject to some buffer. An altitude alert is an example of this type of alert, where descending below or climbing above the altitude set into an altitude command window results in a tone indicating the deviation.

Another type of alerting system monitors the projected behavior of the system and warns based on the prediction that a hazard would be encountered if the no control is applied to the system. In such cases, the alert is, by definition, probabilistic. If the hazard cannot be avoided, the alert is not useful. Therefore, some threshold must be set for the alert to occur. That threshold may be based on a trade-off between certainty and time to respond (Hu et al., 2002; Kuchar and Carpenter, 1997) or may be based on ensuring that some avoidance

maneuver is available (Teo and Tomlin, 2003). Because the system is probabilistic, there are false alarms, where the hazard would not have occurred had the alert not been given, missed detections, where the hazard did not sound but should have, and even induced hazards, where the hazard occurred only because the alert sounded.

Yang and Kuchar (2002) have suggested that alert thresholds can be set to control a specific performance indicator of the system, such as a minimum rate of missed detections or a maximum rate of false alarms. Such systems directly control the performance criteria of interest, rather than requiring them to be computed as is the case with other types of alerts.

***Excessive False Alarms Can Result in Disuse of the System*** Initial implementations of the Ground Proximity Warning System (GPWS), which warned of impending collision with the ground, were prone to false alarms. The alert would sound, even though the hazard did not exist. Moreover, there were a number of cases where the GPWS was ignored despite the hazard existing—a number of controlled flight into terrain (CFIT) accidents occurred as a result.

The incidence of CFIT, even with the GPWS alerting the crew, was troubling. Why would the crew ignore an alert that was informing them of probably the most hazardous situation?

One explanation was that the aircrews had lost their trust in the system due to the high rate of false alarms. That is, the system had "cried wolf" too often and it was no longer being listened to by the aircrew (Bliss et al., 1995). This effect has been found empirically to be affected by workload and the use of redundant information (Bliss and Dunn, 2000; Selcon et al., 1995).

### 2.2.3 Naturalistic Approaches

There is a belief among a number of human factors researchers that the behavior of the operator cannot be

extracted from the environment in which the operator is performing the task. That is, the person's behavior is inextricably linked to the specifics of the situation—it is embedded. Under such a view, the types of analyses, and even the units of analysis, change.

One such view holds that work is "situated" and that the work and how it is done will change and also that the use of the system will be altered by being present in the work environment (Suchman, 1987). In another view, decision making cannot be viewed abstractly from the exact conditions under which the decisions are made by the person, either due to framing issues (Gigerenzer and Goldstein, 1996) or due to the complexity of the environment (Klein et al., 1993). These views have been applied to aviation.

***Analysis of Work Should Contain an Appreciation for the Context in Which the Work Is Conducted*** There are sufficient safeguards built into aviation operations so that simple slips or lapses are unlikely to cause errors; instead, decision errors in aviation tend to be cases where the operator's intent itself is incorrect due to a mistaken interpretation of the environment. Such errors can be viewed in the context of naturalistic decision making, where a true understanding of the mistake can only be obtained by understanding the totality of the circumstances in which the pilot was immersed. Such problems have been found in a majority of cases where crew errors were a causal factor (Orasanu and Martin, 1998; Simpson, 2001).

### 2.2.4 Workload, Vigilance, and Situation Awareness

Workload and vigilance have been significant concerns in aviation since its inception. Situation awareness is a topic that has only been addressed more recently. General principles of workload, vigilance, and situation awareness, as applied to aviation, are presented here.

***Workload Should Be Kept at Moderate Levels for Best Performance*** For most complex tasks, performance has been shown to be poor at both low workload levels, where vigilance is low, and high workload levels, where the operator has difficulty in completing tasks (Hancock et al., 1995).

This simple guideline is effective but may be overly simple. Workload is a general term and has been shown to have somewhat independent aspects of mental workload and physical workload (Hart and Staveland, 1988). Perceptions of workload can also be influenced by time pressure and performance. Moreover, in aviation workload is often dynamic, with periods of high workload interspersed with periods of very low workload.

This latter concern is being addressed by the concept of adaptive or adaptable function allocation, where "adaptive" refers to function allocations controlled by automation and "adaptable" refers to function allocations controlled by the human operator. This concept is discussed further in Chapter 59.

***Crew Vigilance Is Poor for Monitoring Reliable Systems*** Failures of vigilance, particularly for

monitoring automation, have been found to be a significant source of aviation incidents (Malloy and Parasuraman, 1996). Moreover, such vigilance problems seem to occur on either end of the workload spectrum—too little workload or too much workload.

In general, however, research into vigilance has had little impact on the aviation system (Wiener, 1987). Aircrews continue to fail to monitor, such as a recent incident in which an aircraft overflew its destination by 150 miles before air traffic controllers could reach the aircrew on the radio. The crew was reportedly distracted and failed to notice they were not descending toward their destination.

This problem is still significant. Proposals for the next-generation system include substantially more automation and yet do not propose a specific task for the human operators. It is inferred that humans will monitor such automation, despite the findings regarding how poor humans are at such a task.

***Design to Maximize Situation Awareness*** Situation awareness, when viewed as a product of a process of situation assessment, is heightened when the operator is significantly involved in the task. Specifically, when human operators are asked to monitor automation performing a task instead of performing it themselves, their situation awareness suffers (Adams et al., 1995; Endsley and Kaber, 1999). This finding corresponds well to the idea that direct experiences result in deeper cognitive processing and are recalled more easily (Craik and Tulving, 2004).

One part of situation awareness relies on effective perception of information present in the environment. Reducing irrelevant information can improve such perception (Yeh et al., 2003). In addition, a thorough task analysis should be performed to ensure that the necessary information is available to the operator (Endsley et al., 2003; Kirwan and Ainsworth, 1992).

### 2.2.5 Flight Crew Interactions

Flight crews are essentially a team of individuals collaborating to enable the aircraft to reach its destination efficiently and safely. As with any team, the dynamics of interpersonal relationships can have a significant impact on performance.

***Although Ultimate Authority Rests with the Captain, All Flight Crews Should Feel Empowered to Identify Problems and Raise Concerns*** The principles of crew resource management, borne from accidents that may not have occurred if crew members had not deferred to the authority of the captain, insist on full participation of all crew members. Such principles can be particularly difficult to apply in cultures that value deference to authority but are critical to leveraging the expertise and simple redundancy of multiple crew members.

### 2.3 Fatigue and Circadian Rhythms

Aircraft on long flights typically cross numerous time zones, and both passengers and crew are exposed to an atmosphere equivalent in pressure to approximately

8000 ft altitude for nearly the entire duration of flight. These aspects mean that both passengers and crew members frequently suffer from fatigue and disruptions of their circadian rhythms.

The human sleep cycle is regulated by a number of factors, including internal brain processes and external factors such as lighting and the timing of meals. Inconsistencies between the expectations of the internal brain processes and these external factors can result in sleep disruptions, which in turn can result in fatigue.

Numerous remedies or mitigations have been suggested to treat the symptoms of these problems. A number of the factors known to help keep a regular circadian rhythm or to most quickly adjust a circadian rhythm include:

- Regular exercise
- A consistent routine
- Staying well hydrated
- Relaxation techniques such as meditation or reading
- Avoiding strenuous activity while your body adjusts to a new cycle
- Trying to remain exposed to sunlight during daylight hours and avoiding bright lighting at night
- Avoiding stimulants or depressants, such as caffeine or alcohol, if you know your sleep cycle will be disrupted
- Adjusting to the day–night hours as early as possible, including prior to leaving on the trip

In general, it is expected that, for each hour of time zone change, it will take about one day to adjust the internal clock. During the period of adjustment, fatigue is likely to occur. Fatigue has a number of effects on behavior and performance, including reduced vigilance, poor memory, scattered attention, and an increase in response time.

## 3 AIR TRAFFIC CONTROL HUMAN FACTORS

The job of air traffic controllers is highly manual. Their communication with each other and with pilots is through radio voice channels. They manually monitor a radar scope and project aircraft trajectories to detect potential problems. As a result, their performance has not been impacted by automation in the same way as it has for pilots, although under next-generation air traffic plans in Europe and the United States that may change.

### 3.1 Automating Air Traffic Control

As mentioned, air traffic control is highly manual. However, there are a few areas in which automation has been introduced. In general, controllers are, justifiably, highly conservative regarding changes to their displays and work practices, making the introduction of sophisticated automation difficult.

### 3.1.1 Radar

Prior to radar, controllers updated aircraft positions and projections of position on a status board. Using this information, they could identify aircraft whose projections of positions were in conflict. Under such a control scheme, the deconfliction task was primarily a comparison task, where the controller simply needed to compare the positions and note any flights expected to reach common points and the same, or nearly the same, time.

Radar altered this task substantially. The task was now primarily a visual comparison, requiring a projection of the flight path of an aircraft while incorporating knowledge of the intent of the aircraft. (Intent can be inferred from the flight plan and other information.)

### 3.1.2 Flight Strips

***Electronic Representations Are Generally Inferior to the Physical Artifacts They Are Expected to Replace: It Takes Substantial Effort to Create a Suitable Replacement*** After the advent of radar, flight information was kept on specifically formatted strips of paper, called "flight strips." Flight strips would store basic routing and identification information, along with notes made by controllers about the flight. It would be physically handled by the controller in charge of the flight and, if possible, physically passed to the next controller to take charge of the aircraft. An example is shown in Figure 5. Flight strips would be held in an ordered "strip bay."

As automation was added to the air traffic system in the 1960s and 1970s, printers for these strips were also added, but little else was automated. Much of the information was still hand-written on the flight strips, and they were still physically handled by controllers. This physical possession, which was a highly salient cue for who was responsible for the aircraft, proved difficult to replace (Harper and Hughes, 1991).

For controllers, their situation awareness, or "picture," of the air traffic is of paramount importance (Whitfield and Jackson, 1983). The manual handling of flight strips, including the writing of notes, enhances situation awareness. Mediated interaction does not seem to provide the same level of situation awareness and has been heavily resisted by some air traffic controllers.

Flight strips are still used by controllers today, although an automated system, called the "User Request Evaluation Tool" (URET), provides an electronic version of the flight strip (Arthur and McLaughlin, 1998) in the United States, with similar systems having been developed in Europe (Berndtsson and Normark, 1999). URET is used as a flight strip replacement tool. As new controllers, more comfortable with electronic technology, enter the profession, it is expected that manually printed and handled flight strips will be entirely phased out.

### 3.1.3 Conflict Alert

***Not All Alerting Systems with High False-Alarm Rates Result in Underreliance*** One of the few alerting systems available to controllers is the "conflict alert" system. This system projects aircraft trajectories

| 3 4 5 6 7   1   2 | 11 | 15          16 | 20 | 21 | 25 | 27 28 |
|---|---|---|---|---|---|---|
|     8 | 12 |  |  | 22 |  |  |
|     9 | 13 |  |  | 23 |  |  |
|  | 14 | 17   18 |  |  |  |  |
| 10 | 14a | 19 | 20a | 24 | 26 | 29   30 |

| DAL542            1 | 7HQ | 30 / 18 | 330 |  | FLLJ14 ENO 000212 COD PHL | 2675 |
|---|---|---|---|---|---|---|
| H/B753/A | 1827 |  |  |  |  |  |
| T468 G555 |  |  |  |  |  |  |
|     16 16 |  |  |  |  |  |  |
| 486      09 |  | PXT | RA↑1828 |  |  | *ZCN |

**Figure 5**  Flight strip example.

and warns when a procedural separation violation is expected to occur within 3 min. Because the system uses simple dead reckoning to predict future positions, and because of the 3-min look-ahead time, false alarms occur frequently. Despite the high false-alarm rate, controllers still rely on the automation, defying the "cry wolf" effect found in other alerting systems prone to false alarms (Wickens et al., 2009).

If the conflict alert notification is ignored or if it does not occur and aircraft lose separation, the "operational error detection program" (OEDP) alert occurs. The OEDP results in an automatic data dump of information related to the occurrence, and the supervisor is notified. Typically, the controller is immediately removed from the position and is kept away from controlling operational traffic until remedial training occurs. Controllers can be fired or reassigned due to having excessive operational errors. Because of the seriousness of the occurrence of such operational errors, controllers routinely pad separation between aircraft to ensure the OEDP does not activate (Cotton, 2003).

Both conflict alert and the OEDP are fairly old technology. URET, in addition to being a flight strip replacement system, also has a "conflict probe" capability that can predict conflicts along the intended route of flight. Such a system would be less prone to false alarms than the conflict alert system.

### 3.1.4  Traffic Management Advisor

In the early 1990s, the National Aeronautics and Space Administration (NASA) adapted algorithms used in an aircraft's flight management computer to simulate the trajectories of all aircraft in an en route center. The resulting suite of tools, called the Center-TRACON Automation System (CTAS), could generate a prediction of an aircraft's arrival at a given point along its route of flight based on its flight plan, models of the aircraft type, airspace restrictions, winds, and a number of other factors.

From this, expected times of arrival for aircraft at their destinations could be generated. By applying separation requirements, a schedule of aircraft arrivals could be computed to coordinate the arrival of aircraft at a given destination. That coordination would be accomplished by displaying, to each controller involved in the control of the arriving aircraft, the amount of delay that controller was responsible for imparting to the aircraft.

The resulting system is called the Traffic Management Advisor (TMA). After an extensive set of human factors analyses on TMA, the system is in place in the air traffic control system today (Lee et al., 2000). Similar systems have been developed worldwide (Barco Orthogon, 2003; Ljundberg and Lucas, 1992; Nav-Canada, 2003; Robinson et al., 1997).

### 3.2  Conflict Detection and Resolution

A controller's primary function is to separate aircraft according to procedural rules. These rules are designed to ensure that collisions do not occur between aircraft, but they rely heavily on a controller's capability to detect potential problems and resolve them without creating additional conflicts. It has been proposed that, in future systems, automation may be used for this function. Such a shift would have profound human factors implications for the system, implications which are not yet known.

### 3.2.1  Principles for Manual Detection and Resolution

Currently, the process of detecting and resolving conflicts by controllers is manual. They must mentally extrapolate the trajectories of aircraft into the future and predict when aircraft will lose separation. Controllers are extremely adept at this task, but not much is known about exactly how it is done.

***Controller Ability to Predict Conflicts Appears to Be Influenced by the Geometry of the Encounter and by Gestalt Effects***  A number of factors have been shown to influence the detection of conflicts. In one study, novices were found to be influenced by the visual "cluster" of the aircraft, where such clusters were influenced by Gestalt factors of the traffic display (Landry et al., 2001). Specifically, if two conflicting aircraft were in the same visual cluster, as indicated by a high degree of transitions between them as compared to other targets, then the novices were more likely to detect the conflict than if they were in different clusters.

The geometric arrangement of the aircraft also influences whether a conflicting pair will be detected successfully. Specifically, the following factors seem to affect conflict detection performance:

- Speed differences, with larger speed differences having a negative effect on detection (Rantanen and Nunes, 2005)
- Aircraft pairs with acute angles of incidence that are harder to detect than those with obtuse angles (Remington et al., 2000)
- Aircraft pairs that are further from the point of intersection of their flight paths and thus are harder to detect as conflicts (Stankovic et al., 2008)

In general, it is believed that controllers consider three general factors: altitude, where co-altitude aircraft represent a higher risk; convergence of the flight paths; and speeds, where similar speed and distance from the convergence point would represent a higher risk. In general, however, little is known about the psychological processes involved in the controller's conflict detection task (Neal and Kwantes, 2009).

### 3.2.2 Principles for Automating CD&R

The automation of conflict detection and resolution will be extremely challenging from a human factors standpoint. Concepts for such automation include a centralized system, where ground-based automation provides CD&R, a fully decentralized and distributed system, where aircraft provide CD&R, or some combination of these concepts.

***Flight-Deck-Based CD&R Must Overcome the Natural Competition of Aircraft from Different Airlines and Not Impose Substantial Workload on the Flight Deck*** Currently pilots are responsible for the safety of the flight, including the prevention of collisions. However, their capability to do so is limited to tactical avoidance using visual means or mediated by a collision avoidance system. The full deployment of CD&R to the flight deck would necessitate that strategic considerations of traffic be incorporated into the decision making on the flight deck. Currently aircraft from different airlines compete with one another, calling into question their ability to impartially resolve conflicts.

Of particular concern would be the additional workload imposed on the flight deck. During the cruise portion of flight, workload is relatively low and most of the surrounding traffic is in level flight, simplifying CD&R. However, during the climb and descent phases, workload is very high in the flight deck, and the pilots' ability to perform CD&R may be limited.

***Ground-Based CD&R Cannot Be Supervised by a Controller, But Instead Must Be Capable of Graceful Reversion to Manual Control*** Above about 1.5 times the current level of traffic, controller performance drops precipitously (Prevot et al., 2008a). Under ground-based CD&R, traffic would be expected to be substantially above this level. With the accompanying loss in situation awareness, it is unlikely that controllers could successfully monitor a ground-based CD&R system.

In the case of failure of a ground-based CD&R system, backup systems must be capable of moving the system gracefully, and without controller intervention, from the high level of traffic controlled by automation back down to the level of traffic at which the controller is capable of performing the function. Moreover, the controller must be brought back into the loop while this reversion is occurring.

***A Mixed CD&R Must Provide Clearly Delineated Lines of Responsibility and Authority*** In a mixed environment, controllers or ground-based automation would have responsibility for certain aspects of CD&R, while pilots or flight deck automation would have responsibility for other aspects. Similar to an operator's interaction with automation, the authority and expected behavior of the system must be clear to the operators to avoid gaps in perceptions of responsibility.

## 4 AVIATION MAINTENANCE AND SAFETY MANAGEMENT

Apart from "pilot error," a catch-all phrase that includes mistakes resulting from poor human–machine system design, maintenance errors have caused a substantial portion of serious accidents. Recently, there has been a surge of interest in the human factors of maintenance in aviation.

Overall, analysis of the safety of the system has been reactive. When accidents occur, an analysis is done to determine the cause of the accident and remedies are suggested. In the last several decades, aviation regulators have begun using voluntary, anonymous reporting systems such as the Aviation Safety Reporting System (ASRS) in the United States. Such systems track *incidents*, which are situations perceived as hazardous by air traffic controllers or aircrews, but which did not result in an accident. Incidents are considered precursors to accidents, so identifying safety problems based on incidents may prevent accidents from occurring.

### 4.1 Maintenance

In 1927, Brunat (1927) indicated that a substantial fraction of deaths (5%) and injuries (19%) were due to maintenance problems. This compares well with a 1994 report in which 12% of accidents were due to maintenance problems (Marx and Graeber, 1994). (That is not to say there has not been improvement but merely that the relative percentage of accidents caused by maintenance has not substantially changed.)

Maintenance workers must contend with a number of issues, including poor documentation and procedures, fatigue due to shift work, and teamwork issues. In addition, aircraft are complex systems, requiring substantial expertise, and typically maintainers are experts on only one portion of the system. One study of maintenance errors examined the prevalence of

a number of contributing factors: fatigue, pressure, coordination, training, supervision, previous deviation, procedures, equipment, environment, and physiological factors (Hobbs and Williamson, 2003), which was crossed with the type of error committed, such as slip, violation, and perceptual error. Pressure was the leading contributing factor, and the leading type of error was memory lapse. However, many of the contributing factors and types of errors had a substantial percentage of occurrences attributed to them. A more recent report has added "group norms" to the conditions that may contribute to errors (Hobbs, 2008).

In general, little work has been done on the human factors of maintenance work, despite its apparent importance.

### 4.2  Safety

One of the significant challenges in modernizing the air traffic system is the requirement that safety be maintained at extremely high levels. Accident rates, in the number of accidents per flight hour, are in the range of one in every million flight hours. Keeping such a high level of safety is difficult and is compounded by the difficulty in applying traditional engineering methods to complex human-integrated systems. Because of this, human error analysis methods have been developed, although they also suffer from specific weaknesses. The concept of system safety has been applied, although primarily to understanding accidents that have already occurred. Most recently, the FAA has adopted a reporting system called the "safety management system" (SMS) to help track safety problems and identify precursors to accidents so that they can be prevented.

#### 4.2.1  Reliability Methods

***Probabilistic Methods Do Not Produce Reliable Estimates of Safety in Complex Human-Integrated Systems Because the Distributions of Failures Are Not Known***   Many engineering systems undergo reliability analysis based on a probabilistic analysis of the chance of overall system failure based on aggregations of the likelihood of failures of components or subassemblies and the relationship between those components, subassemblies, and the overall system (O'Connor, 2002). That is, one identifies how the failure of individual components relates to the failure of subassemblies, how the failure of subassemblies relates to the failure of higher level subassemblies, and so on, until the probability of failure of the overall system is quantitatively established. Fault trees are a common method used, whereas more complex systems are modeled using such methods as Petri nets or Markov chains (Rauzy, 2008; Schoenig et al., 2006; Volovoi, 2004).

However, such systems are unlikely to produce good estimates of failures in complex human-integrated systems. One reason for this difficulty is that the probability distributions that describe human failings are not known and are likely to be complex and sensitive to minor environmental variation. Moreover, most probability-based methods require an understanding of the dependency of component failures and work best when the failures of components are independent. It is

likely that human failures are highly dependent on the functioning of other components in the system.

#### 4.2.2  Human Operator Error Analysis Methods

***Human Operator Error Analysis Methods Can Be Used to Trace the Consequences of Specific Errors in a Human-Integrated System***   Human operator error analysis methods include process hazard analysis, root cause and barrier analysis, and failure mode and effects analysis (Dhillon, 2007). These methods, which rely on enumerating the various failures that can occur within the system, are most effective when attempting to control the consequences of specific, known human errors. Given a particular type of error, these methods allow for a complete tracing of the error through the system to identify the effect of the system. They also allow for an analysis of specific measures that can be applied to mitigate the consequences of those errors.

However, in these methods, failure to comprehensively enumerate failures can result in significant analysis errors (Johnson, 2007). Moreover, it is impossible to know whether the enumeration used is complete—there are no formal methods for establishing the completeness of the enumeration.

Additional difficulty is encountered when trying to use enumeration methods for predicting the safety of modified or new systems. For such systems, it is difficult to enumerate failures, as it is likely that previously unencountered failures will occur. In safety-critical systems such as the air transportation system, it is typically these difficult-to-identify errors that result in accidents.

#### 4.2.3  System Safety

***Safety Can Be Viewed from a Systems Perspective as the Ability Of Human and Automated Agents to Keep the System from Entering Unsafe States***   An alternative approach to safety analysis is the notion of "system safety" (Leveson, 2004). This approach views safety as a control problem, where one must ensure that agents in the system have sufficient control to prevent the system from entering states that are considered unsafe.

The approach has been almost exclusively applied to accident analysis using the "Systems-Theoretic Accident Model and Processes" (STAMP) approach. This method deconstructs an accident in terms of the features that allowed the system to enter the unsafe state. These features may be related to operators, procedures, regulations, mechanisms, or any combination.

The strength of this approach is the systemic view. It is consistent with the embedded view of human behavior adopted by such methods as naturalistic decision making, and avoids seeking a root or principal cause that tends to oversimplify accident analysis.

However, the method is retroactive. There is as yet few examples of its application as a proactive tool for predicting safety (Landry et al., 2010).

#### 4.2.4  Safety Mangement System

In 2006, the FAA introduced the "safety management system" (SMS). SMS is a framework to capture data on, and manage, safety in the air traffic system. SMS is

a set of processes designed to allow for the accounting and control of risk in aviation. It integrates safety within the management structure of the FAA, providing insight and data for decision making as well as a structure for review and response to safety problems.

With the introduction of SMS, the FAA has formalized processes for establishing safety policies, for the review and management of risk, for continuously evaluating the safety of existing and proposed structures, and for promoting safety.

## 5  EMERGING ISSUES IN AVIATION

Change is being impressed on aviation, a system generally reluctant to change. Technology, such as remotely piloted vehicles and satellite navigation, has been developed that is likely to have significant consequences for the air traffic system. New methods are needed to understand the impact of these new technologies, so that they can be implemented and expand the capacity and capability of the air traffic system.

### 5.1  Remotely Piloted Vehicles

Remotely piloted vehicles (RPVs), also known as uninhabited aerial vehicles (UAVs), are increasing rapidly in number and variety. RPVs, currently used mainly for military and intelligence purposes, are also being considered for a wider range of tasks, including private commercial interests.

This increasing use of RPVs is bringing a number of human factors issues related to the vehicles to the forefront. These issues can be generally grouped into control issues, interface and function allocation issues, and air traffic issues. The first two issues have combined with the less stringent manufacturing requirements to result in a high mishap rate for RPVs, while the latter issue has been a strong impediment to wider usage of RPVs.

### 5.1.1  RPV Control Issues

The control of RPVs differs from that of a conventional aircraft primarily in that (a) the visual field is mediated and (b) there is no vestibular or haptic cueing. (In some cases, a time lag is also present.) These aspects significantly complicate the control of the vehicle by reducing the amount of information available to the operator of the vehicle.

The variety of vehicles complicates analysis in that there are numerous types of vehicles, with very different dynamics and with different control systems. However, the loss of vestibular and haptic cueing is a substantial loss with respect to controllability—pilots can no longer feel things such as the effect of engines failing, of ice accumulation, or of an approach to stall. The delay between the haptic or vestibular cueing a pilot would normally receive and the detection of visual indications of such conditions is significant and will often be the difference between preventing an accident or not.

The mediated visual field restricts field of view and diminishes depth perception. These aspects impair an operator's ability to control the vehicle, particularly upon landing, when peripheral cues are especially valuable.

Some RPVs are operated remotely from takeoff to landing, while others may be launched and recovered visually. (This is particularly true of small, tactical RPVs.) There are substantial differences in the control problems experienced by operators who can see the vehicle; in particular, they often suffer from control reversal problems depending on the orientation of the vehicle with respect to their position.

In a larger sense, there are issues about the level of control that is appropriate for RPVs. Some RPVs are controlled in a fully manual sense, where the pilot's control actions directly move the control surfaces of the vehicle. In other cases, the pilots essentially program the vehicle's trajectory and do not directly control the vehicle's control surfaces. Each of these strategies has advantages and disadvantages, and it is not yet known under what conditions each may be desirable.

### 5.1.2  RPV Interface Issues

RPV interfaces vary widely and, because they are fairly new, do not always use consistent symbology or comply with standard human factors guidance. RPV operators are working with human factors engineers to improve symbology and display elements. However, the capability of providing better visual information to the pilot is limited by the bandwidth of the communications channel connecting the operator's station with the vehicle.

In addition, it is expected that operators may eventually be controlling multiple RPVs, especially once these systems achieve sufficient autonomy. In such a case, RPV operators may become supervisors of multiple semi-independent systems, with the attendant human factors problems associated with such supervisory control.

### 5.1.3  RPVs in the Airspace System

Currently, operating an RPV within the airspace system requires substantial effort and forethought. A waiver must be obtained from the national airspace authorities in most countries, a process that can take a substantial amount of time.

The main reason for this caution is that most vehicles are expected to be able to operate in a predictable manner, and, when losing contact with air traffic control, to be able to remain visually separated from other vehicles. RPVs have no effective equivalent to visual operations when the radio link to the vehicle is compromised. In such cases, the behavior of the vehicle is not considered sufficiently predictable and could therefore pose a substantial hazard to commercial air traffic.

Until RPVs can demonstrate sufficient predictability, it is unlikely that they will be allowed easy access into the airspace system. Even in countries that have lowered some restrictions on RPVs, they have done so only for small vehicles, those weighing just a few pounds, which, even if they were to collide with a manned vehicle, would be unlikely to produce significant damage.

### 5.2  Free Flight

In the 1990s, as satellite-based navigation became possible, it was envisioned that aircraft may be able to fly any desired trajectory as long as they could keep themselves separated from other aircraft. The concept was

dubbed "free flight" and spawned substantial research into the human factors implications of such a concept (e.g., see Bilimoria et al., 2003; Cotton, 2003; Eby et al., 1999; Johnson et al., 1997; Yang and Kuchar, 1997).

It has largely become apparent that uncontrolled flight will not be possible, although it may be possible in portions of the airspace. In such cases, the responsibility for separation assurance will be "distributed" to the flight deck. However, the implications of such concepts are not well known and demand additional human factors analysis.

### 5.3 Automated Conflict Detection and Resolution: Roles and Responsibilities

One of the primary reasons for automating conflict detection and resolution is to overcome the limitation of the controller, who is capable of handling only about 12–15 aircraft within the volume of airspace for which they are responsible. By automating conflict detection and resolution, it is expected that substantially more aircraft could be managed in the same volume. However, controllers will not be capable of controlling the resulting system, which means that they cannot identify automation mistakes and could not take over for the system should it fail. This raises substantial questions about just what controllers would do in a system where the conflict detection and resolution system was automated.

A number of possible operational concepts have been proposed (Dwyer and Landry, 2009; Krozel et al., 2000; McNally and Gong, 2007; Prevot et al., 2008a, 2008b). However, none of these concepts define a specific allocation of function between automation and human based on human factors principles.

### 5.4 System of Systems

The airspace system is made up of a large number of heterogeneous agents who are collaborating, in the sense of sharing the same goal; cooperating, in the sense of working together but toward idiosyncratic goals; and competing, in the sense of having mutually exclusive goals. Moreover, each of these heterogeneous agents exist within a system that could be analyzed and understood separately from the entire air traffic system, such as each aircraft or each airline.

The methods of analysis used for such individual systems, including pilots and controllers, do not seem fully sufficient. In a system as complex as the air transportation system, there are emergent features that can only be understood by taking a systems view and examining the interaction between the heterogeneous agents. New methods are needed to truly understand the behavior and performance of the human agents within the system.

## 6 FUTURE CHALLENGES AND CONCLUSIONS

A substantial amount of human factors research has been applied to the aviation system. Early research work centered on issues of display form and organization, with later work dealing with higher level cognitive issues such as situation awareness and function allocation. Since the early days of aviation, flight decks and air traffic control systems have grown in complexity and capability. Human factors principles have been needed to guide design and ensure the safety of the system as it evolves.

The future appears to be one of an increasing pace of change. If the vision of the U.S. and European plans for the next-generation air traffic system is realized, the system of 2040 is likely to be substantially different than the system of 2010. In order for that vision to be achieved, human factors researchers and engineers must meet the challenge to identify the proper allocation of function in a highly automated system and develop ways to define the safety of the resulting system.

## REFERENCES

Abbott, T. S., Mowen, G. C., Person, L. H., Jr., Keyser, G. L., Jr., Yenni, K. R., and Garren, J. F., Jr. (1980), "Flight Investigation of Cockpit-Displayed Traffic Information Utilizing Coded Symbology in an Advanced Operational Environment," NASA Technical Paper 1684, NASA Langley Research Center, Hampton, VA.

Adams, M. J., Tenney, Y. J., and Pew, R. W. (1995), "Situation Awareness and the Cognitive Management of Complex Systems," *Human Factors*, Vol. 37, pp. 85–104.

Allen, R. B. (1983), "Cognitive Factors in the Use of Menus and Trees: An Experiment," *IEEE Journal on Selected Areas in Communication*, Vol. 1, pp. 333–336.

Arthur, W. C., and McLaughlin, M. P. (1998), *User Request Evaluation Tool (URET): Interfacility Conflict Probe Performance Assessment*, MITRE, McLean, VA.

Barco Orthogon (2003), "Osyris," available: http://www.barco.com/barcoview/downloads/BVW_Osyris_new_6p.pdf.

Beij, K. H. (1933), "Aircraft Speed Instruments," Technical Report No. 420, National Advisory Committee for Aeronautics, Washington, DC.

Beringer, D. B., Willeges, R. C., and Roscoe, S. N. (1975), "The Transition of Experienced Pilots to a Frequency-Separated Aircraft Altitude Display," *Human Factors*, Vol. 17, pp. 401–414.

Berndtsson, J., and Normark, M. (1999, November), "The Coordinative Function of Flight Strips: Air Traffic Control Work Revisited," paper presented at the ACM Group99: International Conference on Supporting Group Work, Phoenix, AZ.

Bilimoria, K. D., Sheth, K. S., Lee, H. Q., and Grabbe, S. R. (2003), "Performance Evaluation of Airborne Separation Assurance for Free Flight," *Air Traffic Control Quarterly*, Vol. 11 (2), pp. 85–102.

Bliss, J. P., and Dunn, M. C. (2000), "Behavioural Implications of Alarm Mistrust as a Function of Task Workload," *Ergonomics*, Vol. 43, No. 9, pp. 1283–1300.

Bliss, J. P., Gilson, R. D., and Deaton, J. E. (1995), "Human Probability Matching Behaviour in Response to Alarms of Varying Reliability," *Ergonomics*, Vol. 38, No. 11, pp. 2300–2312.

Bredereke, J., and Lankenau, A. (2005), "Safety-Relevant Mode Confusions—Modelling and Reducing Them," *Reliability Engineering and System Safety*, Vol. 88, pp. 229–245.

Brunat, H. (1927), "Safety in Airplane Flight," Technical Memorandum No. 438, National Advisory Committee for Aeronautics, Washington, DC.

Bryan, L. A., Stonecipher, J. W., and Aron, K. (1954), "180-Degree Turn Experiment," *University of Illinois Bulletin*, Vol. 54, No. 11, pp. 1–52.

Chapanis, A. (1959), *Research Techniques in Human Engineering*, Johns Hopkins Press, Baltimore.

Civil Aeronautics Board (1953), Civil Air Regulations Part 4b, 4b.611 (b) Code of Federal Regulations.

Cole, E. L., Milton, J. L., and McIntosh, B. B. (1954), "Routine Maneuvers under Day and Night Conditions, Using an Experimental Panel, Eye Fixations of Aircraft Pilots ix," Technical Report No. 53-220, U.S. Air Force Wright Air Development Center, Wright Patterson AFB, OH.

Cotton, B. (2003, July), "For Spacious Skies—The Twenty-First Century Promise of Free Flight," paper presented at the AIAA International Air and Space Symposium and Exhibition: The Next 100 Years, Dayton, OH.

Cowan, N. (2000), "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity," *Behavioral and Brain Sciences*, Vol. 24, pp. 87–185.

Craik, F. I. M., and Tulving, E. (2004), "Depth of Processing and the Retention of Words in Episodic Memory," in Balota, D. A. and Marsh, E. J. (Eds.), *Key Readings in Cognition: Cognitive Psychology*. New York: Psychology Press

Dashevsky, S. G. (1964), "Check-Reading Accuracy as a Function of Pointer Alignment, Patterning, and Viewing Angle," *Journal of Applied Psychology*, Vol. 48, pp. 344–347.

Deaton, J. E., Barnes, M. J., Lindsey, N., and Kern, J. (1989), "The Effect of Windscreen Bows and HUD Pitch Ladder on Pilot Performance during Simulated Flight," paper presented at the Human Factors Society 33rd Annual Meeting, Denver, CO.

Debelack, A. S., Dehn, J. D., Muchinsky, L. L., and Smith, D. M. (1995), "Next Generation Air Traffic Control Automation," *IBM Systems Journal*, Vol. 34, pp. 63–77.

Devaluez, F. (1924), "Investigations of Aviation Accidents and Lessons to be Drawn from Them," Technical Memorandum No. 245, National Advisory Committee for Aeronautics, Washington, DC.

Dhillon, B. S. (2007), *Applied Reliability and Quality*, Springer-Verlag, London.

Dwyer, J., and Landry, S. J. (2009), "Separation Assurance and Collision Avoidance Concepts for the Next Generation Air Transportation System," paper presented at the 13th International Conference on Human-Computer Interaction, San Diego, CA.

Eaton, H. N. (1921), "Aerial Navigation and Navigating Instruments," Technical Report No. 131, National Advisory Committee for Aeronautics, Washington, DC.

Eby, M. S., Kelly Iii, W. E., Inc, S. C. S., and Wichita, K. S. (1999), "Free Flight Separation Assurance Using Distributed Algorithms," in IEEE (Ed.), *Proceedings of the 1999 Aerospace Conference*, Vol. 2, Institute of Electrical and Electronics Engineers, New York.

Elkin, E. H. (1959), "Effect of Scale Shape, Exposure Time and Display Complexity on Scale Reading Efficiency," Technical Report No. 58-472, U.S. Air Force Wright Air Development Center, Wright Patterson AFB, OH.

Endsley, M. R. (1995), "Toward a Theory of Situation Awareness in Dynamic Systems," *Human Factors*, Vol. 37, pp. 32–64.

Endsley, M. R., and Kaber, D. B. (1999), "Level of Automation Effects on Performance, Situation Awareness and Workload in a Dynamic Control Task," *Ergonomics*, Vol. 42, No. 3, pp. 462–492.

Endsley, M. R., and Kiris, E. O. (1995), "The Out-of-the-Loop Performance Problem and Level of Control in Automation," *Human Factors*, Vol. 37, pp. 381–394.

Endsley, M. R., Bolté, B., and Jones, D. G. (2003), *Designing for Situation Awareness: An Approach to User-Centered Design*, New York: Taylor & Francis.

Eriksen, C. W., and Eriksen, B. A. (1974), "Effects of Noise Letter upon the Identification of a Target Letter in a Nonsearch Task," *Perception and Psychophysics*, Vol. 16, pp. 143–149.

Federal Aviation Administration (FAA) (2007), "Operator's Manual: Human Factors in Airport Operations," Report No. HFM-AO-07-01, FAA, Washington, DC.

Fitts P. M. (1951), *Human Engineering for an Effective Air-Navigation and Traffic-Control System*, National Research Council Committee on Aviation Psychology, Washington, DC.

Fitts, P. M., Jones, R. E., and Milton, J. L. (1950), "Eye Movements of Aircraft Pilots during Instrument-Landing Approaches," *Aeronautical Engineering Review*, Vol. 9, No. 2, pp. 1–6.

Fogel, L. (1959), "A New Concept: The Kinalog Display System," *Human Factors*, Vol. 1, No. 1, pp. 30–37.

Francis, G. (2000), "Designing Multifunction Displays: An Optimization Approach," *International Journal of Cognitive Ergonomics*, Vol. 4, pp. 107–124.

Gerovitch, S. (2003), "Interview with Sheridan," available: http://web.mit.edu/slava/space/interview/interview-sheridan.htm.

Gigerenzer, G., and Goldstein, D. G. (1996), "Reasoning the Fast and Frugal Way: Models of Bounded Rationality," *Psychological Review*, Vol. 103, No. 4, pp. 650–669.

Gough, M. N. (1936), "The Measurement of the Field of View from Airplane Cockpits," Technical Report No. 514, National Advisory Committee for Aeronautics, Washington, DC.

Gough, M. N., and Beard, A. P. (1936), "Limitations of the Pilot in Applying Forces to Airplane Controls," Technical Report No. 550, National Advisory Committee for Aeronautics, Washington, DC.

Hacker, E. W. (1984), "Learning from the Past: A Fighter Pilot's Obligation," available: http://www.globalsecurity.org/military/library/report/1984/HEW.htm, accessed July 29, 2005.

Hancock, P. A., Williams, G., and Manning, C. M. (1995), "Influence of Task Demand Characteristics on Workload and Performance," *International Journal of Aviation Psychology*, Vol. 5, pp. 63–86.

Harper, R., and Hughes, J. (1991), "What a F-Ing System! Send 'Em All to the Same Place and Then Expect Us to Stop 'Em Hitting: Making Technology Work in Air Traffic Control," in *Technology in Working Order: Studies of Work, Interaction, and Technology*, G. Burton, Ed., Rank Xerox Cambridge EuroPARC, Cambridge.

Hart, S. G., and Staveland, L. E. (1988), "Development of the NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Human Mental Workload*, P. A. Hancock and N. Meshkati, Eds., North-Holland, Amsterdam, pp. 139–183.

Hawkins, F. (1987), *Human Factors in Flight*, Ashgate, Aldershot, England.

Heglin, H. J. (1973), "NAVSHIPS Display Illumination Design Guide: Human Factors," Report No. NELC/TD223, Naval Electronics Laboratory Center, San Diego, CA.

Hersey, M. D. (1923), "Aeronautic Instruments. Section I: General Classification of Instruments and Problems Including Bibliography," Document No. 125, National Aeronautics and Space Administration, Langley, VA.

Hobbs, A. (2008), "An Overview of Human Factors in Aviation Maintenance," Report No. AR-2008-055, Australian Transport Safety Bureau, Canberra City, Australia.

Hobbs, A., and Williamson, A. (2003), "Associations between Errors and Contributing Factors in Aircraft Maintenance," *Human Factors*, Vol. 45, pp. 186–201.

Horrey, W. J., Wickens, C. D., and Consalus, K. P. (2006), "Modeling Drivers' Visual Attention Allocation While Interacting with In-Vehicle Technologies," *Journal of Experimental Psychology: Applied*, Vol. 12, No. 2, pp. 67–78.

Hu, J., Prandini, M., and Sastry, S. (2002), "Optimal Coordinated Maneuvers for Three-Dimensional Aircraft Conflict Resolution," *Journal of Guidance, Control, and Dynamics*, Vol. 25, pp. 888–900.

Hughes, T. P. (1971). *Elmer Sperry - Inventor and Engineer*. Baltimore: The Johns Hopkins University Press.

Hunt, F. L. (1926), "New Types of Aircraft Instruments," *Journal of the Optical Society of America*, Vol. 12, pp. 227–269.

Johnson, T. L. (2007), "Improving Automation Software Dependability: A Role for Formal Methods?" *Control Engineering Practice*, Vol. 15, pp. 1403–1415.

Johnson, W. W., Battiste, V., Delzell, S., Holland, S., Belcher, S., and Jordan, K. (1997), "Development and Demonstration of a Prototype Free Flight Cockpit Display of Traffic Information," paper presented at the 1997 SAE/AIAA World Aviation Congress, Anaheim, CA.

Kazda, A., and Caves, R. E. (2007), *Airport Design and Operation*, Elsevier, Oxford.

Kirwan, B., and Ainsworth, L. K. (1992), *A Guide to Task Analysis*, Boca Raton, FL: Taylor and Francis.

Klein, G., Orasanu, J., Calderwood, R., and Zsambok, C. E. (1993), *Decision Making in Action: Models and Methods*, Ablex, Norwood, NJ.

Kleinman, D., Baron, S., and Levison, W. (1970), "An Optimal Control Model of Human Response Part I: Theory and Validation," *Automatica*, Vol. 6, pp. 357–369.

Korkmaz, S. V., et al. (2006), "Baggage Handling in an Airplane Cargo Hold: An Ergonomic Intervention Study," *International Journal of Industrial Ergonomics*, Vol. 36, pp. 301–312.

Krozel, J., Peters, M., and Bilimoria, K. (2000), "A Decentralized Control Strategy for Distributed Air/Ground Traffic Separation," paper presented at the 18th AIAA Guidance, Navigation, and Control Conference, Denver, CO.

Kuchar, J. K., and Carpenter, B. D. (1997), "Airborne Collision Alerting Logic for Closely Spaced Parallel Approach," *Air Traffic Control Quarterly*, Vol. 5, pp. 111–127.

Kuchar, J. K., and Drumm, A. C. (2007), "The Traffic Alert and Collision Avoidance System," *Lincoln Laboratory Journal*, Vol. 16, pp. 277–296.

Landry, S. J., Sheridan, T. B., and Yufik, Y. M. (2001), "A Methodology for Studying Cognitive Groupings in a Target Tracking Task," *IEEE Transactions on Intelligent Transportation Systems*, Vol. 2, pp. 92–100.

Landry, S. J., Nickelson, M., Levin, K., and Rowe, D. (2007, October), "Creating Effective Collaboration Tools: Lessons from the Multi-Center Traffic Management Advisor," paper presented at the 51st Annual Meeting of the Human Factors and Ergonomics Society, Baltimore, MD.

Landry, S. J., Lagu, A., and Kinnari, J. (2010), "State-Based Modeling of Continuous Human-Integrated Systems: An Application to Air Traffic Separation Assurance," *Reliability Engineering and Safety Science*, Vol. 95, pp. 345–353.

Lee, K. K., Quinn, C. M., Hoang, T., and Sanford, D. (2000), "Human Factors Report: TMA Operational Evaluations 1996 and 1998," Report No. NASA/TM-2000-209587, NASA Ames Research Center, Moffett Field, CA.

Leveson, N. (2004), "A New Accident Model for Engineering Safer Systems," *Safety Science*, Vol. 42, pp. 237–270.

Levy, J. L., Foyle, D. C., and McCann, R. S. (1998, October), "Performance Benefits with Scene-Linked HUD Symbology: An Attentional Phenomenon?" paper presented at the the 42nd Annual Meeting of the Human Factors and Ergonomics Society, Santa Monica, CA.

Ljundberg, M., and Lucas, A. (1992), "The OASIS Air Traffic Management System," Report No. 28, Australian Artificial Intelligence Institute, Melbourne, Australia.

Malloy, R., and Parasuraman, R. (1996), "Monitoring an Automated System for a Single Failure: Vigilance and Task Complexity Effects," *Human Factors*, Vol. 38, pp. 311–322.

Marx, D. A., and Graeber, R. C. (1994), "Human Error in Aircraft Maintenance," in *Aviation Psychology in Practice*, N. Johnston, N. McDonald, and R. Fuller, Eds., Avebury, Aldershot, United Kingdom, pp. 87–104.

McNally, D., and Gong, C. (2007), "Concept and Laboratory Analysis of Trajectory-Based Automation for Separation Assurance," *Air Traffic Control Quarterly*, Vol. 15, No. 1, pp. 35–63.

McRuer, D., and Graham, D. (1965), "Human Pilot Dynamics in Compensatory Systems," Report No. AAFFDL-TR-65-15, Air Force Flight Dynamics Laboratory, Wright-Patterson AFB, OH.

Mendel, M. B., and Sheridan, T. B. (1986), "Optimal Combination of Information from Multiple Sources," Office of Naval Research, Arlington, VA.

Miller, G. A. (1956), "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psychological Review*, Vol. 63, pp. 81–97.

National Advisory Committee for Aeronautics (NACA) (1921), "Staff Meeting of the Members of the German Scientific Association for Aeronautics," Technical Memorandum No. 20, NACA, Washington, DC.

National Transportation Safety Board (NTSB) (1979), "Aircraft Accident Report: United Airlines, Inc. McDonnell-Douglas, DC-8-61, N8082U," Report No. NTSB-AAR-79-7, NTSB, Washington, DC.

NavCanada (2003), "SASS: Sequencing and Scheduling System," available: http://www.navcanada.ca/content definitionsfiles/TechnologySolutions/products/Stand Alone/sass/SASSen.pdf.

Neal, A., and Kwantes, P. J. (2009), "An Evidence Accumulation Model for Conflict Detection Performance in a Simulated Air Traffic Control Task," *Human Factors*, Vol. 51, pp. 164–180.

Norman, D. A. (1993), *Things That Make Us Smart*, Addison-Wesley, Reading, MA.

O'Connor, P. D. T. (2002), *Practical Reliability Engineering*, 4th ed., Wiley, Chichester..

Orasanu, J., and Martin, L. (1998), "Errors in Aviation Decision Making: A Factor in Accidents and Incidents," paper presented at the Workshop on Human Error, Safety, and Systems Development, Seattle, WA.

Prevot, T., Homola, J., and Mercer, J. (2008a), "Human-in-the-Loop Evaluation of Ground-Based Automated Separation Assurance for NEXTGEN," paper presented at the 8th AIAA AIrcraft Technology, Integration, and Operations Conference, Anchorage, Alaska.

Prevot, T., Homola, J., and Mercer, J. (2008b, August 18–21), "Initial Study of Controller/Automation Integration for NextGen Separation Assurance," paper presented at the Guidance, Navigation, and Control Conference, Honolulu, HI.

Pulat, B. M., and Ayoub, M. A. (1979), "A Computer-Aided Instrument Panel Design Procedure," *Proceedings of the Human Factors Society*, Vol. 23, pp. 191–192.

Rantanen, E. M., and Nunes, A. (2005), "Hierarchical Conflict Detection in Air Traffic Control," *International Journal of Aviation Psychology*, Vol. 15, pp. 339–362.

Rauzy, A. B. (2008), "Guarded Transition Systems: A New States/Events Formalism for Reliability Studies," *Journal of Risk and Reliability*, Vol. 222, pp. 495–505.

Rehmann, A. J. (1995), "Flightdeck Crew Alerting Issues: An Aviation Safety Reporting System Analysis," Federal Aviation Administration, Washington, DC.

Remington, R. W., HJohnston, J. C., Ruthruff, E., Gold, M., and Romera, M. (2000), "Visual Search in Complex Displays: Factors Affecting Conflict Detection by Air Traffic Controllers," *Human Factors*, Vol. 42, pp. 349–366.

Robinson III, J. E., Davis, T. J., and Isaacson, D. R. (1997), "Fuzzy Reasoning-Based Sequencing of Arrival Aircraft in the Terminal Area," paper presented at the 15th AIAA Guidance, Navigation, and Control Conference, New Orleans, LA.

Roscoe, S. N. (1968), "Airborne Displays for Flight and Navigation," *Human Factors*, Vol. 10, No. 4, pp. 321–332.

Roscoe, S. N. (1997a), "The Adolescence of Engineering Psychology," in *Human Factors History Monograph Series*, S. M. Casey, Ed., Vol. 1, Human Factors and Ergonomics Society, Santa Monica, CA.

Roscoe, S. N. (1997b), "Horizon Control Reversals and the Graveyard Spiral," *CSERIAC Gateway*, Vol. 7, No. 3, pp. 1–4.

Roscoe, S. N., Corl, L., and Jensen, R. S. (1981), "Flight Display Dynamics Revisited," *Human Factors*, Vol. 23, No. 3, pp. 341–353.

Sabeh, R., Jorve, W. R., and Vanderplas, J. M. (1958), "Shape Coding of Aircraft Instrument Zone Markings," Technical Note 57-260, U.S. Air Force Wright Air Development Center, Wright-Patterson AFB, OH.

Sanders, and McCormick, E. J. (1993), *Human Factors in Engineering and Design*, McGraw-Hill, New York.

Sarter, N., and Woods, D. (1995), "How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control," *Human Factors: Journal of the Human Factors and Ergonomics Society*, Vol. 37, pp. 5–19.

Schneiderman, B. (1998), *Designing the User Interface*, Addison-Wesley, Cambridge, MA.

Schoenig, R., Aubry, J.-F., Cambois, T., and T., H. (2006), "An Aggregation Method of Markov Graphs for the Reliability Analysis of Hybrid Systems," *Reliability Engineering and System Safety*, Vol. 84, pp. 137–148.

Seidler, K., and Wickens, C. D. (1992), "Distance and Organization in Multifunction Displays," *Human Factors*, Vol. 34, pp. 555–569.

Selcon, S. J., Taylor, R. M., and McKenna, F. P. (1995), "Integrating Multiple Information Sources: Using Redundancy in the Design of Warnings," *Ergonomics*, Vol. 38, pp. 2362–2370.

Sheridan, T. B. (2002), *Humans and Automation: System Design and Research Issues*, Santa Monica, CA: Human Factors and Ergonomics Society.

Simmonds, G. R., Galer, M., and Baines, A. (1981), *Ergonomics of Electronic Displays*, Society of Automotive Engineers, Warrendale, PA.

Simpson, P. A. (2001), "Naturalistic Decision Making in Aviation Environments," Report No. DSTO-GD-0279, DSTO Aeronautical and Maritime Research Laboratory, Fisherman's Bend, Australia.

Stankovic, S., Raufaste, E., and Averty, P. (2008), "Determinants of Conflict Detection: A Model of Risk Judgments in Air Traffic Control," *Human Factors*, Vol. 50, pp. 121–134.

Subcommittee on Instruments (1932), "Present Status of Aircraft Instruments," Report No. 371, National Advisory Committee for Aeronautics, Washington, DC.

Suchman, L. (1987), *Plans and Situated Actions: The Problem of Human-Machine Communication*, Cambridge University Press, Cambridge.

Teo, R., and Tomlin, C. J. (2003), "Computing Danger Zones for Provably Safe Closely Spaced Parallel Approaches," *Journal of Guidance, Control, and Dynamics*, Vol. 26, pp. 434–443.

Thompson, F. L. (1938), "N.A.C.A. Stall-Warning Indicator," Technical Note No. 670, National Advisory Committee for Aeronautics, Washington, DC.

Tipper, S. P., and Driver, J. (1988), "Negative Priming between Pictures and Words in a Selective Attention Task: Evidence for Semantic Processing of Ignored Stimuli," *Memory and Cognition*, Vol. 16, No. 1, pp. 64–70.

VanRullen, and Koch, C. (2003), "Competition and Selection during Visual Processing of Natural Scenes and Objects," *Journal of Vision*, Vol. 3, No. 1, pp. 75–85.

VanRullen, and Thorpe, S. J. (2001), "The Time Course of Visual Processing: From Early Perception to Decision-Making," *Journal of Cognitive Neuroscience*, Vol. 13, No. 4, pp. 454–461.

Ververs, M. P., and Wickens, C. D. (1998), "Head-Up Displays: Effects of Clutter, Symbology, Intensity, and Display Location on Pilot Performance," *International Journal of Aviation Psychology*, Vol. 8, pp. 377–403.

Volovoi, V. (2004), "Modeling of System Reliability Petri Nets with Aging Tokens," *Reliability Engineering and System Safety*, Vol. 84, pp. 149–161.

Warner, E. P. (1922a), "Safety in Flight," Technical Memorandum No. 133, National Advisory Committee for Aeronautics, Washington, DC.

Warner, E. P. (1922b), "Stability of Airplanes," Technical Memorandum No. 92, National Advisory Committee for Aeronautics, Washington, DC.

White, F. (1968), "Airborne Collision Avoidance System Development: An Introduction," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 4, pp. 234–237.

Whitfield, D., and Jackson, A. (1983), "The Air Traffic Controller's Picture as an Example of a Mental Model," paper presented at the IFAC Conference on Analysis, Design, and Evaluation of Man-Machine Systems, Baden-Baden, Germany.

Wickens, C. D. (2003), "Aviation Displays," in *Principles and Practice of Aviation Psychology*, P. Tsang and M. Vidulich, Eds., Lawrence Erlbaum, Mahwah, NJ, pp. 147–199.

Wickens, C. D., and Carswell, C. M. (1995), "The Proximity Compatibility Principle: Its Psychological Foundation and Its Relevance to Display Design," *Human Factors*, Vol. 37, No. 3, pp. 473–494.

Wickens, C. D., and Long, J. (1995), "Object vs. Spaced-Based Models of Visual Attention: Implications for the Design of Head-Up Displays," *Journal of Experimental Psychology: Applied*, Vol. 1, No. 3, pp. 179–194.

Wickens, C. D., Goh, J., Helleberg, J., Horrey, W. J., and Talleur, D. A. (2003), "Attentional Models of Multitask Pilot Performance Using Advanced Display Technology," *Human Factors*, Vol. 45, pp. 360–380.

Wickens, C. D., Rice, S., Keller, D., Hutchins, S., Hughes, J., and Clayton, K. (2009), "False Alerts in Air Traffic Control Conflict Alerting System: Is There a 'Cry Wolf' Effect?" *Human Factors*, Vol. 51, pp. 446–462.

Wiener, E. L. (1987), "Application of Vigilance Research: Rare, Medium, or Well Done?" *Human Factors*, Vol. 29, pp. 725–736.

Yang, L. C., and Kuchar, J. K. (1997), "Prototype Conflict Alerting System for Free Flight," *Journal of Guidance, Control, and Dynamics*, Vol. 20, No. 4, pp. 768–773.

Yang, L. C., and Kuchar, J. K. (2002), "Performance Metric Alerting: A New Design Approach for Complex Alerting Problems," *IEEE Transactions on Systems, Man and Cybernetics,- Part A: Systems and Humans*, Vol. 32, pp. 123–134.

Yeh, M., Merlo, J. L., Wickens, C. D., and Brandenburg, D. L. (2003), "Head Up Versus Head Down: The Costs of Imprecision, Unreliability, and Visual Clutter on Cue Effectiveness for Display Signaling," *Human Factors*, Vol. 45, pp. 390–408.

Zhang, J., and Norman, D. A. (1994), "Representations in Distributed Cognitive Tasks," *Cognitive Science*, Vol. 18, pp. 87–122.

# INDEX